Clim Dyn
DOI 10.1007/s00382-017-3561-9

CrossMark

How accurately do we know the temperature of the surface of the earth?

S. Lovejoy[1]

Received: 31 August 2016 / Accepted: 30 January 2017

Abstract The earth's surface temperature is important in a variety of applications including for quantifying global warming. We analyze 6 monthly series of atmospheric temperatures from 1880 to 2012, each produced with different methodologies. We first estimate the relative error by systematically determining how close the different series are to each other, the error at a given time scale is quantified by the root mean square fluctuations in the pairwise differences between the series as well as between the individual series and the average of all the available series. By examining the differences systematically from months to over a century, we find that the standard short range correlation assumption is untenable, that the differences in the series have long range statistical dependencies and that the error is roughly constant between 1 month and one century—over most of the scale range, varying between ±0.03 and ±0.05 K. The second part estimates the absolute measurement errors. First we make a stochastic model of both the true earth temperature and then of the measurement errors. The former involves a scaling (fractional Gaussian noise) natural variability term as well as a linear (anthropogenic) trend. The measurement error model involves three terms: a classical short range error, a term due to missing data and a scale reduction term due to insufficient space–time averaging. We find that at 1 month, the classical error is ≈±0.01 K, it decreases rapidly at longer times and it is dominated by the others. Up to 10–20 years, the missing data error gives the dominate contribution to the error: 15 ± 10% of the temperature variance; at scales >10 years, the scale reduction factor dominates, it increases the amplitude of the temperature anomalies by 11 ± 8% (these uncertainties quantify the series to series variations). Finally, both the model itself as well as the statistical sampling and analysis techniques are verified on stochastic simulations that show that the model well reproduces the individual series fluctuation statistics as well as the series to series fluctuation statistics. The stochastic model allows us to conclude that with 90% certainty, the absolute monthly and globally averaged temperature will lie in the range −0.109 to 0.127 °C of the measured temperature. Similarly, with 90% certainty, for a given series, the temperature change since 1880 is correctly estimated to within ±0.108 of its value.

Keywords Global temperature · Uncertainty · Scaling · Stochastic modelling

1 Introduction

The atmosphere is a turbulent fluid and the temperature and other state variables fluctuate from the age of the earth down to milliseconds, in space from the size of the planet down to millimeters (see Lovejoy (2015) for a review). Global scale temperature estimates rely on sparse (i.e. fractal), in situ measurement networks (Lovejoy et al. 1986; Nicolis 1993; Mazzarella and Tranfaglia 2000) and mapping them onto regular grids (e.g. with interpolation or Kriging) involves nontrivial space–time homogeneity, smoothness and other assumptions. In the satellite era and with other suppositions, remotely sensed data may also be used (e.g. Mears et al. 2011).

Even the problem of mapping a single spatially point-like in situ measurement onto a finite resolution grid is

✉ S. Lovejoy
lovejoy@physics.mcgill.ca

[1] Department of Physics, McGill University, 3600 University st., Montreal, Que H3A 2T8, Canada

Journal : Large 382	Article No : 3561	Pages : 18	MS Code : CLDY-D-16-00656	Dispatch : 17-3-2017

nontrivial. At first sight it would appear that the problem is even ill-posed because it seems to be an attempt to change the resolution of the data by an infinite factor: from zero to tens or hundreds of kilometers. However, such spatially point-like data are never point-like in space–time and it is the effective space–time resolution that is important. For example in the weather regime (i.e. for time scales up to the lifetime of planetary structures, typically ≈10 days), the space–time relation is linear or 2/3 power law for Eulerian and Lagrangian frames respectively (see Lovejoy and Schertzer 2010, 2013) for both short and extended reviews). However for time scales with resolutions longer than typical (5–10 day) planetary lifetimes (the macroweather regime) to a good approximation the space–time statistics factorize (Sect. 2.4) so that there is quite different time-scale to space-scale relation (Lovejoy and de Lima 2015; Lovejoy et al. 2017). The observed spatial scaling relations (which are also respected by the GCM models—although with slightly different exponents), indicate that the regularity and smoothness assumptions made by classical geostatistical techniques such as Kriging are not applicable. Below, we show that a consequence of the scaling is the existence of "scale reduction factors" that are nonclassical but yet are needed to explain the low frequency part of the observations.

In addition to problems due to sparse networks and unknown or ill-defined space–time resolutions, there are also practical issues such as estimating the temperature over the ocean and over sea ice and with frequent series discontinuities and biases caused by the heat island and cool park effects (Parker 2006; Peterson 2003). Sea surface temperatures series also have nontrivial issues, see Hausfather et al. (2016).

Even high quality surface networks such as the US Historical Climatology Network "have an average of six discontinuities per century and not a single station is homogeneous for its full period of record" (Peterson 2003). Another potential source of bias is the fact that starting at around 1950, the rate of increase of nocturnal (minimum) temperature values on land was almost twice as high when compared to that of diurnal (maximum) temperature values, favouring an increase of duration of the frost-free period in many regions of moderate and high latitudes (Kondratyev and Varotsos 1995; Efstathiou and Varotsos 2010). See also Pielke et al. (2007) who enumerates many other issues and Diamond et al. (2013) who reviews their implications.

Yet in spite of these problems and in order to provide a reliable indicator of the state of the climate, half a dozen centennial, global scale surface air temperature estimates have been produced. The question of their accuracy is essential for many applications, including global warming: indeed, one of the oldest climate skeptic arguments against anthropogenic warming is that the data are unreliable or biased. It is therefore important to quantify their accuracy.

We analyse the six best-documented (at the time of analysis; May 2015) global, monthly averaged time series. Each series was constructed with somewhat different data, with different homogenization and gridding assumptions. Since no absolute ground truth is available, their authors used specific theoretical space–time assumptions and models to quantify the accuracy of each temperature series statistics in order to obtain monthly resolution uncertainty estimates. Yet historically, when confronted with the measurement of a new physical quantity—here the global average surface temperature—the greatest confidence comes from the agreement between qualitatively different ~~and~~ physical consistent approaches. We therefore systematically compare each series with the others determining the relative accuracy as functions of scale [Sect. 2; this idea and an early spectral result were given in Lovejoy et al. (2013a)]. This analysis motivates the development of a model for the absolute accuracy that is developed in Sect. 3. Whereas in Sect. 2, we ask the relative accuracy question: "how well do different methods using different empirical inputs agree with each other as functions of their time scale?", in Sect. 3 we move from relative to absolute estimates of error and bias attempting to answer the question "how accurate are the data as functions of their time scale?"

The explicit treatment of scale is important because over the range of between about 10 days and 10 years (the macroweather regime) the fluctuations (precisely defined below) tend to cancel each other out: increasing fluctuations tend to be followed by decreasing ones so that temporal averages (of essentially all atmospheric quantities) systematically decrease with scale (Lovejoy 2013; Fig. 2 below). At scales beyond ≈10–20 years (the climate regime) the temperature is dominated by anthropogenic effects and the fluctuations increase with scale. In addition, we conventionally expect that lowering the temporal resolution by averaging over longer and longer time intervals will lead to the convergence of each globally averaged temperature series to the actual temperature so that with sufficient averaging (i.e. with low enough temporal resolution) and in accord with the central limit theorem, the different series are expected to mutually converge. The direct way to analyze this is by considering the fluctuations in the differences between the different series and to quantify how rapidly they diminish with temporal resolution. The only technical complication is that we must use an appropriate definition of fluctuation. This is because on average, the classical fluctuations (defined as differences) cannot decrease with scale, so that for our purposes, they are inadequate. Instead, we use the somewhat different Haar fluctuations.

AQ3

Journal : Large 382 | Article No : 3561 | Pages : 18 | MS Code : CLDY-D-16-00656 | Dispatch : 17-3-2017

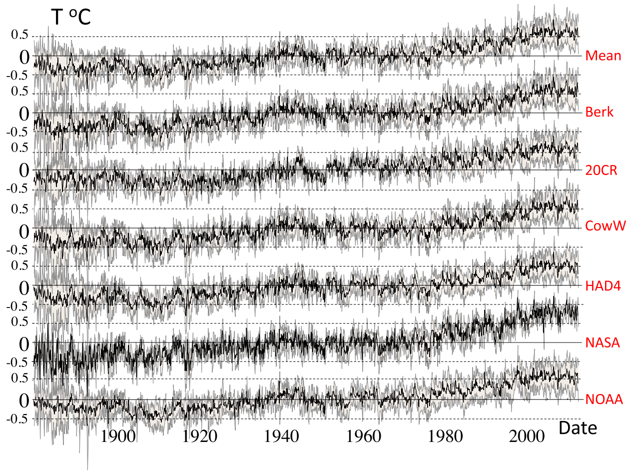


Fig. 1 The six monthly global surface temperature anomaly series from 1880 to 2012 (*black*) with 3 standard deviation uncertainties in grey with the mean of all six (*top*). From bottom to top: NOAA NCDC, NASA GISS, Hadcrutem4, Cowtan and Way, the 20 Century Reanalysis, the Berkeley series and the overall mean. Each series represents the anomaly with respect to the mean of the entire period, indicated by the *black horizontal axes*. For each of the bottom six series, the uncertainties are determined from the standard deviations of the other five

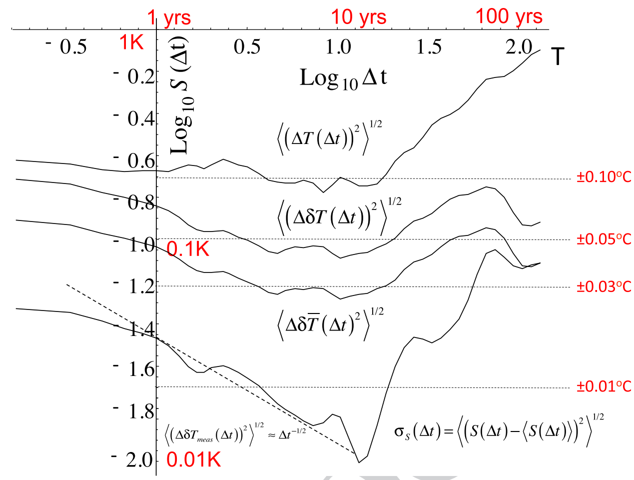


Fig. 2 The RMS Haar fluctuations $S(\Delta t)$ averaged over the six series (*top*), averaged over all the 15 pairs of differences (second from *top*), averaged over the differences of each with respect with the overall mean of the six series (third from *top*), and the standard deviation of the $S(\Delta t)$ curves evaluated for each of the series separately (*bottom*). Also shown for reference (*dashed*) is the line that data with independent Gaussian noise would follow

2 Fluctuation analysis

2.1 The data

The series that we chose were all publically available at monthly resolutions between January 1880 and December 2012 (133 years = 1596 months). They were (a) the NOAA NCDC series GHCN-M version 3.2.0 dataset (Smith et al. 2008), updated in Williams et al. (2012), abbreviated NOAA in the following, (b) the NASA Goddard Institute for Space Studies Surface Temperature Analysis (GISTEMP) series, abbreviated NASA (Hansen et al. 2010), (c) the Combined land and sea surface temperature (SST) anomalies from HadSST3, Hadley Centre–Climatic Research Unit Version 4, abbreviated HAD4 (Brohan et al. 2006; Kennedy et al. 2011), (d) the version 2 series of (Cowtan and Way 2014) (abbreviated CowW), (e) the Twentieth Century reanalysis, version 2 (Compo et al. 2011), (20CR) and (f) the Berkeley Earth series (Rohde et al. 2013) abbreviated Berk. Shortly after these series were analyzed, some of the series were updated (notably by Karl et al. 2015), but we are not trying to establish which series is best, but rather how the errors vary with scale so that the updates are unlikely to alter the conclusions.

Each data set has its particular strengths and weaknesses, we enumerate a few of these in order to underline their diversity. For example, NOAA and NASA use essentially the same land and marine data, but use different methods to fill (some) of the data holes. In contrast the HAD4 series makes no attempt in this direction, thus making fewer assumptions about the spatial statistical properties (especially smoothness, regularity properties). The CowW series takes the contrary view: it uses the HAD4 data but makes strong spatial statistical assumptions (Kriging) to fill in data holes. This is especially significant in the data poor high latitude regions. The 20CR series is of particular interest here because it uses no temperature station data whatsoever. Instead, it uses surface pressure station data and monthly SST data (the same as HADCRUT4) combined with a numerical model (a reanalysis), it is the only series that gives actual temperatures rather than changes with respect to a reference period: "anomalies". The fact that the 20CR agrees well with the other (station based temperature) estimates is strong support for all the series (Compo et al. 2013). Finally, the Berk series uses the same SST data as both HAD4 and CowW but it uses data from many more stations (≈37,000 compared to only 4500 for HAD4 and 7300 for the NOAA series for example), and it uses a number of statistical improvements in the handling of data homogenization and coverage. Our objective here is not to attempt to evaluate which assumptions, or which products are better—or worse—our point is that there is a significant diversity so that the degree of agreement or disagreement between the various series is of itself important.

Figure 1 shows a visual comparison of the series. In addition to the temperature (black), we have shown uncertainty limits (gray). These are not theoretical estimates of intrinsic uncertainty but rather the dispersion of the five temperature records about the given series

Journal : Large 382	Article No : 3561	Pages : 18	MS Code : CLDY-D-16-00656	Dispatch : 17-3-2017

(three standard deviations): it measures the series similitude/dissimilitude. Where gray regions extend far above and below the black lines, they indicate that there is little agreement between the curve in question (black) and the other series. Where the band is narrow, it indicates is strong agreement. Overall we see that each series is very similar to the others (including the particularly significant 20CR series); comparing any individual curve with that of the overall mean of the six (top curve), we see that no particular series stands out. In addition, before 1900—but also after 1980—the series are the most dissimilar so presumably the least reliable. While this is not surprising for the earlier (data poor) epoch, a priori, it is not obvious in the more recent period. In Sect. 3, it is explained by the differing scale reduction factors.

2.2 Anomalies, differences, Haar fluctuations

The uncertainties in Fig. 1 are limited to quantifying the similarities/differences at unique temporal resolutions: 1 month. Since as we go to lower resolutions measurement errors are increasingly averaged, we expect a progressively stronger agreement at longer times. Standard uncertainty analyses (e.g. Kennedy et al. 2011) assume that there are both long term biases and short term errors and that the latter have short-range (exponential) decorrelations (e.g. the errors are auto-regressive or kindred processes). But a growing body of work finds monthly resolution atmospheric fields have long range statistical dependencies (wide range temporal and spatial scaling, power laws, Lovejoy and Schertzer 1986; Bunde et al. 2004; Rybski et al. 2006; Mann 2011; Franzke 2012; Rypdal et al. 2013, see Lovejoy and Schertzer 2013 for a review). The issue of short versus long range correlations also has implications for trend uncertainty analysis, see Lovejoy et al. (2016).

To quantify the resolution effect, denote the true global temperature anomaly by $T(t)$ (i.e. the actual averaged temperature of the entire planet with the annual cycle removed and the overall mean of the series removed so that $<T>=0$ where "<.>" indicates averaging). Define the Δt resolution anomaly fluctuation by:

$$(\Delta T(\Delta t))_{anom} = \frac{1}{\Delta t}\int_{t-\Delta t}^{t} T(t')dt' \quad (1)$$

(We have suppressed the t dependence since we will assume that the fluctuation statistics are statistically stationary; this may be true even though—due to anthropogenic warming—the statistics of the temperature itself are nonstationary). Note that if we have anomaly data at "resolution t", i.e. averaged over time t, $T_\tau(t)$, then $T_\tau = (\Delta T(\tau))_{anom}$ a fact that will use below.

Let us denote the overall deviations from the true value $E_i(t)$ (we use the term "deviation" to include both biases and errors). Now denote the ith measured anomaly by:

$$T_i(t) = T(t) + E_i(t) \quad (2)$$

For large enough averaging interval (Δt), we expect that the deviation E will be increasingly averaged out so that for the ith and jth series $\left(\Delta T_i(\Delta t)\right)_{anom} \approx (\Delta T(\Delta t))_{anom} \approx \left(\Delta T_j(\Delta t)\right)_{anom}$. Alternatively, by defining the difference:

$$\delta T_{ij}(t) = T_i(t) - T_j(t) = \delta E_{ij}(t) \quad (3)$$

(We) have the simple result $\delta T_{ij}(t) = E_i(t) - E_j(t)$. If the deviations $E_i(t)$, $E_j(t)$ are short range processes (i.e. dominated by standard measurement errors with having exponential decorrelations such as autoregressive processes and their kin), we can use the central limit theorem to conclude that at large enough Δt (where $E_i(t)$, $E_j(t)$ are statistically independent) that the rate at which the root mean square (RMS) anomaly fluctuation approaches zero is:

$$\left\langle \Delta\delta T_{ij}(\Delta t)^2\right\rangle^{1/2} = \left\langle \Delta\delta E_{ij}(\Delta t)^2\right\rangle^{1/2} \propto \Delta t^{-1/2} \quad (4)$$

If the separation of the deviations into short term errors and long term biases is at all possible, then for large enough averaging scale (Δt) it should display a $\Delta t^{-1/2}$ regime for the anomaly fluctuations.

Before testing this prediction on the data, we must first discuss different definitions of fluctuations and their limitations. Anomaly fluctuations must on average decrease with averaging scale Δt, so that are only adequate when the fluctuations decrease with scale Δt. For fluctuations that increase with Δt, we can use the classical definition of fluctuation, the differences:

$$(\Delta T(\Delta t))_{diff} = T(t) - T(t-\Delta t) \quad (5)$$

In contrast to anomaly fluctuations, average differences cannot decrease with scale whereas in general, average fluctuations may either increase or decrease as over different ranges of Δt. We must therefore define fluctuations in a more general way; wavelets provide a fairly general framework for this. A simple expedient combines averaging and differencing while overcoming many of the limitations of each: the Haar fluctuation (from the Haar wavelet). It is simply the difference of the mean over the first and second halves of an interval:

$$(\Delta T(\Delta t))_{Haar} = \frac{2}{\Delta t}\int_{t-\Delta t/2}^{t} T(t')dt' - \frac{2}{\Delta t}\int_{t-\Delta t}^{t-\Delta t/2} T(t')dt' \quad (6)$$

(see Lovejoy and Schertzer 2012b for these fluctuations in a wavelet formalism). In words, the Haar fluctuation is the

Journal : Large 382	Article No : 3561	Pages : 18	MS Code : CLDY-D-16-00656	Dispatch : 17-3-2017

difference fluctuation of the anomaly fluctuation, it is also equal to the anomaly fluctuation of the difference fluctuation. In regions where the fluctuations decrease with scale we have:

$$\begin{aligned}(\Delta T(\Delta t))_{Haar} &\approx (\Delta T(\Delta t))_{anom} \quad \text{(decreasing with } \Delta t\text{)}\\ (\Delta T(\Delta t))_{Haar} &\approx (\Delta T(\Delta t))_{diff} \quad \text{(increasing with } \Delta t\text{)}\end{aligned} \tag{7}$$

In order that Eq. 7 is reasonably accurate, the Haar fluctuations need to be multiplied by a "calibration" factor; here we use the "canonical" value 2 although a more optimal value could be tailored to individual series.

Over ranges where the dynamics have no characteristic time scale, the statistics of the fluctuations are power laws so that:

$$\langle|\Delta T(\Delta t)|^q\rangle \propto \Delta t^{\xi(q)} \tag{8}$$

The left hand side is the qth order structure function and $\xi(q)$ is the structure function exponent. "<>" indicates ensemble averaging; for individual series this is estimated by temporal averaging (over the disjoint fluctuations in the series). The first order ($q = 1$) case defines the "fluctuation exponent" H:

$$\langle|\Delta T(\Delta t)|\rangle \propto \Delta t^H \tag{9}$$

In the special case where the fluctuations are quasi-Gaussian, $\xi(q)=qH$ and the Gaussian white noise case corresponds to $H = -1/2$ (i.e. $\xi(q) = -q/2$). More generally, there will be "intermittency corrections" so that $qH - \xi(q)=K(q)$ where $K(q)$ is a convex function with $K(1)=0$. $K(q)$ characterizes the multifractality associated with the intermittency.

Equation 9 shows that the distinction between increasing and decreasing fluctuations corresponds to the sign of H. It turns out that the anomaly fluctuations are adequate when $-1<H<0$ whereas the difference fluctuations are adequate when $0<H<1$ (Lovejoy and Schertzer 2013, ch. 5). In contrast, the Haar fluctuations are useful over the range $-1<H<1$ which encompasses virtually all geoprocesses, hence its more general utility. When H is outside the indicated ranges, then the corresponding statistical behaviour depends spuriously on either the extreme low or extreme high frequency limits of the data.

2.3 Temporal analysis and the relative measurement errors

Figure 2 (top curve), shows the result when we estimate the Haar temperature fluctuations and average them over all the available disjoint intervals Δt and over all the series, calculating the RMS Haar fluctuation:

$$S(\Delta t) = \left\langle (\Delta T(\Delta t))^2_{Haar} \right\rangle^{1/2} \tag{10}$$

the "structure function": below we drop the subscripts, all fluctuations are Haar. In a scaling regime, we therefore have:

$$S(\Delta t) \propto \Delta t^{\xi(2)/2} \tag{11}$$

If the intermittency is small ($K(q)\approx 0$), then $\xi(2)/2\approx H$ and $S(\Delta t) \propto \Delta t^H$. Note that we estimate $S(\Delta t)$ using all available disjoint intervals of size Δt. Since the number of disjoint intervals decreases as Δt increases, so does the sample size, hence the statistics are less reliable at large Δt explaining the somewhat "noisy" appearance of plots such as Fig. 2 or 3. The only way to completely quantify this effect is with a stochastic model of the process; this is done in Sect. 3.

Starting at the smallest (monthly) scales with fluctuations $\approx\pm 0.14$ K, the latter decrease slowly to $\approx$10 years, (roughly as Δt^H, with $H \approx -0.1$ see Lovejoy and Schertzer (2012a) and below) whereas for $\Delta t > \approx 10$ years they increase. This increase reflects the increasing dominance of anthropogenic forcing over the natural variability (Lovejoy et al. 2013b). How accurate is this curve? Figure 3 (top set) shows the individual $S(\Delta t)$ functions for each of the series, we see that they are very close. The bottom curve in Fig. 2 quantifies this closeness by determining the standard deviation σ_S of the $S(\Delta t)$ curves about the ensemble mean at the top of Fig. 2. We see that—as expected—σ_S decreases as $\Delta t^{-1/2}$—but only over the range over which natural variability is dominant—becoming as low as 0.01 °C (±0.005 °C) at decadal scales. At the longer time scales, the standard deviation increases implying a disagreement over the magnitude of multi-decadal and centennial variability

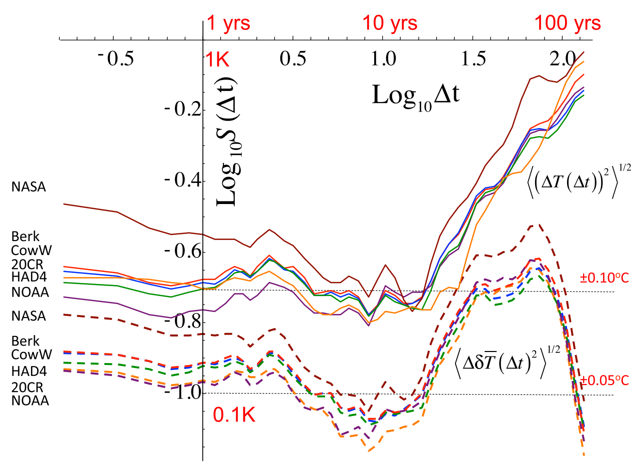


Fig. 3 The top set of curves (*solid*) are $S(\Delta t)$ for each of the different series, the bottom set (*dashed*) are the differences of each with respect to the mean of all the others: NOAA dark purple, NASA (*brown*), HAD4 (*green*), Cow (*blue*), 20CR (*orange*), Berk (*red*) (indicated at the *left* in the order of the curves)

i.e. disagreements of the order 0.06 to 0.1 K (±0.03 to ±0.05 K) for the total anthropogenic ~~forcing~~.

In most applications, we are interested the accuracy of the *temperature anomalies* themselves whereas the bottom curve in Fig. 2 only tells us about the accuracy of our estimate of their RMS *statistics*. To characterize the former, we analyze the fluctuations of the differences between series: $\Delta\delta T_{i,j}(\Delta t)$ (Eq. 3) or alternatively, between the ith series and the mean $\langle T(t)\rangle$ of all the series:

$$\delta \bar{T}_i(t) = T_i(t) - \langle T(t)\rangle \tag{12}$$

The second curve from the bottom is the RMS of the latter over all the series is: $\left\langle \Delta\delta\bar{T}(\Delta t)^2 \right\rangle^{1/2}$. In Fig. 2, the third curve from the bottom is the RMS of $\Delta\delta T_{i,j}(\Delta t)$ averaged over all the pairs of series: $\left\langle (\Delta\delta T(\Delta t))^2 \right\rangle^{1/2}$ (for N series, there are N (N − 1)/2 pairs, here N = 6 so that there are 15 pairs). Whereas $\left\langle (\Delta\delta T(\Delta t))^2 \right\rangle^{1/2}$ quantifies the typical difference between any two randomly chosen series at resolution Δt, $\left\langle \Delta\delta\bar{T}(\Delta t)^2 \right\rangle^{1/2}$ is the typical Δt resolution deviation of a series when the mean of all the series is considered the truth. A similar approach was recently used to estimate relative errors in climatological precipitation series in (de Lima and Lovejoy 2015).

Figure 2 shows a rather surprising result. While at first (from months to about 3–4 years), as expected—at least initially—the series do converge (they become closer to the overall mean), they do so considerably more slowly than expected for series with short range correlations. Rather than converging as $\Delta t^{-1/2}$ (Eq. 4), they converge as $\approx \Delta t^{-0.2}$ indicating long range statistical dependencies, confirming earlier results obtained using spectra (Lovejoy and Schertzer 2013) (appendix 10 C; scaling fluctuations imply power law spectra $E(\omega) \approx \omega^{-\beta}$ with $\beta = 1 + \xi(2)$ where ω is the frequency). Ignoring small intermittency corrections, $\beta = 1 + 2H$ so that a "flat" $S(\Delta t)$ curve ($\xi(2) \approx 0$) indicates a spectrum $E(\omega) \approx \omega^{-1}$. However, in the scale range $\Delta t > \approx 10$–20 years dominated by anthropogenic effects, the differences begin to *increase* and over the entire range of time scales, there is an irreducible (minimum) error ≈ ±0.03 °C to ±0.05 °C. Since the standard theory predicts a $\Delta t^{-1/2}$ fall-off: it fails at all scales so that different sources of error must be dominant (the effect of the finite sample size that decreases at larger Δt slightly increases the "noisiness" of the curves at larger Δt, and is probably responsible for the small downturn in the $S(\Delta t)$ curves of the differences at $\Delta t \approx > 100$ years). Indeed, the standard theory predicts centennial scale deviations of ≈ ±0.002 °C rather than the observed ± 0.03 °C to ± 0.05 °C (third curve, from the top, extreme right). Figure 2 also brings into question the utility of attempting to break the deviation into distinct short term measurement error and long term measurement bias components. The combination of error and bias is apparently present at all scales.

Figure 2 shows how any series differs from any other as well from the best estimate of the truth: the average over all of them. However, we may further quantify the monthly spreads in Fig. 1: for any given series, how close is it to the mean of the others (the relative measurement errors)? Fig. 3 shows the result: the top gives $\left\langle \Delta T_i(\Delta t)^2 \right\rangle^{1/2}$ for each of the six curves; we see that these statistics are indeed very similar (their dispersion is quantified in the bottom curve of Fig. 2). Note that the NASA curve has the steepest slope in the macroweather region ($\Delta t \approx < 10$–20 years) corresponding to $H \approx -0.2$ rather than $H \approx -0.1$ for the others. More interesting is the bottom set of curves $\left\langle \delta\bar{T}_i(\Delta t)^2 \right\rangle^{1/2}$, the difference between the ith series and the mean of the other five. From the top set we see that generally the NOAA and 20CR series have the weakest variability (top curves) whereas the NASA and Berk series have the strongest. From the bottom set we see that the NOAA and 20CR series are the closest to the other series whereas the NASA, Berk and CowW are the furthest (the most different). Aside from its obvious interpretation in terms of similitude and difference from one series to another, the statistics of $\left\langle \delta\overline{T}_i^2 \right\rangle^{1/2}$ will later be compared with the same quantity from stochastic simulations (Sect. 3) in order to validate them.

2.4 Space–time fluctuations, statistical factorization and the scale reduction factor

The differences between the series are due to the quantity and quality of the data that they use and the assumptions they use in order to grid them and then to space–time average them. In terms of the statistics of the resulting series, the former effect is largely associated with different amounts of missing data while the latter will affect the effective space–time resolution of the data. Both of these effects are important in modelling the errors; to model their effects, we require knowledge of the space–time statistics.

A space–time analysis of the 20th C reanalysis of the absolute temperatures (with only annual detrending) was already given in ch. 10 of (Lovejoy and Schertzer 2013). However, for our present purposes, the statistics of temperature anomalies—not temperature data—is needed; we therefore used the HADcrut anomaly data from 1880 at 5° × 5° spatial resolution. Figure 4 shows the result of estimating the RMS Haar spatial fluctuations over various spatial resolutions in the zonal direction, for the latter, the difference in the longitudinal angle Δθ was used, the fluctuation statistics being averaged over all latitudes from 60°S to 60°N (weighted by the grid box size—the latitude dependent map factor).

The top (monthly) resolution curve shows that the fluctuations decrease with increasing spatial scale. Since only

Journal : Large 382 | Article No : 3561 | Pages : 18 | MS Code : CLDY-D-16-00656 | Dispatch : 17-3-2017

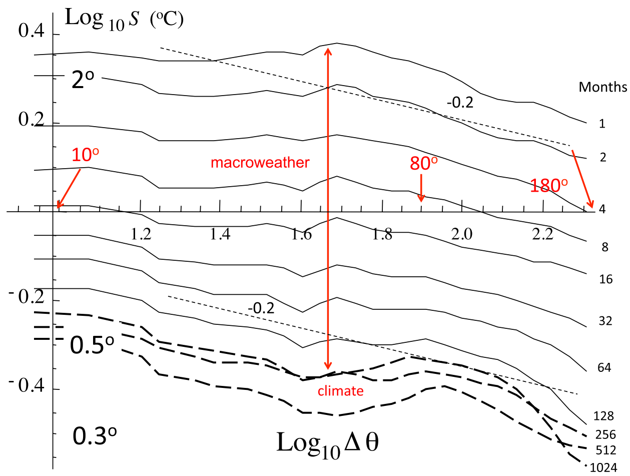


Fig. 4 The zonal spatial analysis of the HADCrut surface data (on a 5° × 5° grid) as functions of temporal averaging (systematically doubling from 1 month to 1024 months ≈ 85 years, *top* to *bottom*). Although it is "noisy", ~~roughly~~ the effect of temporal averaging is the decrease the amplitude of the fluctuations at all spatial scales. This is as predicted by the macroweather space–time factorization property. The double headed arrow shows the predicted downward shift from 1 to 128 months (*red* curves) with temporal $H_t = -0.3$. The reference line has slope $\xi_x(2)/2 = -0.2$

≈40% of the pixels had data, we used a Haar fluctuation algorithm that takes into account the missing data ("Appendix A" of Lovejoy 2015). This is important since if the data are interpolated, then the result is too smooth and can give spurious scaling (a smooth curve will have a Haar exponent $H=1$ rather than $H<0$ as in the data).

From Fig. 4 we can see that as the spatial resolution (Δθ) is increased, the anomaly fluctuations decrease with scale roughly as: $S_\theta(\Delta\theta) \propto \Delta\theta^{\xi(2)/2}$ with $\xi(2) = -0.4$. To interpret this result, recall that the spatial fluctuation exponent $H_s = \xi(1)$ is defined in terms of the mean (i.e. first order moment): $\langle|\Delta T(\Delta\theta)|\rangle \propto \Delta\theta^{H_s}$. Whereas in the macroweather regime the temporal RMS and mean fluctuation exponents are nearly equal $\left(\left\langle \Delta T(\Delta t)^2 \right\rangle^{1/2} \propto \langle \Delta T(\Delta t)\rangle \propto \Delta t^H\right.$; low intermittency, $K(q)=0$; see the discussion after Eq. 9)—the spatial fluctuations are on the contrary highly intermittent (see e.g. section 10.3.1 of Lovejoy and Schertzer (2013) so that the quasi Gaussian approximation no longer holds. In space there is an intermittency correction $\xi(2)/2-\xi(1) = \xi(2)/2\text{-}H_\sigma \approx -0.1$ so that $\left\langle \Delta T(\Delta\theta)^2 \right\rangle^{1/2} \propto \langle \Delta T(\Delta\theta)\rangle^{-0.1} \propto \Delta\theta^{H_x-0.1}$; the graphical estimate in Fig. 4 ($\xi(2)/2 \approx -0.2$) thus implies $H_x \approx -0.1$. Since $H_x < 0$, both the mean—and the RMS fluctuations—decrease with scale Δθ. (the spatial subscript "x" is used since we presume that the zonal angular separation Δθ is approximately ~~equal~~ to the great circle distance Δx).

Also shown in Fig. 4 is the effect of increasing the temporal averaging, systematically doubling it from 1 month to 1024 months (≈85 years). The temporal fluctuations have $H<0$, so that the temporal fluctuation is simply the anomaly at that scale (equal to the temporal average) so that Fig. 4 effectively represents the joint space–time RMS fluctuations $S_{x,t}(\Delta\theta, \Delta t)$. In ch. 10 of (Lovejoy and Schertzer 2013; Lovejoy and de Lima 2015) it is argued on both theoretical and empirical grounds (monthly temperatures from the 20CR) that to a good approximation, the space–time statistics factorize. For the second order statistics, this implies:

$$S_{x,t}(\Delta\theta, \Delta t) \propto S_x(\Delta\theta) S_t(\Delta t) \tag{13}$$

Where $S_x(\Delta\theta)$ and $S_t(\Delta t)$ are respectively the space only and time only RMS structure functions (we have temporarily added the subscript "t": elsewhere we continue to denote the time only RMS structure function simply by $S(\Delta t)$). Since $\log S_{x,t}(\Delta\theta, \Delta t) \approx Const. + \log S_x(\Delta\theta) + \log S_t(\Delta t)$, plot of log Δθ versus $\log S_{x,t}(\Delta\theta, \Delta t)$, factorization implies that for various time resolutions Δt, the curves for $\log S_{x,t}(\Delta\theta, \Delta t)$ are simply displaced downwards by $\log S_t(\Delta t)$. We can see that this is relatively well confirmed in Fig. 4. In addition, due to the temporal macroweather scaling (Fig. 3 for the global series up to about ≈10–20 years), we expect $S_t(\Delta t)$ also to be a power law so that the in macroweather regime, the curves will be roughly equally spaced as the averaging time Δt is doubled. From the figure, we find (up to Δt≈256 months, i.e. ≈ 20 years and from ≈20° to 180° longitude):

$$S_{\theta,t}(\Delta\theta, \Delta t) \propto \Delta\theta^{H_x} \Delta t^{H_t};\ H_x \approx -0.2;\ H_t \approx -0.3 \tag{14}$$

i.e. we have factorization *and* space–time scaling. Similar space–time factorization (but with different exponents) was found to hold in historical precipitation data (Lovejoy and de Lima 2015).

In order to understand the physical meaning of space–time factorization, recall that in the weather regime the appropriately nondimensionalized structure function has a form (very roughly): $S_{s,t}(\Delta\theta, \Delta t) \propto \left(\Delta\theta^2 + \Delta t^2\right)^{s/2}$ (see Pinel et al. 2014 for more precise, general results). This implies that the same amplitude of fluctuation $S_{x,t}$ will typically result from either an instantaneous spatial displacement L (i.e. with space–time lag (L,0)), or from a temporal lag τ at a fixed location (with space–time lag (0, τ)). Mathematically, it implies that there is a size (L)—lifetime (τ) relationship which is the solution of implicit equation $S_{x,t}(L, 0) = S_{x,t}(0, \tau)$; in this nondimensionalized example the relation is: $L = \tau$. In contrast, in the macroweather regime, due to factorization, the corresponding implicit relation between L and τ is $S_x(L)S_t(0) = S_x(0)S_t(\tau)$ whose solution will depend on the spurious small L and small τ behaviours (where for example, the scaling laws break down). To avoid this technical issue (in this case with both H_x and H_t <0), instead of structure functions, we can use autocorrelation functions to obtain new (nondimensional) macroweather space–time relations: $\tau \propto L^{H_x/H_t}$ (Lovejoy et al. 2017).

Notice that the above temporal exponent ($H_t = -0.3$)—which is the exponent of 5° × 5° resolution data—is smaller than the corresponding exponent of the globally averaged series (in Fig. 2 it is $H_t \approx -0.1$, see Sect. 3 for a more accurate estimate). The reason for this apparent discrepancy is that the temporal exponent H_t—while remaining in the range $0 > H_t > -1/2$—varies considerably from region to region with the oceans typically having $H_t \approx -0.1$ whereas land typically has $H_t \approx -0.3$. As we increase spatial averaging from 5° × 5° to global, the higher (ocean) exponents tend to dominate so that for globally averaged temperatures $H_t \approx -0.1$.

The space–time macroweather statistics will be more fully investigated elsewhere, for this paper, the key point is that both the spatial and temporal H's are negative. When $H<0$, then we saw (Eq. 1) that the temperature at resolution τ will scale with exponent H, i.e. as τ^H ($H<0$). Hence if a measured series "m" is not sufficiently averaged or on the contrary, perhaps over-smoothed by interpolation, then it's effective resolution τ_m will be different from the nominal resolution τ_n and $T_{\tau_m}/T_{\tau_n} \approx \lambda_t^{H_t}$ where $\lambda_t = \tau_m/\tau_n$ is the resolution scale ratio. Since the spatial exponent $H_x<0$, the same argument applies in space (resolutions Θ_m, Θ_n) so that overall the statistics of the measured anomalies differ from the true anomalies by the multiplicative factor:

$$T_{\tau_m,\Theta_m}/T_{\tau_n,\Theta_n} \approx \lambda_t^{H_t}\lambda_x^{H_x} = e^{\delta u} \tag{15}$$

we have introduced δu which is a convenient characterization of the overall space–time factor $\lambda_t^{H_t}\lambda_x^{H_x}$. The "δ" is to remind us that δu is due to a difference in the logarithms of the scaling factors. When δu is not too far from zero—as here—we have $e^{\delta u} \approx 1 + \delta u$, below we empirically estimate δu. Note that conventional geostatistical methods such as Kriging assume that at small scales, the fields are smooth—that there are no resolution dependencies. This implies that $\delta u = 0$ and as we see below, it explains their inability to explain the low frequency divergences of the series.

In the precipitation literature, this type of resolution dependent multiplicative factor (when of purely of spatial origin) is called an "areal reduction factor" (for scaling approaches to this, see e.g. (Bendjoudi et al. 1997; Veneziano and Langousis 2005). The analysis in Fig. 4 shows that more generally we may expect analogous "scale reduction" factors to appear when comparing two different anomaly temperature series that have different effective space–time resolutions. Two global time series with different effective resolutions will have statistics that multiplicatively differ over their entire range of scales, this scale reduction factor therefore leads to an overall bias in the statistics.

3 The absolute errors

3.1 Fractional Gaussian Noise (fGn)

The previous section compared the relative errors of six global monthly temperature series. We found that the dominant statistical behavior of the differences between the series δT_{ij} cannot be explained by the usual dichotomy of (short term) error and (long term) bias. In order to understand this and to estimate the absolute measurement errors, we need a model of both the actual temperature and the measurement process. We have cited now numerous studies that show that the temperature is scaling over the macroweather regime (Lovejoy and Schertzer 2013), has argued macroweather temporal intermittency is low and (Lovejoy et al. 2015b) has shown that for macroweather time series, the simplest scaling model; fractional Gaussian noise (fGn) is a reasonable approximation (at least if we ignore the extremes) and that the long range memory implicit in the scaling can be used for forecasting purposes. It may be useful to note that fGn is related by differentiation to the more familiar Fractional Brownian motion (fBm) process.

For our purposes, an fGn process $G_H(t)$ with parameter H, is defined as:

$$G_H(t) = \frac{c'_H}{\Gamma(1/2+H)} \int_{-\infty}^{t} \left(t - t'\right)^{-(1/2-H)} \gamma\left(t'\right) dt'; \; -1 < H < 0 \tag{16}$$

$\gamma(t)$ is a unit Gaussian "δ correlated" white noise with $<\gamma> = 0$ and:

$$\left\langle \gamma(t)\gamma\left(t'\right) \right\rangle = \delta\left(t - t'\right) \tag{17}$$

where "δ" is the Dirac function and Γ is the usual gamma function. The constant c'_H is a constant chosen so as to make the expression for the statistics particularly simple. Details of this and other, useful properties of fGn are briefly summarized in "Appendix A". A longer review of the properties relevant for macroweather modelling and forecasting are given in (Lovejoy et al. 2015b) and full mathematical treatment is available in (Biagini et al. 2008). From Eq. 16, it can be seen that in our range of interest ($-1/2 < H < 0$), G_H is a smoothed white noise; like the Dirac function and $\gamma(t)$ it is a generalized function that is strictly only meaningful when integrated over a finite set.

The properties of fGn needed below are:

1. $G_H(t)$ is statistically stationary.
2. The mean vanishes: $\left\langle G_H^{(s)}(t) \right\rangle = 0$.
3. When $H = -1/2$, the process $G_{-1/2}^{(s)}(t)$ is simply a Gaussian white noise.

Journal : Large 382 | Article No : 3561 | Pages : 18 | MS Code : CLDY-D-16-00656 | Dispatch : 17-3-2017

4. Anomaly fluctuations: $G_{H,\tau}(t) = \frac{1}{\tau}\int_{t-\tau}^{t} G_H(t')dt'$ satisfy: $\left\langle G_{H,\tau}(t)^2 \right\rangle \propto \tau^{2H};\ -1 < H < 0$.
5. It follows that in the small scale limit (τ->0), the variance diverges and H is scaling exponent of the root mean square (RMS) value. This singular small scale behaviour is responsible for the strong power law resolution effects in fGn.
6. Sample functions $G_{H,\tau}(t)$ fluctuate about zero with successive fluctuations tending to cancel each other out.
7. Differences: in the large Δt limit: $\left\langle \left(\Delta G_{H,\tau}(\Delta t)\right)^2_{diff} \right\rangle \propto 2\tau^{2H}\left(1-(H+1)(2H+1)\left(\frac{\Delta t}{\tau}\right)^{2H}\right)$.
8. Haar fluctuations $\left\langle \left(\Delta G_{H,\tau}(\Delta t)\right)^2_{Haar} \right\rangle = \Delta t^{2H};\ \Delta t \geqslant 2\tau$. [illegible] the normalization c'[illegible] Appendix A"), this result is exact.
9. This implies that Haar fluctuations at time scale Δt scale as Δt^{2H} and do not depend on the resolution τ, H is the fluctuation exponent (Eq. 9).
10. In usual treatments,of fGn, the parameter H is the fluctuation exponent of the fBm whose increments are the corresponding fGn. This conventional fGn parameter H is thus one larger and is confined to the range $0 \le H \le 1$. Here, we define H more generally as the fluctuation exponent (Eq. 9), this allows the definition to also be valid for nonGaussian, intermittent multifractal processes.

3.2 ~~3.2~~ Modelling the earth's temperature

Having defined the basic statistically stationary scaling process (fGn), we need only add a nonstationary process to represent the anthropogenic warming. In Lovejoy (2014) it was shown that anthropogenic effects were roughly linear in the CO_2 radiative forcing ($\log CO_2$) rather than linear in time. The theoretical justification was that—due to economic activity—CO_2 concentration is a reasonable proxy for all the anthropogenic effects. It would thus be better to model the anthropogenic part as a contribution linear in $\log CO_2$—i.e. to replace the time axis by $\log CO_2$. However for simplicity, here we will use a term linear in time:

$$T(t) = \sigma_T G_H(t) + At \tag{18}$$

where t is the time in units of months and σ_T is the RMS Haar month to month fluctuation, G_H is an fGn process and A is a linear approximation to the anthropogenic trend. With this model, the temperature fluctuates about the mean $\langle T(t)\rangle = At$. However, as analyzed and underlined in Lovejoy et al. (2016), even though on (ensemble) average, fGn ~~in~~ trendless, on each realization, it displays a random trend that will contribute some uncertainty to estimates of global warming.

Using Eqs. 17, 18, the Haar structure function of the model earth temperature yields:

$$S^2(\Delta t) = \left\langle \Delta T(\Delta t)^2 \right\rangle = \sigma_T^2 \Delta t^{2H} + A^2 \Delta t^2 \tag{19}$$

(we have used property 8 of Sect. 3.1 and the fact that the Haar fluctuation of the function At is $A\Delta t$). From the empirical structure functions (Figs. 2, 3) if we regress $S(\Delta t)$ between 8 months and 12 years (this avoids the low frequency part dominated by the anthropogenic contribution), we get the H estimate:

$$H = -0.090 \pm 0.042 \tag{20}$$

Taking $H = -0.1$ and fitting the other parameters, we obtain:

$$A = (5.83 \pm 0.073) \times 0^{-4}\ \mathrm{K/month};\ \sigma_T = 0.142 \pm 0.01\ \mathrm{K} \tag{21}$$

Where the uncertainty estimates come from the six different series. This value of A corresponds to 0.700 ± 0.009 K/century. With these parameters, in the model (Eq. 18), we made the simulation in Fig. 5.

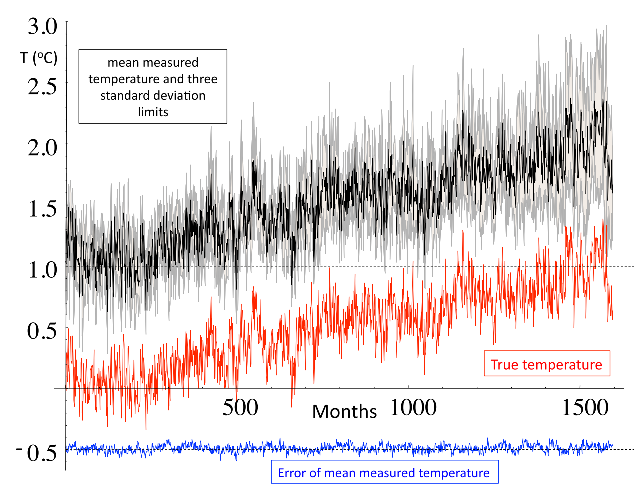


Fig. 5 *Red* is "true earth" (model) temperature using Eqs. 18 the parameters of Eqs. 20, 21. *Black* is the mean of six simulations of the measurement process (Sect. 3.4) with 3 standard deviation spreads (*gray*) and shifted one unit upwards. *Blue* is the difference between the mean measured temperature and the true temperature (displaced 0.5 downward)

3.3 Modelling the measurement errors and biases

The usual approach to temperature measurement uncertainties is to consider measurement errors that are essentially white noises i.e. $G_{-1/2}(t)$, (i.e. $H = -1/2$). This includes those with short range (exponential) decorrelations such as Auto Regressive (AR) processes and their kin. The latter are essentially white noises for scales larger than their decorrelation distances/times. In addition, from the discussion in Sect. 2.4, due to the scale reduction factors, we expect there to be multiplicative biases $e^{\delta u}$ effective over the entire range of time scales. Since these are close to unity, $e^{\delta u} \approx 1 + \delta u$. Although δu does depend on how missing data is dealt with, it does not exhaust the effects of sparse measurements. Recall that over the period 1880-present, at $5° \times 5°$ resolution there are typically >50% missing data and different series have different degrees of missing data, this is an important additional effect. Since (roughly) the space–time statistics factor and are scaling (Sect. 2.4), the effect of the missing data is thus to add a third component to the error, one which is expected to be of the same statistical type as the natural variability i.e. to be proportional to an fGn process. These considerations suggest the following measurement model:

$$T_i(t) = T(t)\left(1 + \delta u_i\right) + \sigma_T B_i G_H^{(i)}(t) + \sigma_T \varepsilon_i G_{-1/2}^{(i)}(t) \tag{22}$$

Where T_i is measured temperature from the ith global temperature series (here $i = 1, 6$ for the six series discussed in Sect. 2) and $T(t)$ is true global temperature (Eq. 18). The first term on the right is the scale reduction factor, the second term is the missing data term and the third is the short range measurement error term. The latter terms have been nondimensionalized using the typical monthly (Haar) variance σ_T (Eq. 18) and the nondimensional amplitudes of these noises are denoted B_i, ε_i respectively.

Taking $T(t)$ as the earth model (Eq. 18), we obtain:

$$T_i(t) = \sigma_T\left(1 + \delta u_i\right)G_H^{(0)}(t) + A\left(1 + \delta u_i\right)t + \sigma_T B_i G_H^{(i)}(t) + \sigma_T \varepsilon_i G_{-1/2}^{(i)}(t) \tag{23}$$

The $G_H^{(0)}$ is the realization of the fGn that determined the true temperature of the earth (Eq. 18); in the following we use the empirical estimate (Eq. 20) $H = -0.1$ throughout. Since $\langle G_H \rangle = 0$, $T_i(t)$ fluctuates around a line with slope $A\left(1 + \delta u_i\right)$.

In order to statistically test the full model (i.e. the model of the earth temperature plus measurement errors; Eqs. 22, 23) we only need the statistical distribution of the parameters δu_i, B_i, ε_i. For this, we will make some simplifying assumptions: (a) that each has a Gaussian distribution, mean μ, standard deviation σ, (b) that for each individual series, the parameters δu_i, B_i, ε_i are statistically independent of each other. In the development below, there is a further independence assumption, that for pairs of different series *i, j*, these terms are statistically independent of each other. Since the series share much data, this last assumption is clearly not fully justified. However, this really affects our interpretation of the results, we are in fact making a statistical estimate of the *effective* parameters, i.e. the parameters that *would* be needed in order to explain the observations *if the series were indeed independent*.

3.4 Estimating the measurement errors and biases

A simple way to estimate the measurement model parameters δu_i, B_i, ε_i is to consider the temporal (Haar) fluctuation for each series:

$$\Delta T_i(t) = \sigma_T\left(1 + \delta u_i\right)\Delta G_H^{(0)}(\Delta t) + A\left(1 + \delta u_i\right)\Delta t + \sigma_T B_i \Delta G_H^{(i)}(\Delta t) + \sigma_T \varepsilon_i \Delta G_{-1/2}^{(i)}(\Delta t) \tag{24}$$

In the following, we attempt to estimate the statistics of δu_i, B_i, ε_i from structure functions estimated from the time intervals from single series rather than ensemble (statistical) averaging. To make this distinction clear for time averaging we use the overbar "–". For example, the time averaged (squared) fluctuation (structure functions) are thus:

$$\begin{aligned} S_i^2(\Delta t) \approx \overline{\Delta T_i(\Delta t)^2} &= S^2(\Delta t) + \delta u_i^2 S^2(\Delta t) + \sigma_T^2 B_i^2 \Delta t^{2H} + \sigma_T^2 \varepsilon_i^2 \Delta t^{-1} \\ &= \sigma_T^2 \varepsilon_i^2 \Delta t^{-1} + \sigma_T^2\left(1 + \delta u_i^2 + B_i^2\right)\Delta t^{2H} + A^2\left(1 + \delta u_i^2\right)\Delta t^2 \end{aligned} \tag{25}$$

(the cross terms disappear because of the independence assumption). The "≈" is used because we estimated the ensemble average from the temporal averages on the individual series so that for example, $\left(\overline{\Delta G_H(\Delta t)^2}\right)^{1/2} = \Delta t^{2H}$ (see property 8, Sect. 3.1). Equation 25 shows that there are three zones: a high frequency classical error measurement term, $\sigma_T^2\varepsilon_i^2\Delta t^{-1}$ a medium frequency missing data and scale reduction term $\sigma_T^2\left(1 + \delta u_i^2 + B_i^2\right)\Delta t^{2H}$, and a low frequency scale reduction term $A^2\left(1 + \delta u_i^2\right)\Delta t^2$. In "Appendix B", we show how the measurement model parameters can be estimated from their structures functions and the structures functions of the pairwise series differences (as in Figs. 2, 3). The results are that δu, B, ε are Gaussian random variables with estimated means and standard deviations (μ, σ):

$$\begin{aligned} \mu_{\delta u} &= 0.114;\ \sigma_{\delta u} = 0.077 \\ \mu_B &= 0.347;\ \sigma_B = 0.175 \\ \mu_\varepsilon &= 0.132;\ \sigma_\varepsilon = 0.062 \end{aligned} \tag{26}$$

Since the different random variables are somewhat correlated, using the above equation yields the "effective" values needed for the simulations below. For completeness, recall that we have already estimated $H = -0.1$, $A = (5.83 \pm 0.073) \times 10^{-4}$ K/month and $\sigma_T = 0.142 \pm 0.01$ K (Eqs. 20, 21).

In order to judge the implications, we can determine, the contribution of each of the three effects.

3.4.1 The scale reduction bias

This term is:

$$\left\langle \Delta T(\Delta t)^2_{red} \right\rangle^{1/2} = \left(\sigma_T^2 \mu_{\delta u^2} \Delta t^{2H} + A^2 \mu_{\delta u^2} \Delta t^2\right)^{1/2} \quad (27)$$

From Eqs. 26, 27, we have: $\left\langle \Delta T(\Delta t = 1\ \mathrm{month})^2_{\mathrm{red}} \right\rangle^{1/2} = 0.020K$ (i.e. ± 0.01 K) (where Δt are in units of months). Conversely, at the longest scales (133 years), we find $\left\langle \Delta T(\Delta t = 133\ \mathrm{yrs})^2_{\mathrm{red}} \right\rangle^{1/2} = 0.134K$ (± 0.067 K). In terms of the true earth temperature, from Eq. 22 we see that it implies a multiplicative bias of a factor $1 + \mu_{\delta u}$, i.e. $\left(\langle T_i(t)\rangle - T(t)\right)/T(t) = \mu_{\delta u} \approx 11.4\%$ (recall that $T(t)$ is the true model temperature). The series to series variation in δu, is given by $\sigma_{\delta u} = \pm 7.7\%$; it is significant. We can also check that it is plausible that it originates in variations in the effective space–time resolutions. To see this, recall that in Sect. 2.4 we argued that if two series differed in temporal resolution by a factor λ_t and spatial resolution by a factor λ_x, then the overall RMS scale reduction factor between the two would be $e^{\mu_{\delta u}} \approx 1 + \mu_{\delta u} = \lambda_t^{-0.3} \lambda_x^{-0.2}$. Therefore, the mean scale reduction factor $\mu_{\delta u} = 0.114$ could be explained by perfect spatial resolution ($\lambda_x = 1$) but inadequate temporal resolution $\lambda_t \approx 0.7$, by perfect temporal resolution ($\lambda_t = 1$) but inadequate spatial resolution $\lambda_x \approx 0.6$, or by some intermediate combination of imperfect spatial and temporal resolutions. These values correspond to differences in the effective degree of temporal and spatial resolutions and they seem reasonable. This scale reduction factor most strongly affects the scale ranges dominated by anthropogenic effects. This can explain the observation (Fig. 1) that the global series differs most strongly from each other in the recent (post≈1980) which is the period that has the strongest rate of anthropogenic warming.

3.4.2 The bias due to missing data

We have:

$$\left\langle \Delta T(\Delta t)^2_{miss} \right\rangle^{1/2} = \sigma_T \mu_{B^2}^{1/2} \Delta t^H \quad (28)$$

so that at 1 month, $\left\langle \Delta T(\Delta t = 1\ \mathrm{month})^2_{miss} \right\rangle^{1/2} = \pm 0.028K$ whereas at 133 years $\left\langle \Delta T(\Delta t = 133\ yrs)^2_{miss} \right\rangle^{1/2} = \pm 0.013K$ To put this in perspective, ignoring the low frequency anthropogenic term, the small short-range error term, and the scale reduction factor (this is a good approximation for resolutions $\tau \approx \leq 10$ years, see Fig. 6) then the missing data error variance is 15% of the true temperature variance: $\left\langle \left(T_\tau(t) - T_{i,\tau}(t)\right)^2 \right\rangle / \left\langle T_\tau(t)^2 \right\rangle = \mu_{B^2} = 0.15$ (including the scale reduction factor increases this to $\mu_{B^2} + \mu_{\delta u^2} = 0.17$).

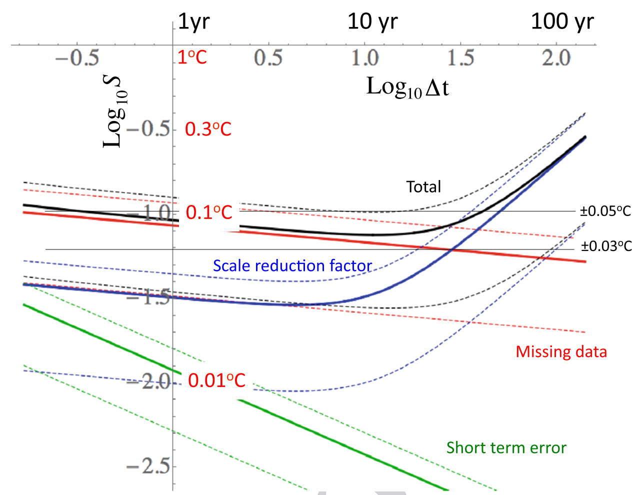


Fig. 6 The structure functions of the various measurement errors with one standard deviation limits shown as *dashed lines* (corresponding the variation from one measurement series to another). The *blue curve* is the contribution of the scale reduction factor, the red is from missing data (slope $= H = -0.1$) and the *green* is the short-range measurement error (slope $-1/2$). The *black curve* is the sum of all the contributions. Notice that most of the contribution to the errors are from the scaling parts. These Haar structure functions have been multiplied by a canonical factor of 2 so that the fluctuations will be closer to the anomalies (when decreasing) or differences (when increasing). Note that these show essentially the difference between the true earth temperature and the measurements; the difference between two different measured series will have double the variances, the difference structure function should thus be increased by a further factor $2^{1/2}$ before comparison with Figs. 2, 3 or the figures below

Using $\sigma_{B^2} = 0.104$ we see that the series to series variation about the 15% mean is about $\pm 10\%$.

3.4.3 The short-term error

We have:

$$\left\langle \Delta T(\Delta t)^2_{error} \right\rangle^{1/2} = \sigma_T \mu_{\varepsilon^2}^{1/2} \Delta t^{-1/2} \quad (29)$$

so that at 1 month we have: $\left\langle \Delta T(\Delta t = 1\ \mathrm{month})^2_{error} \right\rangle^{1/2} = \pm 0.010K$ whereas for 133 years, it is: $\left\langle \Delta T(\Delta t = 133\ \mathrm{yrs})^2_{error} \right\rangle^{1/2} = \pm 0.0003K$. The total variance of the biases and errors is the sum of the three so that $\left\langle \Delta T(\Delta t = 1\ \mathrm{month})^2_{all} \right\rangle^{1/2} = \pm 0.032K$ and $\left\langle \Delta T(\Delta t = 133\ \mathrm{yrs})^2_{all} \right\rangle^{1/2} = \pm 0.068K$. The latter provides a good estimate of the centennial scale temperature errors relevant for evaluating the amplitude of the industrial epoch warming. Converting this to 90% certainty limits (≈1.6 standard deviations) we can say that with 90% certainty, for a given series, that the temperature change since 1880 is correct to within ±0.108 °C.

It is useful to graphically assess the esult by comparing the individual terms that contribute to the error and bias at each scale Δt; this is shown in Fig. 6. Starting with the short term error, we see that the smallest temporal resolution, it is roughly equal to the scale reduction factor but becomes quickly negligible at longer times. Until 10–20 years when the anthropogenic contribution becomes important, the errors are dominated by the missing data term, after that, by the scale reduction term. We can see that the total error is mostly in the range ± 0.03 to $\pm 0.05\,°\mathrm{C}$, although it is a little higher at centennial scales. In the next subsection, we make stochastic simulations of the series and further evaluate the realism of the model.

3.5 Stochastic modelling the measurement process

We can now use the simulated "true" earth temperature (Fig. 5) with these parameters and Eq. 23 to create six simulations of the measured earth series. Figure 6 shows the result when they are presented in the same way as Fig. 1 (i.e. the grey "errors" are actually three standard deviations of the difference of the given series with respect to all the others). Since in this case the true temperature is known, we can also display the true errors (Fig. 7), which show that due to the variable scale reduction factors and variable missing data terms, some series have errors that are significantly different from the others. Figure 5 also shows the errors when the mean of the six simulations is used as the overall temperature estimate. From these simulations we can deduce some fairly simple statistics; for example at monthly resolutions, the RMS difference between the measured series and the truth is $\pm(0.057 \pm 0.025)$ °C so that we can say that the series are "typically" in error by this amount (this is also roughly the amplitude of the error curve with respect to the mean of the series shown at the bottom of Fig. 5). Also, the difference in the mean of each series with respect to the true mean (the bias in the temporal means) is: $0.0087 \pm 0.040\,°\mathrm{C}$ and the corresponding bias with respect to the mean of the six is: $0 \pm 0.020\,°\mathrm{C}$ (Fig. 8).

This means that if we choose a series at random, then there is 90% chance (1.6 standard deviations) that its bias is in the range -0.056 to $0.073\,°\mathrm{C}$ and that it monthly RMS variation about its biased mean is in the range 0.017 to 0.082 °C. If we want to determine the absolute earth temperature, we ~~can~~ now choose the 20CR (the others only give anomalies). The preceding statistics indicate that for a given month its temperature will be in error by $0.010 \pm 0.074\,°\mathrm{C}$ (one standard deviation) so that with 90% certainty, the true monthly and globally averaged temperature is the range -0.109 to $0.127\,°\mathrm{C}$ of the 20CR absolute temperature value for that month.

In order to test the model, we can use it to reconstruct the various structure function statistics discussed in Figs. 2, 3: the mean structure function $\left\langle \Delta T(\Delta t)^2 \right\rangle^{1/2}$, the mean difference structure function with respect to the mean $\left\langle \Delta\delta\overline{T}(\Delta t)^2 \right\rangle^{1/2}$, the mean differences between pairs $\left\langle \Delta\delta T(\Delta t)^2 \right\rangle^{1/2}$ and the standard deviation of the difference of the individual structure functions with respect to the mean of the others ($\sigma_S(\Delta t) = \left\langle (S(\Delta t) - \langle S(\Delta t) \rangle)^2 \right\rangle^{1/2}$). The results are shown in Fig. 9; we can see that it well reproduces the empirical curves (Fig. 2); these are superposed for ease of comparison. Note that since the simulated series are analyzed in exactly the same way as the measurement series, ~~that~~ all nontrivial sampling and

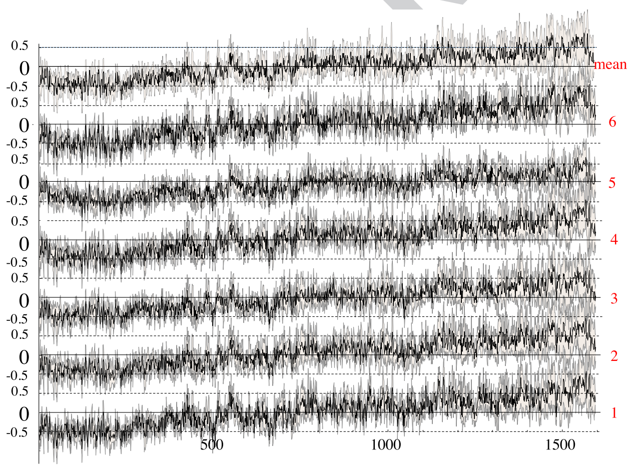


Fig. 7 The six simulated earth temperature measurement series are shown using the same presentation as for the data in Fig. 1 i.e. with the *grey* indicating the three standard deviation limits of the excluded series. The top is the mean of all and the three standard deviation spread is the is due to spread of all the others

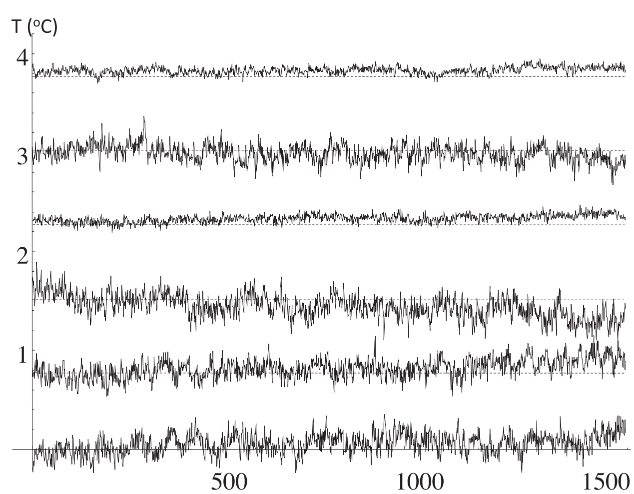


Fig. 8 The absolute errors of the simulated measurement process, with each curve separated by 0.75 K for clarity. Perhaps the most obvious difference between the series is due to their differing scale reduction factors, these factors amplify all the errors by a given factor $1 + \delta u$

Journal : Large 382 | Article No : 3561 | Pages : 18 | MS Code : CLDY-D-16-00656 | Dispatch : 17-3-2017

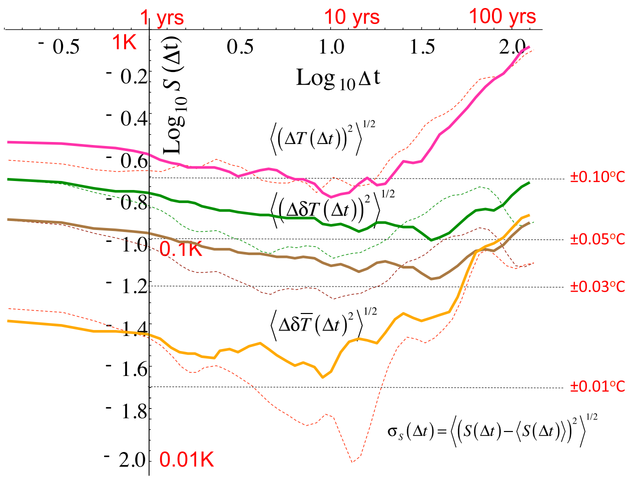


Fig. 9 The *dashed curves* are the empirical curves reproduced from Fig. 2, the thick curves are the corresponding simulated curves using the simulations from Fig. 5

analysis issues are accounted for in the simulations so that the simulation—data agreement is highly significant.

Another way of evaluating these effects is shown in Fig. 10. This displays the same series of structure functions and structure functions of differences that were shown in Fig. 9, except that we systematically remove one of the terms so as to gauge its effect on the statistics. The upper right graph shows that although the short range error term is small, that it nevertheless gives a noticeable contribution especially to the differences $\left\langle \Delta\delta T(\Delta t)^2\right\rangle^{1/2}, \left\langle \Delta\delta\bar{T}(\Delta t)^2\right\rangle^{1/2}$ (green and brown respectively). With no missing data (bottom left), the difference curves are (unrealistically) very close to each other. Finally (lower right), we see that the scale reduction factor is essential for explaining the statistics at long Δt. Rather than displaying simply the means of the six simulations, we can also shows the statistics of the individual realizations that were used in calculating the means (Fig. 11); we see that the series to series variability is fairly realistic (c.f. Fig. 3).

4 Conclusions

Accurate global scale temperature estimates are important in many applications, especially global warming. Deviations of estimated global scale surface temperatures from the true global mean (i.e. errors plus biases) arise not only from human induced inhomogeneities but also because of objective difficulties in determining (spatial) temperature fields from point-like station values. The difficulties are fundamental since the temperature field has nonclassical space–time statistical behaviours (especially scaling and intermittency), and the measuring networks are also sparse (fractal) in both time and in space (they have "holes" at all scales). Rather than attempting to directly quantify the uncertainty with the help of classical statistical assumptions and models, we therefore exploited the fact that a

Fig. 10 The various contribution to $\left\langle \Delta T(\Delta t)^2\right\rangle^{1/2}$ (*pink, top*), $\left\langle \Delta\delta T(\Delta t)^2\right\rangle^{1/2}$ (*green*, 2nd from *top*), $\left\langle \Delta\delta\bar{T}(\Delta t)^2\right\rangle^{1/2}$ (*brown*, third from *top*) and σ_S (*orange, bottom*) with statistics averaged over the six simulated series ($\left\langle \Delta T(\Delta t)^2\right\rangle^{1/2}$, $\left\langle \Delta\delta \ldots^2\right\rangle^{1/2}$, σ_S), and pairs of differences ($\left\langle \Delta\delta T(\Delta t)^2\right\rangle^{1/2}$). The *upper left* graph shows the result with all three error terms present, the *upper right* when the short term error is removed (=0), *lower left* when the missing data term is removed ($B=0$) and *lower right* after the scale reduction factor is removed ($\delta u=0$)

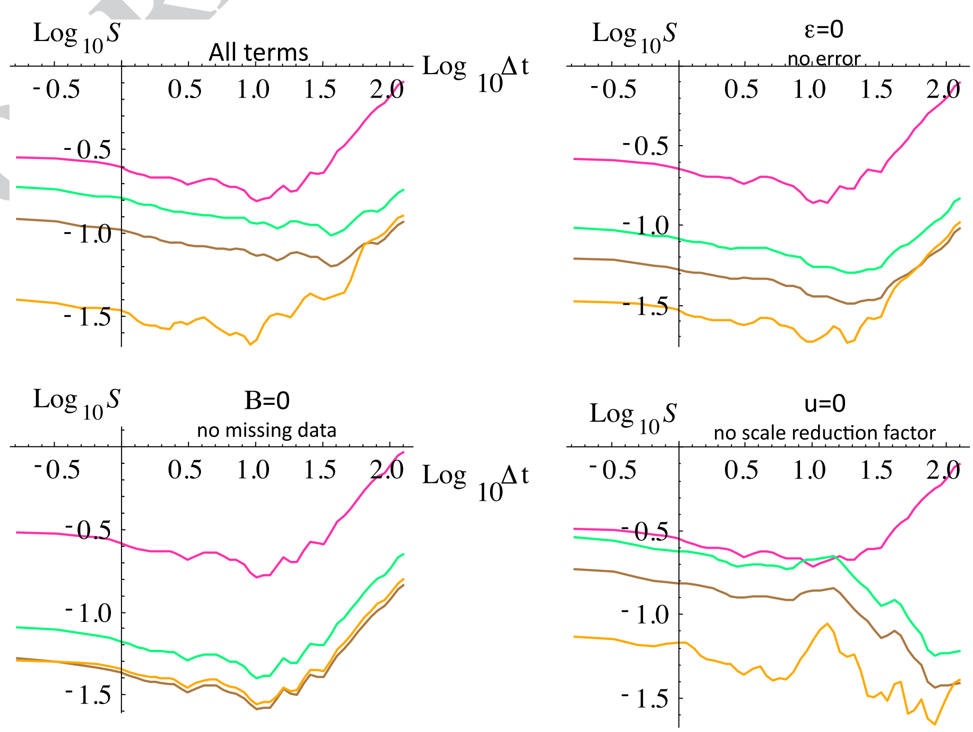
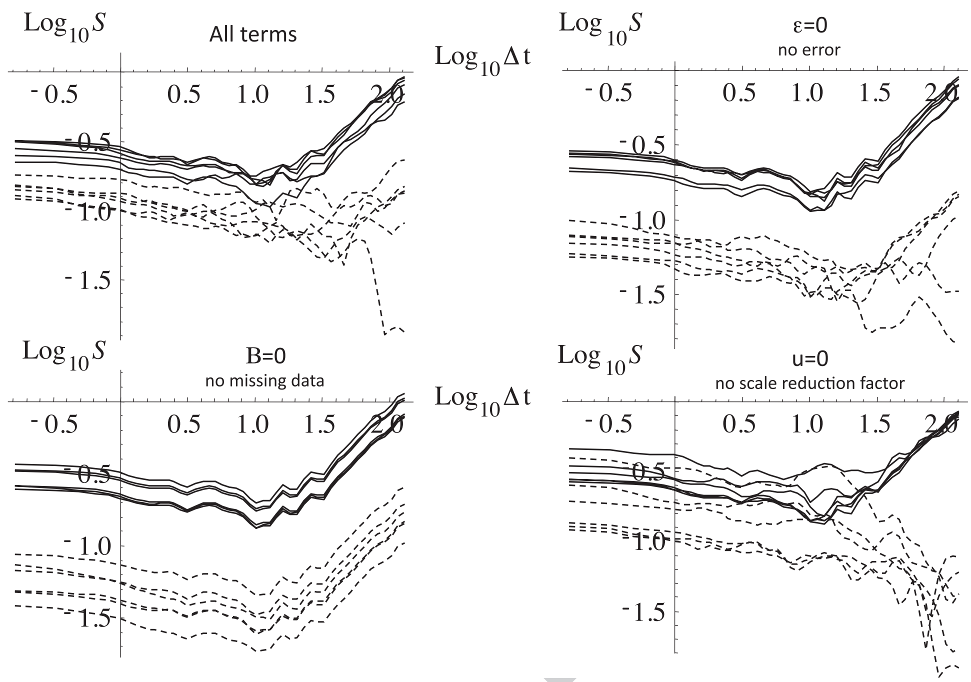


Fig. 11 Similar to Fig. 10 for $\left\langle \Delta T(\Delta t)^2 \right\rangle^{1/2}$ and $\left\langle \Delta\delta\overline{T}(\Delta t)^2 \right\rangle^{1/2}$ except that the results for each of the six simulated measurement terms are shown separately. The structure functions $\left\langle \Delta T(\Delta t)^2 \right\rangle^{1/2}$ (*thick, top*), and differences with respect to the mean $\left\langle \Delta\delta\overline{T}(\Delta t)^2 \right\rangle^{1/2}$ (*bottom, dashed*) for each of the six individual realizations used shown in Fig. 6 and used in Figs. 9, 10. Compare this to Fig. 3 for the data



half dozen or more series have been produced, each using somewhat different data and methodologies. Before making specific assumptions about the errors and biases in the data and attempting to directly quantify them with respect to the real world, we first ask (Sect. 2) how well do different approaches agree with each other as functions of time scale (what are the relative errors)?

In order to isolate the deviations at different time scales we estimated fluctuations and determined their average root mean square values from two months to 133 years (from 1880 to 2012). Perhaps the most obvious conclusion was that although each series was quite similar to the others—and this includes one that was based on only monthly SST and surface pressure observations (the 20CR)—that *even at long time scales differences between the series did not converge*. This is surprising since classical theory shows that for short range correlated errors (e.g. AR(1) processes or kindred processes that are essentially Gaussian white noises at long enough time scales) their RMS differences diminish as $\Delta t^{-1/2}$. Instead of this, from months to centennial scales, the RMS fluctuations stayed nearly constant, mostly between $\approx\pm0.03\,°C$ and $\pm0.05\,°C$ (one standard deviations); they slightly increased at long times, Figs. 2, 3. Since the variability at scales $>\approx10$ years is dominated by ~~the~~ anthropogenic ~~forcing~~, this is a direct estimate of the accuracy with which the latter can be estimated. Also significant is the finding that the *statistics* of the fluctuations can be estimated with much higher relative accuracy (e.g. between 3 and 10 years to better than $\pm0.0005\,°C$).

The fact that the differences between the series have nearly constant deviations—independent of the time scale—demonstrates the existence long-range statistical dependencies in the series errors and biases that are outside conventional geostatistical uncertainty assumptions requires the development of new methodologies.

In order to go beyond relative errors (Sect. 2), so as to estimate absolute errors (Sect. 3), we need models of both the earth's true temperature and of the measurement process itself. For the former, we assumed a combination of natural variability modelled by a scaling, fractional Gaussian noise (fGn) process combined with a linear trend representing the anthropogenic warming. While the former is the simplest scaling model (it is nonintermittent), the latter is an approximation to an anthropogenic contribution (in reality, the latter is much more linear as a function of the CO_2 radiative forcing than as a function of time).

For the measurement errors, although we included a classical short range error term, in order to account for the dominant high and low frequency errors, we need two new sources of error: we introduced both missing data and scale reduction factors. The error due to missing data must have the same type of temporal statistics as the nonmissing data, so that it was also modelled as an fGn process. However, as fGn processes are averaged to lower and lower resolutions, their amplitudes diminish (this affects all the frequencies) so that by itself, missing data is not sufficient for explaining the low frequency errors. For the latter, we relied on the observation (Sect. 2.4) that the temperature anomalies are highly sensitive to their space–time resolutions: in both space and in time, fluctuations systematically decrease in amplitude with increasing scale (in roughly scaling, power law manners). This means that if a series is insufficiently averaged—in space and/or in time—then its effective resolution will be different from the nominal resolution (here, one month, globally averaged). This scale/resolution effect is multiplicative so that it affects all frequencies. Following the hydrology literature's analogous "areal reduction factor" (due to spatial resolution effects), this more general (space–time) effect is a "scale reduction factor".

In order to test the model we need to estimate its parameters; two for the earth model (the amplitude of the natural variability and the anthropogenic trend), and three for the measurement process: ε, B, δu (the amplitudes of the short term error, the missing data and the scale reduction factor). Since the measurement process is stochastic with each series characterized by a different triplet of amplitudes we only need their statistics (assumed to be Gaussian, we need their means and standard deviations). We showed how to make robust parameter estimates using structure function analyses of the $6\times5/2=15$ pairs of series differences. We found for example that the measurement error was about ±0.01 K at one month decreasing rapidly for longer times. That the missing data term was dominant and contributed about 15% to the variance of the temperature at all resolutions up to about 10–20 years (the series to series variability is about 10% around this mean value). Beyond this, ($\Delta t\approx>$10–20 years) the scale reduction factor was dominant, so that temperature anomalies (due to inadequate space–time averaging) were on average about 11% too large with a series to series variability of about 8% around this value.

Finally, using the estimated parameters, we made stochastic simulations of both the "true" earth temperature and the measurement process (including all the sampling issues in the statistical analysis) and showed that all the fluctuation statistics as functions of time—including the pairwise difference fluctuations—were very close to the observations so that the model quantitatively accounts for all the differences between the series and all sampling issues. We thus have confidence that we have an accurate estimate of the absolute temperature errors, and—as for the relative errors—these are generally in the range ±0.03 to ±0.05 K over almost all the range of time scales (month to 133 years). More precisely, at monthly scales, we found that for a given month and series, its temperature will be in error by 0.010 ± 0.074 °C (one standard deviation) so that with 90% certainty, the true monthly and globally averaged temperature is the range −0.109 to 0.127 °C of the temperature value for that month. At centennial scales, we estimated that with 90% certainty, that the corresponding temperature change since 1880 is correct to within ±0.108 °C (i.e. about 10% of the industrial epoch warming).

In order to give a satisfactory estimate of the accuracy of global temperatures, we showed that a new approach was needed and we suggested a simple stochastic temperature and measurement model based on the observed scaling of global temperatures. This approach can readily be extended in a number of directions for quantifying measurement uncertainties. For example, for the temperature, it could be extended to varying spatial resolutions, indeed the relative accuracy method—using pairwise series differences but at 5°x5° resolution—has already been applied to global precipitation (de Lima and Lovejoy 2015). In future it may also be applied to determining the accuracy of pre-industrial multiproxies.

Acknowledgements The author thanks R. Hébert, L. del Rio Amador and David Clarke for useful discussions. This work was unfunded, there were no conflicts of interest. The data were downloaded from the publically accessible sites to be found in the corresponding references (first paragraph, Sect. 2).

Appendix A: some useful properties of fractional Gaussian noise

In this appendix, we give a brief summary of some useful properties of fGn; a longer review is given in (Lovejoy et al. 2015b) and a full mathematical exposé in (Biagini et al. 2008). The standard ("s") fGn process $G_H^{(s)}(t)$ with parameter H, can be defined as:

$$G_H^{(s)}(t)=\frac{c_H}{\Gamma(1/2+H)}\int_{-\infty}^{t}\left(t-t'\right)^{-(1/2-H)}\gamma\left(t'\right)dt';\ -1<H<0 \quad (30)$$

$\gamma(t)$ is a unit Gaussian "δ correlated" white noise with $<\gamma>=0$ and:

$$\left\langle\gamma(t)\gamma\left(t'\right)\right\rangle=\delta\left(t-t'\right) \quad (31)$$

where "δ" is the Dirac function. The constant c_H is a constant chosen so as to make the expression for the statistics particularly simple, see below. It may be useful to note that fGn is related by differentiation to the more familiar Fractional Brownian motion (fBm) process. We can see by inspection of Eq. 16 that $G_H^{(s)}(t)$ is statistically stationary and by taking ensemble averages of both sides of Eq. 16 we see that the mean vanishes: $\left\langle G_H^{(s)}(t)\right\rangle=0$. When $H=-1/2$, the process $G_{-1/2}^{(s)}(t)$ is simply a Gaussian white noise.

Now, take the average of G_H over τ; the "τ resolution anomaly fluctuation":

$$G_{H,\tau}^{(s)}(t)=\frac{1}{\tau}\int_{t-\tau}^{t}G_H^{(s)}\left(t'\right)dt' \quad (32)$$

If c_H is now chosen such that:

$$c_H=\left(\frac{\pi}{2cos(\pi H)\Gamma(-2H-2)}\right)^{1/2} \quad (33)$$

Journal : Large 382 | Article No : 3561 | Pages : 18 | MS Code : CLDY-D-16-00656 | Dispatch : 17-3-2017

then we have:

$$\left\langle G_{H,\tau}^{(s)}(t)^2 \right\rangle = \tau^{2H};\ -1 < H < 0 \tag{34}$$

This shows that a fundamental property of fGn is that in the small scale limit ($\tau \geq 0$), the variance diverges and H is scaling exponent of the root mean square (RMS) value. This singular small scale behaviour is responsible for the strong power law resolution effects in fGn. Since $\left\langle G_H^{(s)}(t) \right\rangle = 0$, sample functions $G_{H,\tau}(t)$ fluctuate about zero with successive fluctuations tending to cancel each other out; this is the hallmark of macroweather.

A comment on the parameter H is now in order. In treatments of fBm, it is usual to use the parameter H confined to the unit interval i.e. to characterize the scaling of the increments of fBm. However, fBm (and fGn) are very special scaling processes, and even in low intermittency regimes such as macroweather—they are at best approximate models of reality. Therefore, it is better to define H more generally as the fluctuation exponent (Eq. 9); with this definition H is also useful for more general (multifractal) scaling processes although the common interpretation of H as the "Hurst exponent" is only valid for fBm in the usual fGn literature, the parameter H is the fluctuation exponent of it's integral, fBm, i.e. it is larger by unity that that used here.

Anomalies

An anomaly is the average deviation from the long term average and since $\left\langle G_H^{(s)}(t) \right\rangle = 0$, the anomaly fluctuation over interval Δt is simply G_H at resolution Δt rather than τ:

$$\left(\Delta G_{H,\tau}^{(s)}(\Delta t)\right)_{anom} = \frac{1}{\Delta t}\int_{t-\Delta t}^{t} G_{H,\tau}^{(s)}(t')dt' = \frac{1}{\Delta t}\int_{t-\Delta t}^{t} G_H^{(s)}(t')dt' = G_{H,\Delta t}^{(s)}(t);\ \Delta t > \tau \tag{35}$$

Hence using Eq. 34:

$$\left\langle \left(\Delta G_{H,\tau}^{(s)}(\Delta t)\right)_{anom}^2 \right\rangle = \Delta t^{2H};\ -1 < H < 0 \tag{36}$$

Differences

In the large Δt limit we have:

$$\left\langle \left(\Delta G_{H,\tau}^{(s)}(\Delta t)\right)_{diff}^2 \right\rangle \approx 2\tau^{2H}\left(1 - (H+1)(2H+1)\left(\frac{\Delta t}{\tau}\right)^{2H}\right);\ \Delta t >> \tau \tag{37}$$

Since $H<0$, the differences asymptote to the value $2\tau^{2H}$ (double the variance). Notice that since $H<0$, the differences are not scaling with Δt.

Haar fluctuations

For the Haar fluctuation we obtain:

$$\left\langle \left(\Delta G_{H,\tau}^{(s)}(\Delta t)\right)_{Haar}^2 \right\rangle = 4\Delta t^{2H}\left(2^{-2H} - 1\right);\ \Delta t \geqslant 2\tau \tag{38}$$

this scales as Δt^{2H} and does not depend on the resolution τ (Lovejoy et al. 2015a).

Since we will use Haar fluctuations throughout, it is convenient to define the fGn $G_H(t)$ with a nonstandard normalization replacing the constant c_H in Eq. 30 by c'_H:

$$c'_H = \frac{c_H}{2\sqrt{2^{-2H} - 1}} \tag{39}$$

With this we can define $G_{H,\tau} = \frac{G_{H,\tau}^{(s)}}{2\sqrt{2^{-2H}-1}}$ so that:

$$\left\langle \left(\Delta G_{H,\tau}(\Delta t)\right)_{Haar}^2 \right\rangle = \Delta t^{2H};\ \Delta t \geqslant 2\tau\,. \tag{40}$$

Appendix B: estimating the parameters of the measurement model

In this appendix, we describe how we estimated the statistics of the amplitudes of the measurement series noises (δu, B, ε, for the scale reduction factor, missing data and conventional measurement error respectively).

The idea is to use second order structure functions (Sect. 3), however from structure functions we can only estimate the squared quantities (δu^2, B^2, ε^2). We therefore used an easily verifiable result, valid for a Gaussian random variable x:

$$\mu_x = \pm\left(\mu_{x^2}^2 - \frac{\sigma_{x^2}^2}{2}\right)^{1/4}$$
$$\sigma_x = \left(\mu_{x^2}^2 - \mu_x^2\right)^{1/2} \tag{41}$$

where μ_x, σ_x are respectively the means and standard deviations of x and μ_{x^2}, σ_{x^2} of x^2. Finally, the sign of μ_x is not determined. In the case of B, ε, this is unimportant since they are multiplied by sign symmetric random functions so that without loss of generality we can we take $\mu_B > 0$, $\mu_\varepsilon > 0$, but for δu, there is an ambiguity. However, since presumably the series are insufficiently averaged, we expect $\delta u > 0$ so that below, we use the plus sign.

The error in the squared fluctuation variance at each scale Δt is therefore:

$$
\begin{aligned}
S_i^2(\Delta t) - S^2(\Delta t) &= \delta u_i^2 S^2(\Delta t) + \sigma_T^2 B_i^2 \Delta t^{2H} + \sigma_T^2 \varepsilon_i^2 \Delta t^{-1} \\
&= \sigma_T^2 \varepsilon_i^2 \Delta t^{-1} + \sigma_T^2 \left(\delta u_i^2 + B_i^2\right)\Delta t^{2H} + A^2 \delta u_i^2 \Delta t^2
\end{aligned} \tag{42}
$$

where $S(\Delta t)$ is the ensemble averaged true earth structure function (see Eq. 25). Since at large Δt the Δt^2 term is dominant, regression of this equation against Δt^2 can conveniently be used to estimate $\mu_{\delta u} = 0.114$ and $\sigma_{\delta u} = 0.077$. However the other terms are smaller and to obtain robust estimates it is advantageous to consider the pairwise differences as in Figs. 2, 3. Since there are six series, we have $6\times 5/2=15$ pairs, giving us substantially more statistics with which to estimate the missing data and error amplitudes B_i, ε_i of the ith series (here, the index i runs from 1 to 6). Therefore, consider the differences between the ith and jth series of measurements:

$$
\delta T_{ij}(t) = \sigma_T \delta u_{ij} G_H^{(0)}(t) + A\delta u_{ij} t + \sigma_T B_{ij} G_H^{(ij)}(t) + \sigma_T \varepsilon_{ij} G_{-1/2}^{(ij)}(t) \tag{43}
$$

where $\delta u_{ij}^2 = \delta u_i^2 + \delta u_j^2$ and we have used the mathematical result:

$$
\begin{aligned}
&B_{ij} G_H^{(ij)}(t) \overset{d}{=} B_i G_H^{(i)}(t) - B_j G_H^{(j)}(t);\ B_{ij}^2 = B_i^2 + B_j^2 \\
&\varepsilon_{ij} G_{-1/2}^{(ij)}(t) \overset{d}{=} \varepsilon_i G_{-1/2}^{(i)}(t) - \varepsilon_j G_{-1/2}^{(j)}(t);\ \varepsilon_{ij}^2 = \varepsilon_i^2 + \varepsilon_j^2
\end{aligned} \tag{44}
$$

where "$\overset{d}{=}$" indicates equality in probability distributions (so that $G_H^{(ij)}(t) \overset{d}{=} G_H^{(i)}(t) \overset{d}{=} G_H^{(j)}(t)$). These results follow since sums and differences of independent Gaussian variables are also Gaussian and their variances add.

Therefore the fluctuations in the differences are:

$$
\begin{aligned}
\delta\Delta T_{ij}(\Delta t) =& \sigma_T \delta u_{ij} \Delta G_H^{(0)}(\Delta t) + A\delta u_{ij} \Delta t + \sigma_T B_{ij} \Delta G_H^{(ij)}(\Delta t) \\
&+ \sigma_T \varepsilon_{ij} \Delta G_{-1/2}^{(ij)}(\Delta t)
\end{aligned} \tag{45}
$$

With this, squaring and averaging, we obtain for the corresponding squared structure function:

$$
S_{ij}^2(\Delta t) = \overline{\delta\Delta T_{ij}(\Delta t)^2} = \sigma_T^2 \varepsilon_{ij}^2 \Delta t^{-1} + \sigma_T^2\left(\delta u_{ij}^2 + B_{ij}^2\right)\Delta t^{2H} + A^2 \delta u_{ij}^2 \Delta t^2 \tag{46}
$$

We can now estimate the parameters by regression of $S_{ij}^2(\Delta t)$ on the fifteen i, j pairs of difference structure functions against Δt^{-1}, Δt^{2H} (with $H = -0.1$) and Δt^2. To make the problem numerically more robust, we used the fact that the trend A was estimated earlier from regressions on the individual series $T_i(t)$. Similarly, for each of the six $S_i(\Delta t)^2$ functions, we estimated the trends $A^2 \delta u_i^2$; using the estimates for A this leads to estimates of $\mu_{\delta u}$, $\sigma_{\delta u}$, $\delta u_{ij}^2 = \delta u_i^2 + \delta u_j^2$. These trends were then removed to obtain the (quadratically) detrended difference structure function $S_{ij,\mathrm{det}}^2(\Delta t) = \sigma_T^2 \varepsilon_{ij}^2 \Delta t^{-1} + \sigma_T^2\left(\delta u_{ij}^2 + B_{ij}^2\right)\Delta t^{2H}$; when regressed against Δt^{-1}, Δt^{2H}, these gave robust estimates of the prefactors $\sigma_T^2 \varepsilon_{ij}^2$ and $\sigma_T^2\left(\delta u_{ij}^2 + B_{ij}^2\right)$. Combined with the trend based estimates of δu_{ij}^2, we thus obtain 15 estimates for each of the random variables, ε_{ij}^2, B_{ij}^2. If we assume that the parameters are independent identically distributed random variables then Eq. 38 shows that:

$$
\begin{aligned}
B_{ij}^2 &\overset{d}{=} 2B_i^2 \overset{d}{=} 2B_j^2 \\
\varepsilon_{ij}^2 &\overset{d}{=} 2\varepsilon_i^2 \overset{d}{=} 2\varepsilon_j^2
\end{aligned} \tag{47}
$$

Therefore, we use the estimates of ε_{ij}^2, B_{ij}^2 to obtain estimates of the statistics of ε_i^2, B_i^2, and then from Eq. 35, by assuming the variables are Gaussian, we obtain estimates for the means and standard deviations of ε_i, B_i. For completeness, we give the means and standard deviations of δu_i, obtained from $S_i(\Delta t)$ as explained earlier.

$$
\begin{aligned}
\mu_{\delta u} &= 0.114;\ \sigma_{\delta u} = 0.077 \\
\mu_B &= 0.347;\ \sigma_B = 0.175 \\
\mu_\varepsilon &= 0.132;\ \sigma_\varepsilon = 0.062
\end{aligned} \tag{48}
$$

(due to the ambiguity in the sign, we did not take the square root of Eq. 41 to more directly yield B_i, ε_i). Since the different random variables are somewhat correlated, using the above equation yields the "effective" values needed for the simulations below. For completeness, recall that we have already estimated $H = -0.1$, $A = (5.83 \pm 0.073)\times 10^{-4}$ K/month and $\sigma_T = 0.142 \pm 0.01$ K (Eqs. 20, 21).

References

Bendjoudi, H., Hubert, P., Schertzer, D., Lovejoy, S. (1997)~~,~~ Interprétation multifractale des courbes intensité-durée-fréquence des précipitations, Multifractal point of view on rainfall intensity-duration-frequency curves, C.R.~~ A.~~ S., (Sciences de la terre et des planetes/Earth and Planetary Sciences). 325:323–326

Biagini F, Hu Y, Øksendal B, Zhang T (2008) Stochastic Calculus for Fractional Brownian Motion and Applications. Springer-Verlag, London

Brohan P, Kennedy JJ, Harris I, S. F. B. Tett, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. J Geophys Res 111:D12106 doi:10.1029/2005JD006548

Bunde A, Eichner JF, Havlin S, Koscielny-Bunde E, Schellnhuber HJ, Vyushin D (2004) Comment on ''scaling of atmosphere and ocean temperature correlations in observations and climate models''. Phys Rev Lett 92:039801–039801

Compo GP et al (2011) The twentieth century reanalysis project. Quarterly J Roy Meteorol Soc 137:1–28 doi:10.1002/qj.776

Compo GP, Sardeshmukh PD, Whitaker JS, Brohan P, Jones PD, McColl C (2013) Independent confirmation of global land warming without the use of station temperatures. Geophys Res Lett 40:3170–3174 doi:10.1002/grl.50425

Cowtan K, Way RG (2014) Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. Q J R Meteorol Soc 140:1935–1944. doi:10.1002/qj.2297

de Lima MIP, Lovejoy S (2017) Macroweather precipitation variability up to global and centennial scales. Wat Resour Res AQ5

Diamond HJ et al (2013) US climate reference network after one decade of operations: status and assessment. Bull Amer Meteor Soc 94:485–498 doi:10.1175/BAMS-D-12-00170.1

Efstathiou MN, Varotsos CA (2010) On the altitude dependence of the temperature scaling behaviour at the global troposphere. Int J Remote Sens 31(2):343–349

Franzke C (2012) Nonlinear trends, long-range dependence and climate noise properties of temperature. J Clim 25:4172–4183. doi:10.1175/JCLI-D-11-00293.1

Hansen J, Ruedy R, Sato M, Lo K (2010) Global surface temperature change. Rev Geophys 48:RG4004 doi:10.1029/2010RG000345

Hausfather Z, Cowtan K, Clarke DC, Jacobs P, Richardson M, Rohde R (2017) Assessing recent warming using instrumentally-homogeneous sea surface temperature records. Sci Adv

Karl TR, Arguez A, Huang B, Lawrimore JH, McMahon JR, Menne MJ, Peterson TC, Vose RS, Zhang H-M (2015) Possible artifacts of data biases in the recent global surface warming hiatus. Sci Expr 1–4 doi:10.1126/science.aaa5632

Kennedy JJ, Rayner NA, Smith RO, Saunby M, Parker DE (2011) Reassessing biases and other uncertainties in sea-surface temperature observations measured in situ since 1850 part 2: biases and homogenisation. J Geophys Res 116:D14104. doi:10.1029/2010JD015220

Kondratyev KY, Varotsos C (1995) Atmospheric greenhouse effect in the context of global climate change. Il Nuovo Cimento C 18(2):123–151

Lovejoy S (2013) What is climate? EOS 94(1)

Lovejoy S (2014) Scaling fluctuation analysis and statistical hypothesis testing of anthropogenic warming. Clim Dyn 42:2339–2351. doi:10.1007/s00382-014-2128-2

Lovejoy S (2015) A voyage through scales, a missing quadrillion and why the climate is not what you expect. Climate Dyn 44:3187–3210 doi:10.1007/s00382-014-2324-0

Lovejoy S, de Lima MIP (2015) The joint space-time statistics of macroweather precipitation, space-time statistical factorization and macroweather models. Chaos 25:075410. doi:10.1063/1.4927223

Lovejoy S, Schertzer D (1986) Scale invariance, symmetries, fractals and stochastic simulations of atmospheric phenomena. Bulletin of the AMS 67:21–32

Lovejoy S, Schertzer D (2010) Towards a new synthesis for atmospheric dynamics: space-time cascades, Atmos Res. doi:10.1016/j.atmosres.2010.01.004

Lovejoy S, Schertzer D (2012a). Low frequency weather and the emergence of the Climate. In: Sharma AS, Bunde A, Baker DN, Dimri VP (eds) Extreme events and natural hazards: the complexity perspective, AGU monographs, Washington DC, pp. 231–254

Lovejoy S, Schertzer D (2012b) Haar wavelets, fluctuations and structure functions: convenient choices for geophysics. Nonlinear Proc Geophys 19:1–14. doi:10.5194/npg-19-1-2012

Lovejoy S, Schertzer D (2013) The Weather and Climate: Emergent Laws and Multifractal Cascades. Cambridge University Press, Cambridge

Lovejoy S, Schertzer D, Ladoy P (1986) Fractal characterisation of inhomogeneous measuring networks. Nature 319:43–44

Lovejoy S, Scherter D, Varon D (2013a) How scaling fluctuation analyses change our view of the climate and its models (Reply to R. Pielke sr.: Interactive comment on "Do GCM's predict the climate... or macroweather?" by S. Lovejoy et al.). Earth Syst Dynam Discuss 3:C1–C12

Lovejoy S, Schertzer D, Varon D (2013b) Do GCM's predict the climate…. or macroweather? Earth Syst Dynam 4:1–16. doi:10.5194/esd-4-1-2013

Lovejoy S, del Rio Amador L, Hébert R (2015a) The Scaling LInear Macroweather model (SLIM): using scaling to forecast global scale macroweather from months to decades. Earth System Dyn Disc 6:489–545 doi:10.5194/esdd-6-489-2015

Lovejoy S, del Rio Amador L, Hébert R (2015b) The ScaLIng Macroweather Model (SLIMM): using scaling to forecast global-scale macroweather from months to Decades. Earth Syst Dynam 6:1–22. doi:10.5194/esd-6-1-2015

Lovejoy S, del Rio Amador L, Hebert R, de Lima I (2016) Giant natural fluctuation models and anthropogenic warming, Geophys Res Lett. doi:10.1002/2016GL070428

Lovejoy S, Del Rio Amador L, Hébert R (2017) Harnessing butterflies: theory and practice of the Stochastic Seasonal to Interannual Prediction System (StocSIPS). In: Tsonis AA (ed) Nonlinear Advances in Geosciences,. Springer Nature

Mann ME (2011) On long range dependence in global surface temperature series. Clim Change 107:267–276

Mazzarella A, Tranfaglia G (2000) Fractal characterisation of geophysical measuring networks and its implication for an optimal location of additional stations: an application to a rain-gauge network. Theor Appl Climatology 65:157–163 doi:10.1007/s007040070040

Mears CA, Wentz FJ, Thorne PW, Bernie D (2011) Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo estimation technique. J Geophys Res Atmos 116:2156–2202

Nicolis C (1993) Optimizing the global observational network—a dynamical-approach. J Appl Meteor 32:1751–1759

Parker DE (2006) A demonstration that large-scale warming is not urban. J Clim 19:2882–2895 doi:10.1175/JCLI3730.1

Peterson TC (2003) Assessment of urban versus rural in situ surface temperatures in the contiguous United States: No difference found. J Clim 16:2941–2959

Pielke RA et al (2007) Unresolved issues with the assessment of multidecadal global land surface temperature trends. J Geophys Res (Atmos). 112, 2156–2202. doi:10.1029/2006JD008229

Pinel J, Lovejoy S, Schertzer D (2014) The horizontal space-time scaling and cascade structure of the atmosphere and satellite radiances. Atmos Resear 140–141:95–114 doi:10.1016/j.atmosres.2013.11.022

Rohde R, Muller RA, Jacobsen R, Muller E, Perlmutter S, Rosenfeld A, Wurtele J, Groom D, Wickham C (2013) A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011. Geoinfor Geostat: An Overview. doi:10.4172/2327-4581.1000101

Rybski D, Bunde A, Havlin S, von Storch H (2006) Long-term persistance in climate and the detection problem. Geophys Resear Lett 33:L06718-06711-06714 doi:10.1029/2005GL025591

Rypdal K, Østvand L, Rypdal M (2013) Long-range memory in Earth's surface temperature on time scales from months to centuries. JGR Atmos 118:7046–7062 doi:10.1002/jgrd.50399

Smith TM, Reynolds RW, Peterson TC, Lawrimore J (2008) Improvements to NOAA's Historical Merged Land-Ocean Surface Temperature Analysis (1880–2006). J Clim 21:2283–2293

Veneziano D, Langousis A (2005) The areal reduction factor: a multifractal analysis. Water Resour Res. doi:10.1029/2004WR003765

Williams CN, Menne M, Lawrimore JH (2012) NCDC Technical Report No. GHCNM-12-02 Modifications to Pairwise Homogeneity Adjustment software to address coding errors and improve run-time efficiency Rep., NOAA, Washington DC

Journal : Large 382 | Article No : 3561 | Pages : 18 | MS Code : CLDY-D-16-00656 | Dispatch : 17-3-2017