Pre-publication GRL supplement

This is the pre-publication version of the supplementary material to the paper:
Lovejoy, S., L. del Rio Amador, R. Hébert, and I. de Lima, 2016: Giant natural fluctuation models and anthropogenic warming, **Geophys. Res. Lett., 43**, doi:10.1002/2016GL070428.
It focuses on how to win the $100,000 contest.

Supplement

1. Introduction

1.1 Data description

In this supplement we give a more complete analysis and discussion of the models discussed in the main text as well as several contest details.

On Keenan's contest website (http: // www.informath.org/Contest1000.htm), he gives the following explanation of the contest:

> "The file Series1000.txt contains 1000 simulated time series. Each series has length 135: the same length as that of the most commonly studied series of global temperatures (which span 1880-2014). The 1000 series were generated as follows. First, 1000 random series were obtained (for more details, see below). Then, some of those series were randomly selected and had a trend added to them. Each added trend was either 1°C/century or -1°C/century. For

> comparison, a trend of 1°C/century is greater than the trend that is claimed for global temperatures. A prize of $100,000 (one hundred thousand U.S. dollars) will be awarded to the first person who submits an entry that correctly identifies at least 900 series: which series were generated without a trend and which were generated with a trend."

Significant additional details are:

> "During the generation of the 1000 series, ... the initial 1000 random series were obtained via a trendless statistical model, which was fit to a series of global temperatures. The trendless statistical model is preferable to the trending statistical model relied upon by the IPCC, when the models are compared via relative likelihood.) "

Although the contest started on Nov. 18, 2015, there were problems so that within only a few days, Keenan had been forced to retract his original series and reimburse the applicants. The contest was relaunched on Nov. 22 stating: "the prize will be awarded to anyone who can demonstrate, via statistical analysis, that the increase in global temperatures is probably not due to random natural variation". Unfortunately (!), at an unknown later date (but after Nov. 30th; see the Corbett Report posted on November 30 2015 and available at: https://www.youtube.com/watch?v=YFiCKBgA_eQ), this original winning criterion was quietly modified so that the statement is now the quite different: "Anyone who

can demonstrate, via statistical analysis that the increase in global temperatures is probably not due to random natural variation should be able to win the contest".

1.2 Gelman' s contribution

Before taking a closer look at the contest, we browsed the internet and found a post by Andrew Gelman only days after the contest was started (http : // andrewgelman.com/2015/12/09/why - i - decided - not - to - enter - the - 100000 - global - warming - time - series - challenge/). Gelman straightforwardly analyzed Keenan's series and concluded that the requirement of 900 correct trend assignments was about 5 standard deviations above what was possible. Since this corresponds to probabilities of success of less than one in a million, Gelman concluded that it wasn't worth the $10 entry fee, and strongly discouraged further participants.

Before elaborating on several key improvements, let us first review his contribution. Gelman used standard linear regression to obtain slope estimates for each series, he then plotted the histogram (fig. S1).

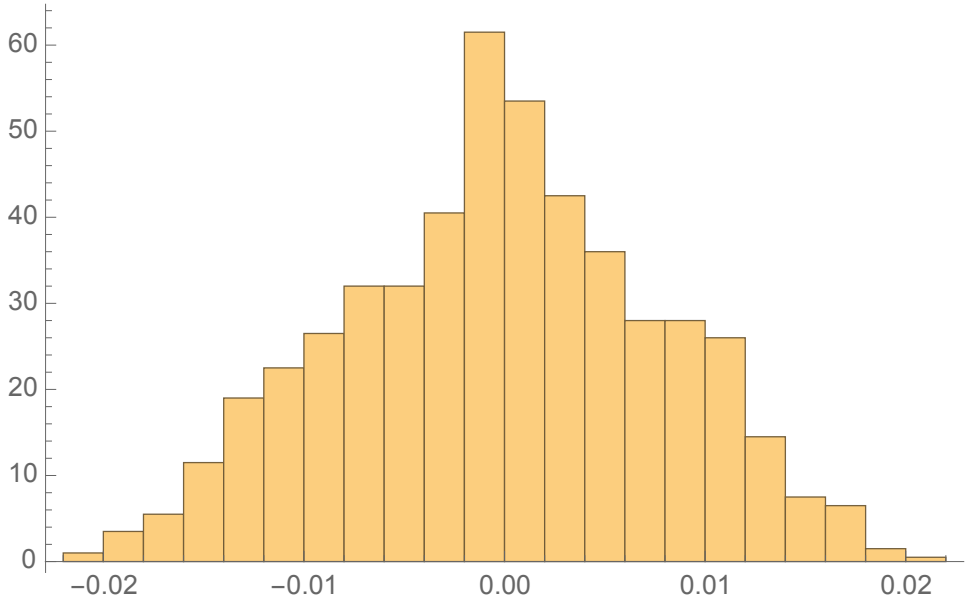

Fig. S1: The histogram (the vertical axis is the probability density function, the horizontal axis is the trend in °C/year) of the regression slopes of $T(t)$ (mean= -0.000218±0.00793). Notice that the ± 0.01 trends that correspond to the imposed 1°C/century trends are nearly invisible.

Gelman reasonably assumed that the resulting distribution was a mixture of Gaussian distributions centred at slopes of -0.01, 0, +0.01 (corresponding to an added slope of -0.01, no added slope and an added slope of +0.01). Note that the spread around these imposed values is due to both real (but random) trends in the (on average only) "basic trendless" process ($T_{init}(t)$) and to an additional spread introduced by the uncertainties inherent in the slope estimation process: here, the regression procedure. Since Keenan's second "pure trend" process with trends of -0.01, 0, or +0.01 °C/yr, is a three-state Bernoulli process for the trends, in the following we refer to it simply as the "Bernoulli process" (see section 2 for a more formal mathematical description).

From the trend histogram in fig. S1, Gelman estimated the parameters of this Bernoulli process. A general mixture of three Gaussians involves 8 independent parameters: three means, three standard deviations and two amplitudes. However, if we use a bias-free trend estimator, then the three means are known (-0.01, 0, 0.01). Also, the spreads about the means due to the regression/trend estimator uncertainties are the same for each. If we finally assume that the process is symmetric with respect to the sign (+, -) we are led to two parameter probability densities of the form:

$$p(x)=\frac{P_0}{A}G\left(\frac{x}{A}\right)+\left(\frac{1-P_0}{2A}\right)\left(G\left(\frac{x+a}{A}\right)+G\left(\frac{x-a}{A}\right)\right);\quad G(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2};\quad a=0.01 \tag{S1}$$

Where the only parameters are P_0, A and $G(x)$ is a "unit" Gaussian (i.e. mean 0, standard deviation = 1). From this, we can easily find the standard deviation:

$$\sigma_{trend}=\left\langle x^2\right\rangle^{1/2}=\left(A^2+a^2\left(1-P_0\right)\right)^{1/2} \tag{S2}$$

which relates the spread in the trends σ_{trend} to the distribution parameters. Although the original problem did not specify that the series were symmetric with respect to the sign of the trends, this is not important since by analyzing the absolute slopes, the distribution becomes symmetrized anyway (furthermore, the statistics of the series are indeed nearly sign symmetric[1]).

[1] For example, there are 495 series that have $T(135) > T(1)$ whereas sign symmetry would lead to 500±16.

Gelman then straightforwardly estimated the number that would be correct if an optimal threshold was used to classify all the trends within certain bounds as untrended and all those outside as trended (see fig. S6 for an illustration of this idea, but using a different trend estimation technique). Using the above form for $p(x)$, and using the maximum likelihood estimator, one obtains P_0 = 0.537 and A = 0.00401. If we now classify the series as trended whenever $|x| > x_c$ where x_c is a critical threshold, then we find that the optimal x_c (see table S1) is 856±10 correct (see fig. S7 for a P_0 - A diagramme[2]).

To put these results into perspective, if the only information that we had was that P_0 = 0.54, then the probability of getting 900 correct can be estimated using the theory of Bernoulli processes which tells us that for large N the result is a Gaussian with theoretical standard deviation $(N\,P_0\,(1-P_0))^{1/2}$ where N = 1000 is the number of Bernoulli "trials" (here, trends). Putting in the numbers we find that 900 correct classification corresponds to a 22.8 standard deviation event i.e. to a probability of $9\text{x}10^{-116}$.

1.3 Estimating the trends by assuming long range correlations in the residuals

Gelman had fallen into Keenan's trap. The standard regressions that he used, assume that the residuals (what's left over after the linear trend is removed) are uncorrelated. Although the IPCC AR5 does obliquely mention the issue of strong

[2] Gelman allowed the -0.01 trends and +0.01 trends to have different weights (rather than the same value $(1-P_0)/2$ as in the equation above and obtained an estimate of 854±10 correct solutions (the nearly 5 standard deviation result alluded to above).

(power law) versus weak (exponentially decorrelated) residuals, in the end they assume that the correlations are indeed weak (they allow for residuals to be related by an Auto Regressive order 1, AR(1) process, but this makes only a minor difference with respect to the assumption of independent residuals) and it is this assumption of weak or absent correlations in the residuals that Keenan was criticizing. However, as pointed out (and exploited notably) in L2014 - as well as in two other papers statistically explaining the "pause" [*Lovejoy*, 2014b], [*Lovejoy*, 2015] - the residuals are in fact scaling (scale invariant) with long-range (power law, not exponential decorrelations, statistical dependencies) so that on the contrary they are highly correlated. If one assumes that the basic process is Gaussian or nearly Gaussian (as it indeed is, see fig. S8) - and this is combined with long range dependencies - then a good model for the process is either fractional Brownian motion (fBm, denoted $B_{H,s}(t)$) or its increments, fractional Gaussian noise (fGn)[3].

---

[3] The fractional Brownian motion (fBm) process $B_{H,s}(t)$ is defined as:

$$B_{H,s}(t)=\frac{c_H(s)}{\Gamma(H+1/2)}\int_{-\infty}^{0}\left((t-t')^{H-1/2}-(-t')^{H-1/2}\right)\gamma(t')dt'+\frac{c_H(s)}{\Gamma(H'+1/2)}\int_{0}^{t}(t-t')^{H-1/2}\gamma(t')dt' \quad \text{(S3)}$$

where $\gamma(t)$ is a "unit" Gaussian white noise process with $\langle\gamma\rangle=0, \langle\gamma^2\rangle=1$ and $0\le H\le 1$ and the constant $c_H(s)$ is chosen so that the covariance is:

$$\left\langle B_{H,s}(t)B_{H,s}(t')\right\rangle = s^2\left(t^{2H}+t'^{2H}+\left|t-t'\right|^{2H}\right)/2 \quad \text{(S4)}$$

where s is the volatility parameter. Note that fBm has the special point at the origin: $B_{H,s}(0) = 0$, it is nonstationary. The fluctuation exponent $H>0$ fluctuations tend to grow with scale so that individual realizations tend to "wander (the classical drunkard's walk – the usual Brownian motion has $H=1/2$). Additionally – and this is important for Keenan's model - realizations have random trends.

In comparison, the increments of $B_{H,s}$ are fractional Gaussian noises (fGn):

$$G_{H',s}(t)\propto s\int_{-\infty}^{t}(t-t')^{H'-1/2}\gamma(t')dt' \quad \text{(S5)}$$

where H' is scaling exponent of the fluctuations of $G_{H,s}$ and $H' = H-1$ so that $-1\le H'\le 0$. From equation S5 it is obvious that fGn is a statistically stationary process and because $H'<0$

That Keenan used fBm is quite plausible since it enables him to make his point about the inadequacies of the usual trend analyses while making the contest unwinnable with conventional regression techniques such as those employed by Gelman and the IPCC. Actually, given that in his papers, Keenan regularly uses Auto Regressive (AR), and the related AR Integrated Moving Average (ARIMA) type models, in reality, he probably used the fractional derivative extensions: Fractional ARIMA (FARIMA) but these are implicitly based on fBm. However, we saw that the increments of Keenan's process are nearly stationary (fig. 2) and fig. S8 shows that they are nearly Gaussian (at least for the first differences at the finest resolution, the increments at the far left in fig. S8). Such processes have statistical properties that are only determined by their spectra - or equivalently by the second order structure function (fig. 1) - so that we can continue to analyse and model the process in terms of an fBm with superposed additional trends. We underline "additional" since each realization of an fBm produces a random trend (even thought the process itself is trendless), and this was surely part of Keenan's idea to exploit this fact. All of this just implies that a good initial model for Keenan's process is fractional Brownian motion (fBm) so that we should use the maximum likelihood trend estimate from an fBm model. Thanks to a handy function built into the *Mathematica* software that we used, this was easy.

---

fluctuations tend to diminish with scale (fig. 2 indicates that $H' \approx -0.1$ for real global temperatures) and fGn has no random trends.

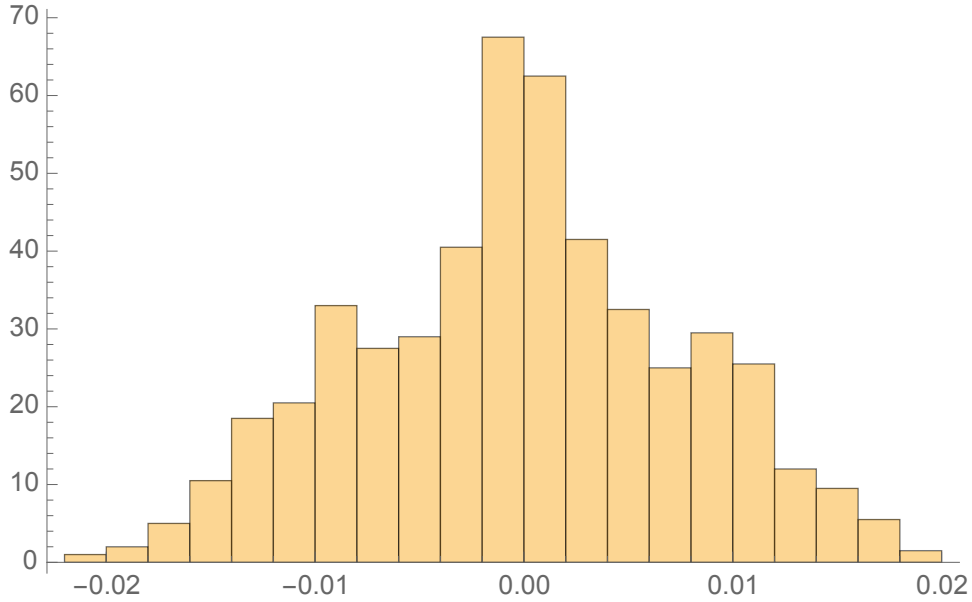

Fig. S2: The histogram (the vertical axis is the probability density function, the horizontal axis is the trend in ºC/year) of the fBM analysed trends for $T(t)$. Notice that the ± 0.01 trends are now noticeable (the mean = -0.00024± 0.007726).

When one estimates trends assuming that the residuals are from such a power law correlated fBm model, then one gets the slightly narrower distribution shown in fig. S2. As can be seen in the figure, the ±0.01 imposed trends are now a bit more visible; the overall standard deviation has shrunk from 0.00793 to 0.00772; figure S3 shows a direct comparison of the regression and fBm trend histograms.

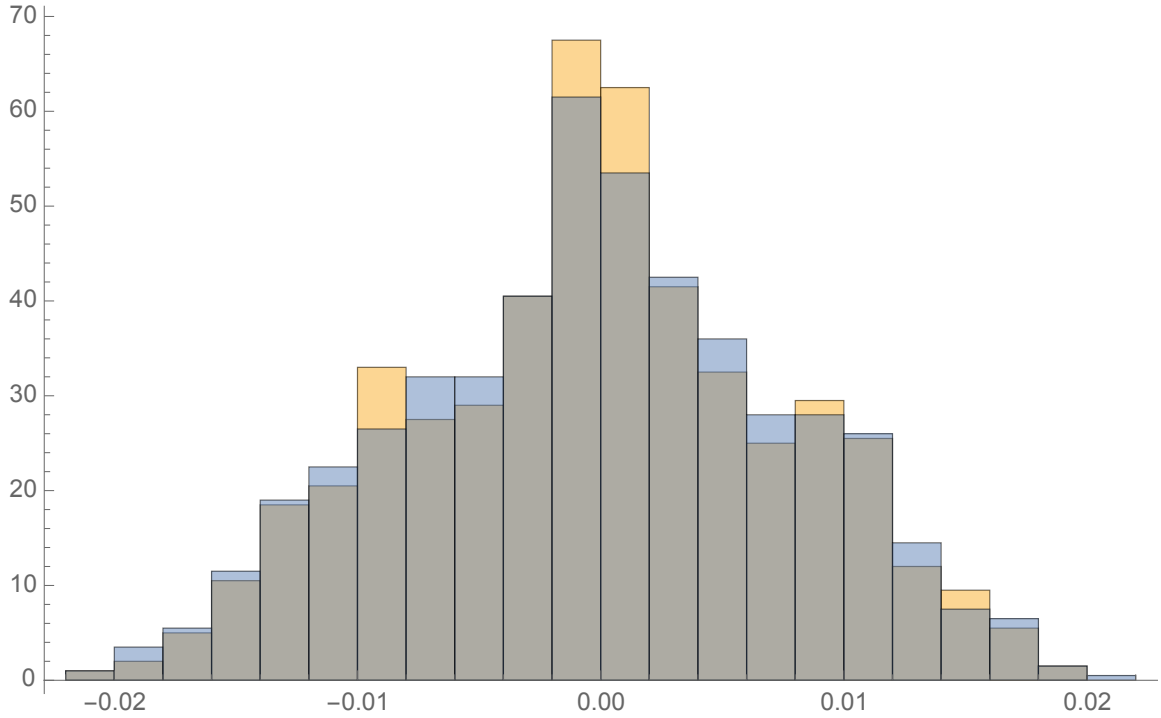

Fig. S3: The histogram (the vertical axis is the probability density function, the horizontal axis is the trend in ºC/year) of the regression trends (gray), yellow are fBm trends. Notice the slight difference: the yellow spikes at ±0.01 and at 0. The underlying distribution is slightly narrower and this makes a big difference.

To show that this difference between the regression and the fBm trends is actually quite significant, we can now repeat Gelman' s analysis (estimating the P_0, A parameters of a 3 Gaussian mixture) on the fBm trends (this is equivalent to assuming long range statistical dependencies in the residuals rather than an absence of correlations in the residuals). Repeating the analysis of the previous subsection but with the fBm trends, we find parameters P_0 = 0.544, A = 0.00361, (shown on the histogram in fig. S6, table S1) and shown as the big red point in the P_0-A diagram (fig. S7) which shows contours of the expected number of correct

classifications as a function of P_0, A. The P_0-A diagram is a useful representation since the P_0, A are all that is required in order to make an optimal classification. From fig. S7, we can see that this small change in parameters from Gelman's standard regression (P_0 = 0.537, A = 0.00401), makes a large difference in the probability of success.

From the point of view of obtaining 900 correct guesses, this distribution in fig. S6 is already an improvement. If we use the criterion that the error should be minimized and select a cutoff, we find that it should be at critical slope x_c = 0.00613, this gives an expectation of 886 correct answers (fig. S7, table S1). A theoretical estimate of the process standard deviation (given below) in this estimate is ±9. Therefore, we are close to the 900. In this case we find 886±9 correct so that now the odds are considerably better than one in a million of winning: they are now about a one in 13 chance (see table S1 for the parameters A, P_0).

Before continuing, how do we know that there are indeed these long range statistical dependencies that lead to this subtle - but for the sake of the contest highly significant - difference in success? The next two figures show that the distribution of fBm parameters H (the key scaling exponent) and volatility parameter s that controls the amplitude of the variability) are nearly identical for simulated and analyzed H, s - the only other parameter is the fBm trend discussed above, further confirmation comes from the stochastic model that we develop in section S2.

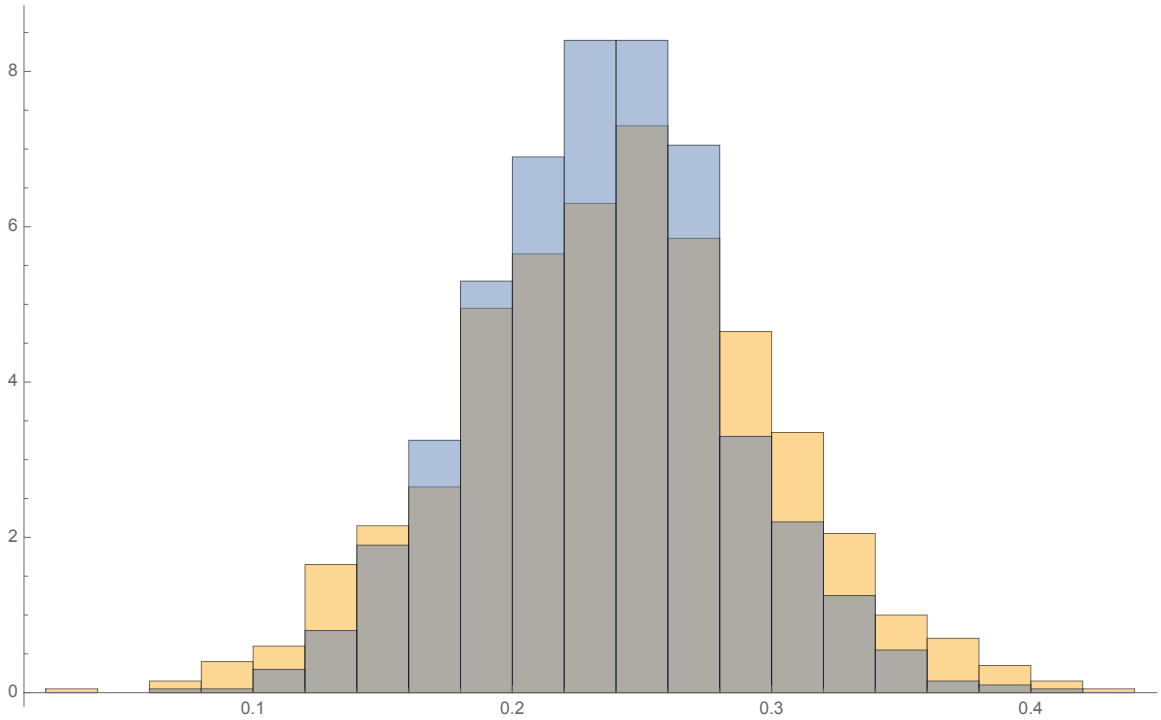

Fig. S4: The histogram (the vertical axis is the probability density function, the horizontal axis is the parameter H) of the distribution of H' s using H = 0.25 for the simulations (yellow), data, blue). Mean of the simulations: 0.234±0.048, of data: 0.240±0.050.

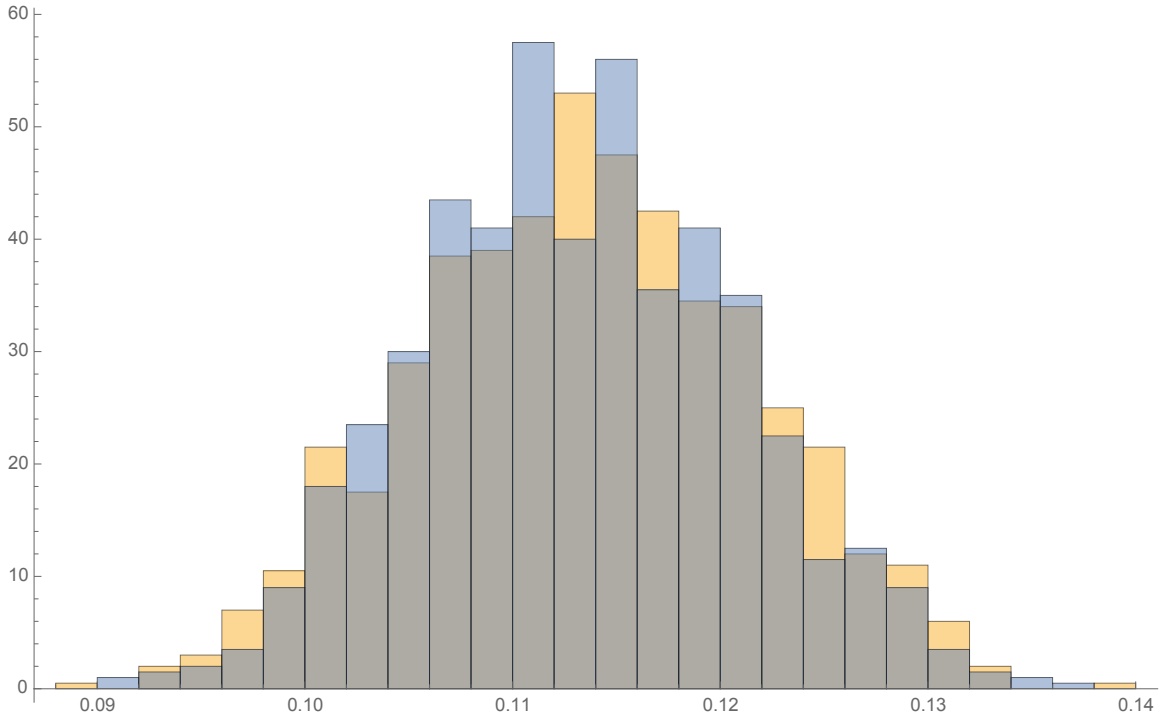

Fig. S5: The histogram (the vertical axis is the probability density function, the horizontal axis is the volatility parameter *s*) of the distribution of *s*' s (data yellow, model blue - note the interchange of colours with respect to the previous). For simulations: *s* =0.1132±0.007; for the data: 0.1135±0.008.

Figures S4, S5 also show that the analysis using the maximum likelihood method of the simulated pure fBm process with $H = 0.25$ gives a slightly lower mean (0.24) very close to the data (including the entire distribution which is simply the dispersion which is presumably due to the relatively short (135 point) length of the series. Certainly, the data and fBm with $H = 0.25$ and $s = 0.113$ are very close. The only exception is the distribution (and standard deviation) of the trends that are a bit too small.

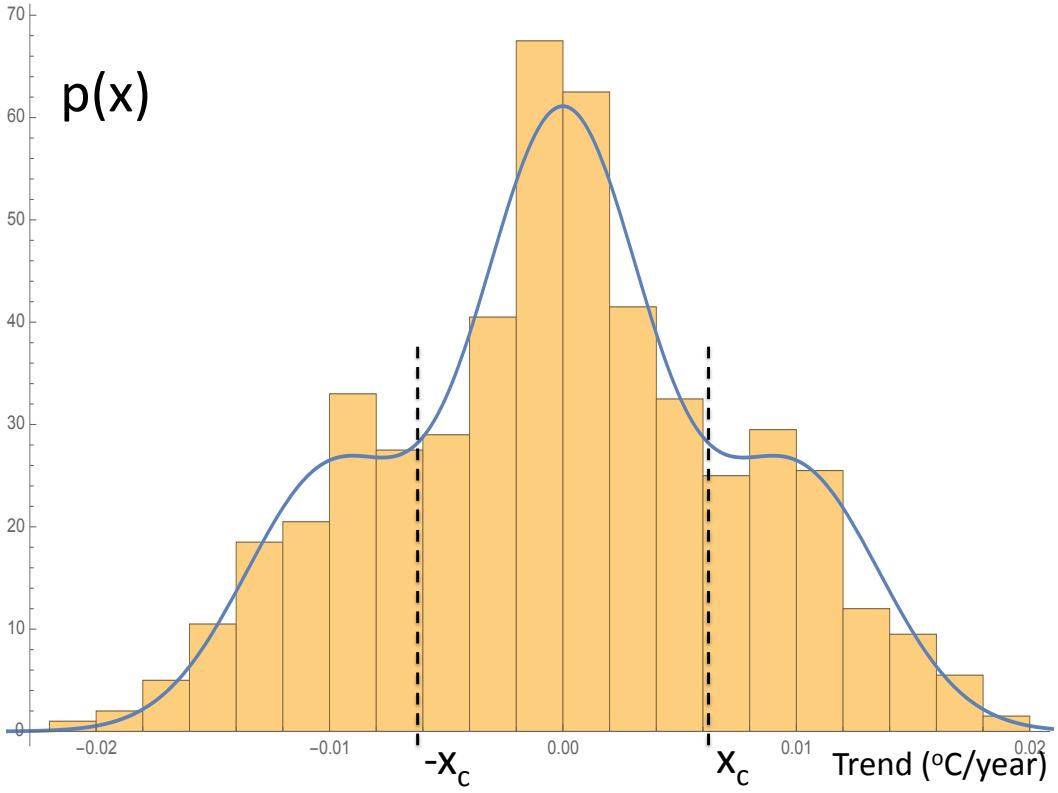

Fig. S6: The histogram (the vertical axis is the probability density function, the horizontal axis is the trend in °C/year) showing the distribution of the fBm trends with the fit, it is reasonably good. The critical classification boundaries x_c that minimize the classification error are shown as dashed lines.

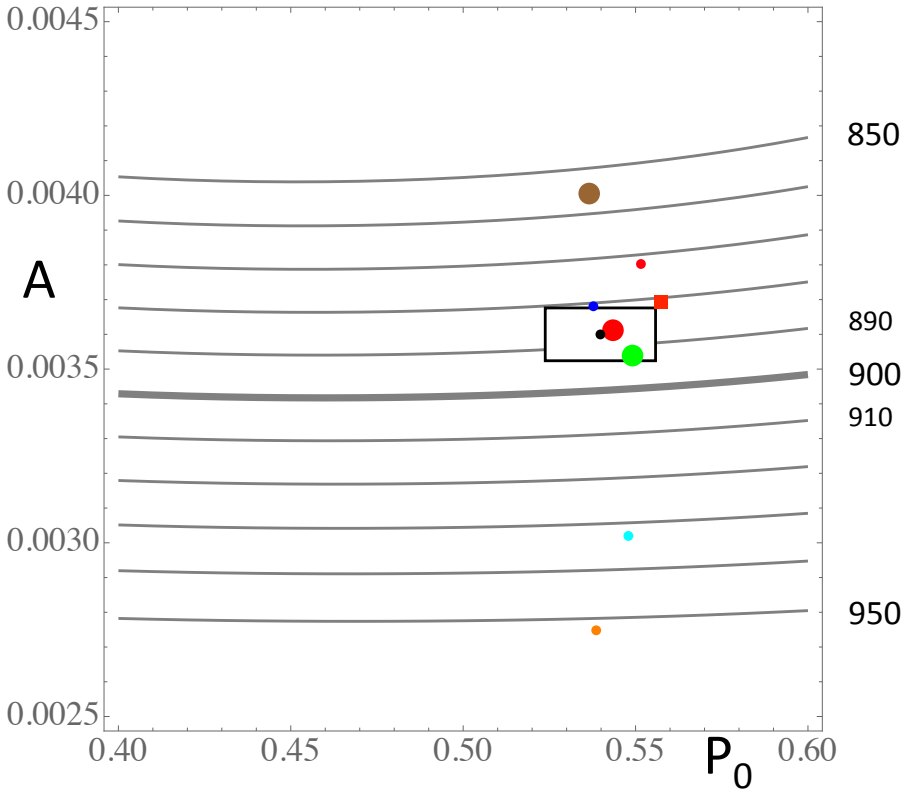

Fig. S7: A contour plot of the number of correct choices based on using the optimum regression slope or fBm trend as a function of the Gaussian width A (vertical axis) and probability of an untrended series (P_0), the horizontal. The contour lines indicate the following number of expected correct answers: 850, 860, 870... 940, 950 (top to bottom, roughly at intervals of one standard deviation), the thick line corresponds to 900 correct answers. Trend analyses of the data are shown as large circles, brown using regression (Gelman's result), and red from the fBm trends. The red square is the trend deduced simply from the difference of the temperature change over the entire series: (T(135) - T(1))/134. Interestingly it is much better than the regression and nearly as good as the fBm maximum likelihood method (see also table S1). This shows that using the fBm trends brings the data near the critical thick line corresponding to 900 correct answers. The green

is the data analysed but bootstrapped using 100 realizations of a process with 1000 random variables from the mixture of three Gaussian distributions (one around each of the -0.01, zero, +0.01 slopes); it is an attempt to take into account the biases in the (P_0, A) parameter estimates from regressions on the slope histograms.

Model results are shown as small circles all with scaling exponent for the fBm H = 0.25, and volatility s =0.113. The black circle is the model point with randomslope process standard deviation σ_{sproc}= 0.002375 and the rectangle centred on it indicates the 1 standard deviation limits as inferred by the model running 10 samples of 1000 realizations each; the blue is the same but with σ_{sproc} = 0.0024 showing the sensitivity to this parameter. The upper red point is the same simulation as the black point but using regressions to estimate the slopes. The orange point at the bottom is the simulation but without the slope process (fBm trends), the cyan point is the same but with slopes estimated from regression. In both cases the contest would be easy to win.

How to do better? The above estimated 3 Gaussian mixture with P_0, A, are the *a posteriori* parameter estimates. They are from the distribution of trends from Keenan's model, and then with the parameters P_0, A estimated from regression, or from fBm trend estimates. The estimated P_0, A parameters will thus be slightly biased. We need to use Bayes theorem... or an equivalent to estimate the original model parameters. To do this, we followed a "bootstrap" procedure. We made a model using the above probabilities and a Bernoulli process to take 1000 random trends from the above "three hump" distribution and then analyzed the result. We did this 100 times and found that every time, P_0 decreased a little and A increased a little... always by nearly the same amount (about 1% of the values). To first order,

these changes would be the same if we had started at the unknown original distribution. Therefore, to estimate the original parameters, we increased P_0 and decreased A in order to estimate the "original" distribution corrected for biases in the estimation procedure: $P_{0\text{original}}$ =0.549, A_{original} =0.00354. With these new parameters, one finds the critical slope dividing the untrended and trended populations, x_c = 0.006115 and this predicts 893±9 correct guesses! We are now close... but – unless we're lucky - we'll need more than one guess to win (see table S1)!

S2. Simulating Keenan's processes

S2.1 Statistically characterizing the processes

To fully understand Keenan's problem, we need to find a way to make simulations of his processes. Recall that a stochastic process with specific statistical properties can typically be obtained through many different construction mechanisms – the latter are not unique (think of the central limit theorem: sums of random variables with varied initial distributions end up have a Gaussian distribution). In this case, it is pertinent to recall that a process with statistically stationary increments (i.e. one that has first differences with the same statistical properties at all instants in time, see fig. 1) and which also has Gaussian statistics (fig. S8), is entirely determined by its second order statistics, such as the (Fourier) spectrum. However, at the level of spectra, the difference between the increments and the process is simply a power law filter so that if the increments are Gaussian,

then the process is still determined by its spectrum. In practice this means that we needn't find an identical recipe to Keenan, only an equivalent one.

An implication of the nonstationarity of the process (fig. 1) is that the autocorrelation function does not converge, (note that this is also true of the underlying fBm process), so that the autocorrelation must be replaced by its (near) equivalent, the second order structure function (unfortunately, this technical point is often ignored).

The classical structure function is the second order moment (mean square) fluctuation $\Delta T(\Delta t)$ over a lag Δt with the fluctuation defined as the difference in the function (here, the temperature $T(t)$) over the lag: $\Delta T(\Delta t) = T(t) - T(t-\Delta t)$. However, a complication here is that the definition of fluctuations in terms of differences is not adequate, we use a slightly different definition: the Haar fluctuation (already discussed in section S2), which is defined as the difference between the average of the first and second halves of the interval.

S2.2 Structure Functions and Probability Distributions

The previous subsection we indicated that if the increments are Gaussian and stationary, then the process is defined by the second order structure function. Let us therefore begin with a basic characterization of the process as a function of scale, using the Haar structure function. In fig. 2 we already showed the comparison of the Haar structure function for T, T_{init} determined from Keenan's simulations and the average of three temperature global, annual series analysed in L2014.

We now confirm that the process does indeed have nearly Gaussian increments. We calculated the probability distributions for increments with lags increasing by factors of two: Δt = 1 year, Δt = 2, 4, ...128 years. Note that the usual probability distribution function is the "Cumulative Probability Distribution " (CDF) which is the integral of the probability density starting at - infinity. The distributions used below are instead from a threshold s to + infinity i.e. they are equal to 1-CDF.

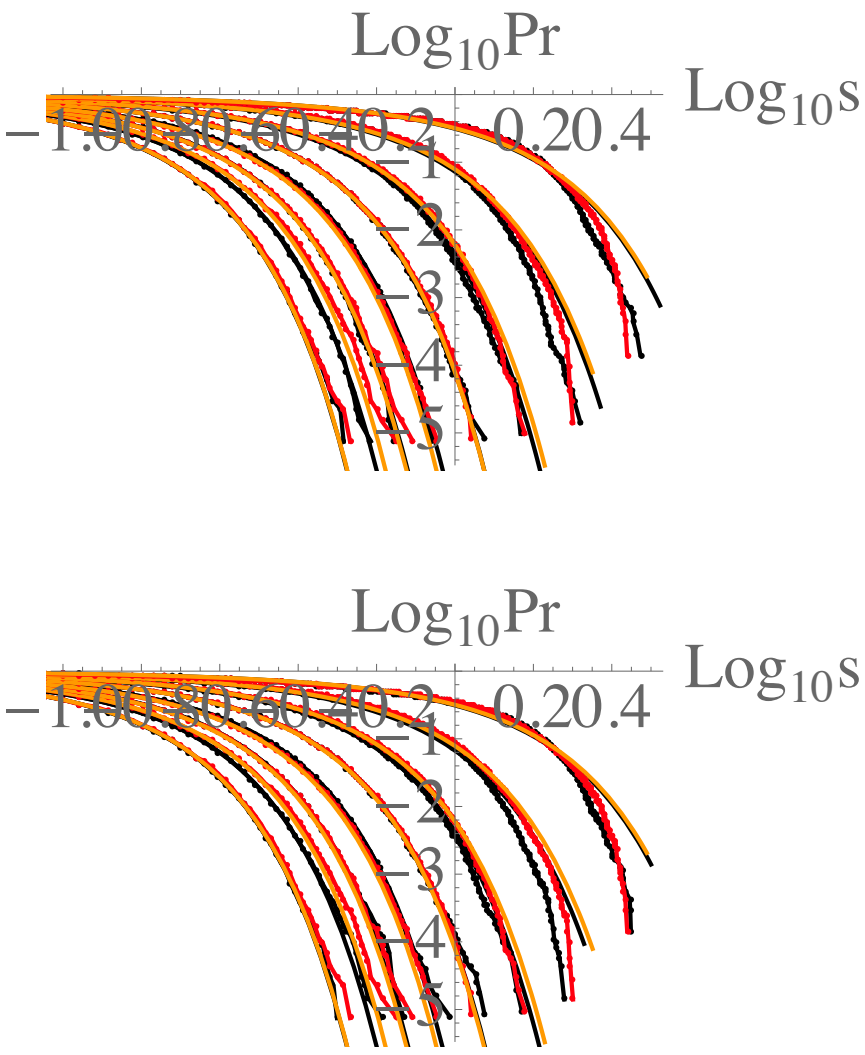

Fig. S8: The cumulative probability distributions of a temperature difference exceeding a fixed threshold s for Keenan's $T(t)$ process (red) and the simulation (black) with parameters H = 0.25 (fBm scaling exponent), s = 0.113 (volatility, i.e. the RMS variability at a given reference scale), P = 0.54 (the probability of an untrended series), σ_{sproc} =0.002375 (the standard deviation of the random slope process). The figure shows 2 different samples of 1000 series so that it is plausible that most of the differences between the simulations and the data are from random sample to sample variations.

The curves from left to right are for the differences at lags 1, 2, 4, 8... 128 years. Although there are some differences, to within sample to sample variations, (except at the largest scales) Keenan's and the simulated distributions are very close at all scales, and are also close to the Gaussians (indicated by the thin lines). The smooth curves are the best fitting Gaussians. Note that the amplitude of the distributions increase systematically with the lag (the curves move systematically to the right at longer and longer lags Δt). Also, the main discrepancy between the distributions from $T(t)$ and from the model appear to be at a time scale of about 64 years.

It is also of interest to compare Keenan's probability distributions with those of the real world, see fig. S9; these were used to estimate the probability density in fig. 3. This shows that the differences are nearly independent of scale[4]; that the true exponent is $H \leq 0$ ($H \approx -0.1$; see e.g. fig. 2 it is approximately an fGn) unlike Keenan's model that has H = 0.25 (i.e. >0, for the fBm). Also, notice that the extremes are not Gaussian, but rather "fat tailed" bounded here between s^{-4} and s^{-6} (green) where s is

[4] When comparing fig. S8 and S9, it is important to notice that for clarity S9 has been multiplied by the lag Δt: without this multiplication, all the curves in S9 would be superposed on top of each other unlike the model distribution S8 whose amplitude systematically grows with lag.

a temperature threshold (see section 2 for a discussion of these "black swan" events).

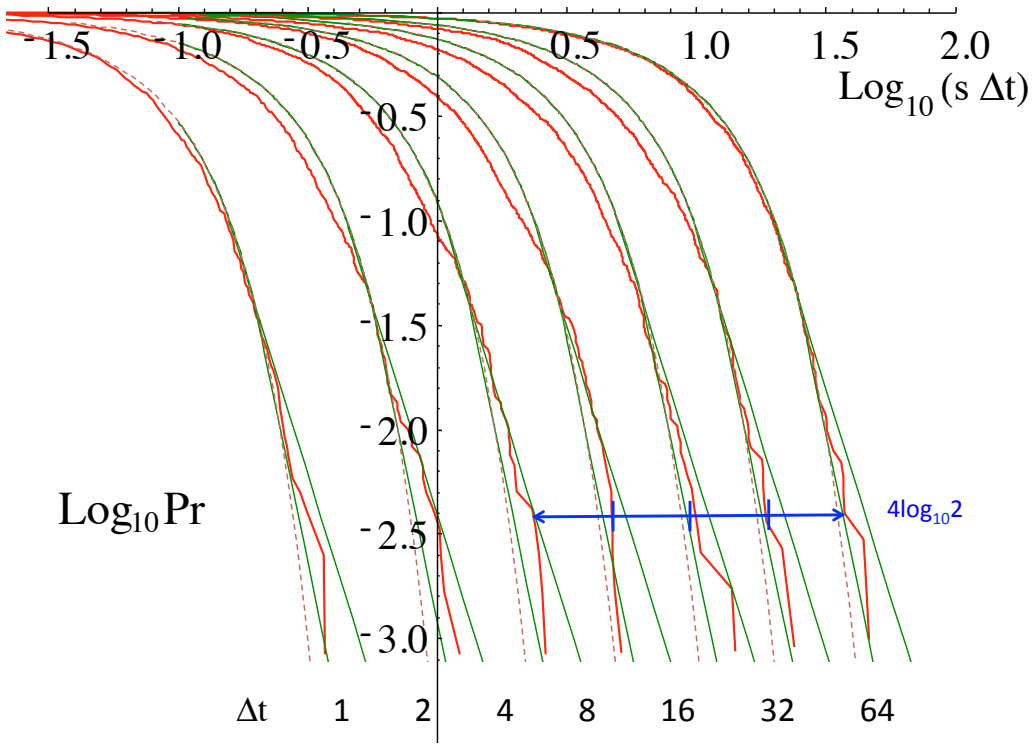

Fig. S9: The cumulative probability distribution of preindustrial (1500-1900) northern hemisphere temperature changes exceeding a threshold *s* after being *rescaled by the time lag* Δt using three multiproxies, reproduced from [*Lovejoy*, 2014a] where the series used are also discussed. The rescaling was used to separate the distributions, otherwise they would all be nearly on top of each other. This shows that the scaling of these differences[5] has nearly H =0 (unlike Keenan's that has H = 0.25 for the fBm). Also, notice that the

[5] A technical point is that if the properly defined fluctuations have $H<0$ (the data have $H \approx -0.1$), then the differences - as here - will have an exponent $H = 0$.

extremes are not Gaussian, but "fat tailed" bounded here between s^{-4} and s^{-6} (green) where s is a temperature threshold.

S2.3 Simulating the process, a missing extra random trend process

In the earlier section, we argued that the process $T_{init}(t)$ could be approximated by a fractional Brownian motion (fBm, denoted $B_{H,s}(t)$, see eq. S3, footnote 3, with scaling exponent H =0.25±0.004, and volatility s =0.113±0.007) and we used the maximum likelihood method on an fBm (i.e. assuming that residuals of trends had strong, long range statistical dependencies/correlations) in order to get a tighter estimate of the trends than would have been possible using standard regressions. In the above section we confirmed that the process has (nearly) Gaussian increments and we have characterized it's second order correlation structure.

The simplest model would thus appear to be $B_{H,s}(t)$ (fBm, footnote 3) with a Bernoulli process $b(t)$ for the added trends (or absence of added trends):

$$\begin{aligned} & \Pr(B_n = 0) = P_0 \\ b(t) = B_n t; \quad & \Pr(B_n = a) = \left(\frac{1-P_0}{2}\right); \quad a = 0.01 \\ & \Pr(B_n = -a) = \left(\frac{1-P_0}{2}\right) \end{aligned} \tag{S6}$$

where P_0 is the probability that the process is untrended. If there were only the Bernoulli trends and the fBm, then we would have:

$$T(t) = B_{H,s}(t) + b(t) \tag{S7}$$

and (see fig. S7, the orange dot at the bottom) we find that the resulting A value is much too small to reproduce the data: it would be able to correctly classify 940 of the series. It seems that Keenan - presumably realizing this – added an extra random slope process (or an equivalent complication) so as to effectively increase A. If this is correct, then the contest may have been deliberately made unwinnable. Certainly, with only the fBm and Bernoulli processes (with H = 0.25, s = 0.113 and P_0 =0.54), the width of the distributions of trends is σ_{trend} =±0.00735, and this may be compared with the direct determination (using the fBm trend estimates) of σ_{trend} =±0.00772±0.0014 which is compatible with eq. S2 estimate 0.00768 (using A =0.00361±0.000076, P_0 = 0.054±0.016).

The simplest random slope process $r(t)$ is one that has a trend α with a Gaussian random trend; it should be symmetric (zero mean) so that only it's standard deviation σ_{sproc} needs to be determined. It is of the form:

$$r(t)=\alpha t;\quad p(\alpha)=\frac{1}{\sigma_{sproc}}G\left(\frac{\alpha}{\sigma_{sproc}}\right) \tag{S8}$$

where $G(x)$ is the unit Gaussian introduced above and α is the random trend of $r(t)$ (i.e. $<\alpha>=0$, σ_{sproc} is the standard deviation of α).

The overall temperature process, the "full" model is thus:

$$T(t)=T_{init}(t)+b(t);\quad T_{init}(t)=B_{H,s}(t)+r(t) \tag{S9}$$

Since the variances of the trend of the new process will just add to the variance of the previous process (fBm plus Bernoulli trends) we have the variance of the whole process:

$$\left\langle T(t)^2 \right\rangle = s^2 t^{2H} + a^2 t^2 \left(1 - P_0\right) + \sigma_{sproc}^2 t^2 \tag{S10}$$

and for the trends (see eq. S2):

$$\sigma_{trend}^2 = A^2 \left(1 - P_0\right) = A'^2 \left(1 - P_0\right) + \sigma_{sproc}^2 \tag{S11}$$

where A' is the width parameter without the additional trend process $r(t)$. Equation S11 shows how the width A increases (from A') with the addition of the $r(t)$ process. It is thus easy to estimate σ_{sproc}; we find the optimal value σ_{sproc} = 0.002375. The resulting process P_0, A values are shown in fig. S7, they are extremely close (within the uncertainty limits, see the black point and the rectangle) to the data (the big red point).

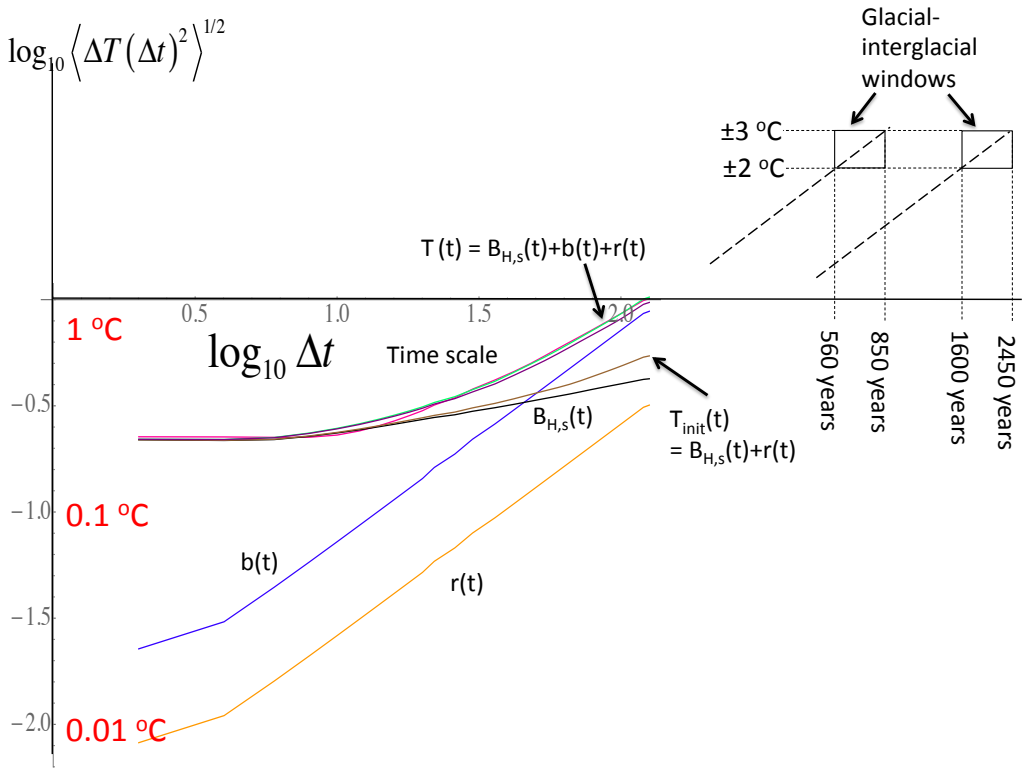

Fig. S10: This is root mean square Haar structure function for the data and simulation and the different components of the simulation (statistical independence of the various processes implies that the variances add (the square of the RMS $S(\Delta t)$ functions). Pink is Keenan's $T(t)$ (reproduced from fig. 2), black is the pure fBm process ($B_{H,s}(t)$, s =0.113, H =0.25), blue is pure Bernoulli slope process ($b(t)$ with P_0 = 0.54) and yellow is pure random slope process with (σ_{sproc} = 0.002375). Brown is $T_{init}(t)$ =$B_{H,s}(t)$+$r(t)$ (the fBM plus the random slope process), purple (near the top) is the fBm plus the Bernoulli slope process ($B_{H,s}(t)$+$b(t)$), and green is the full process ($T(t)$ = $B_{H,s}(t)$+$b(t)$+$r(t)$, i.e. fBm plus Bernoulli plus random slope).

Also shown (upper right) is the extrapolation of $T(t)$ to longer time scales (see fig. 2 for $T_{init}(t)$). The model $T(t)$ predicts typical temperature fluctuations of ±2 °C at 560 years and ±3 °C at 850 years and $T_{init}(t)$ at 160 and 2450 years (the rectangle). Since going in and

out of an ice age is a change of roughly this order this is the "glacial- interglacial window". Realistic time scales from paleo data are roughly 100 times longer [*Lovejoy and Schertzer*, 1986].

Let us now consider the full model, eq. S9 with parameters $H = 0.25$, $s = 0.113$ for the fBm process and a random trend process with Gaussian distribution with standard deviation $\sigma_{sproc} = 0.002375$ for $T_{init}(t)$. Superposed on this (to yield $T(t)$) is the sign symmetric Bernoulli trend process with parameters $P_0 = 0.54$ corresponding to 54% untrended series.

To check that the model is realistic we can calculate the structure functions figs. S10, S11 or the probability distributions at different lags/scales, fig. S9). To understand fig. S10, start with the fBm (black) curve and the pure Bernoulli process (blue curve). We see that high frequencies (small Δt's), that the fBm simulation gives nearly exactly the same behaviour as the data (pink) out to about 20 years. At large Δt's, (beyond about 30 years (i.e. $\log_{10}\Delta t \approx 1.5$) the entire process is dominated by the Bernoulli slope process ($b(t)$, blue). Finally, the random slope process ($r(t)$ orange) is much smaller than the Bernoulli slope process and only gives a very small contribution to the structure function (the difference between the purple and the green). However, as we saw, it makes it just a little tougher to determine which series are trended, and this makes the difference between winning and losing the contest. Finally, a blow-up comparison of the full process and $T(t)$ is shown in fig. S11; we see that over the entire range the structure functions of our model and $T(t)$ are nearly identical; the only region over which they are noticeably different is near

10 years where the logs are different by about 0.03, corresponding to the simulations being about 7% more variable. But the fluctuations at this scale are practically irrelevant for determining slopes so that this small difference is probably not significant for estimating the trends.

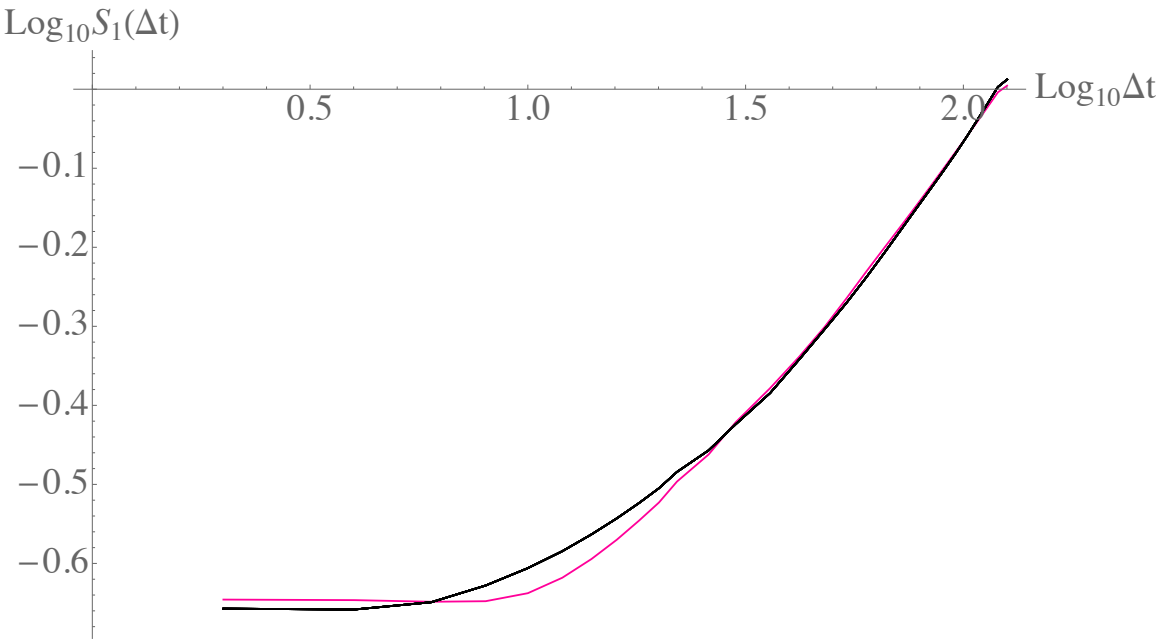

Fig. S11: A blow up of the root mean square Haar structure functions in fig. S10 for Keenan's data (red). The black curve is the full model with the optimal values H =0.25, s = 0.113, P_0 = 0.54, σ_{sproc} =0.002375.

Before concluding let us make a comment. Keenan only stated that he started with a "trendless process" and of course fBm models already have random trends. However the contest would have been too easy so that he clearly did something else that effectively increased (a little bit) the amplitude of the random trends. At first

sight it might seem that the method that we propose to reproduce his statistics – the addition of the random slope process $r(t)$ to the fBm - seemingly contradicts his statement that he first made 1000 series of a trendless process and then added in the Bernoulli process. However, since the added random slope process also has stationary increments and is Gaussian, the sum of this process with the fBm is indeed itself a trendless process. Ironically, since his data $T(t)$ are essentially sign symmetric, it would appear that his overall process – i.e. including the Bernoulli slope process - is also trendless (indeed in fig. 1, $\mu_T(t) \approx 0$)! In any event, it is quite possible – even likely – that Keenan produced his series with a different construction mechanism, one that has statistical properties very close to the one described here. It is possible that the small difference – associated with the model's slight excess of variability near 10 years is enough to allow the contest to be won- at least in principle. Good luck to any new attempts!

S2.4 Using the model to better understand the problem

With the model, we have full information about the $T(t)$ process and we know exactly which series were trended and which were untrended. Table S1 shows some of the results that we obtained by running the model for ten samples of 1000 series of 135 points each with slightly different σ_{sproc} parameters (it was sensitive to this!). Aside from the derived parameters A and x_c, σ_{trend}, we can make a posteriori estimates of P_0.

We can also estimate other interesting properties. For example, if the absolute trends are sorted in increasing order, then the optimum index is the rank of the

trend that should be chosen to yield the maximum number correct (third row); note the large sample to sample fluctuations. Since we know in advance which series are trended and which are not, we can also estimate the maximum possible number of correct choices in a sample of 10,000 (fourth row); or the equivalent for 1000 series samples 10 times (fifth row). If only the ensemble best threshold was known (i.e. we had no information about the underlying trends/no trends (this is the realistic case), then we obtain the number correct, in row 6 (out of 1000 series).

Simulations					fBm (Data)		Regression		T(135) - T(1)
Model σ_{sproc}	**0.00230**	**0.00235**	**0.002375**	**0.0024**	**Raw**	**Boot-strap**[i]	**Data**	**Sims**	**Data**
Optimum index[a]	550±32	547±17	552±18	541±29	551	550±17			
No. correct in 10,000[b]	8933	8886	8896	8842					
No. correct optimum[c]	897±9	893±8	892±9	888±10					
No. correct ensemble threshold[d]	892±10	887±9	888±9	887±9	886±9	893±9	856±9	872±9	880±9
Critical threshold[e]	0.00606± 0.00059	0.00593± 0.00032	0.00569± 0.00031	0.00639± 0.00055	0.00613	0.00612	0.00635	0.00630	0.00627
A[f]	0.00354	0.000358, 0.00356	0.00360, 0.00356	0.00368	0.00361	0.00354	0.00401	0.00380	0.00373
P_0 [g]	0.535	0.527, 0.543	0.540, 0.541	0.538	0.5435	0.549	0.537	0.551	0.554
σ_{trend} [h]	0.00772	0.00781, 0.00763	0.00771± 0.00036	0.00773± 0.0037	0.00772		0.00793	0.00775	0.00772

Table S1: The model parameters are: H = 0.25, s = 0.113 (H is the fBm exponent, s is the volatility parameter of the fBm, see equation in footnote 3). P_0 = 0.54 for the Bernoulli process and σ_{sproc} for the random slope process (with the exception of the third line (a single sample of 10,000), all the numbers are with respect to a sample of 1000 series, each 135 points long and the uncertainties are with respect to such samples). With the exception of those indicated "regression" (two far right columns) all the trends were estimated using the maximum likelihood method assuming an fBm process. The far right column is from the data when slopes are estimated by the difference between the first and last member of the series (see fig. 3).

[a] If the absolute trends are sorted in increasing order, then the optimum index is rank of the trend that should be chosen for the maximum number correct. Note the large sample to sample fluctuations.

[b] This is the maximum possible number of correct choices in a sample of 10,000; it was calculated with knowledge of the correct classifications.

[c] Same as b except for mean (and standard deviation) of the ten samples.

[d] If only the ensemble optimum threshold was known (this is the realistic case), this is number correct, samples of 1000. It is slightly less than the "optimum" number in the row above.

[e] This is the mean (and standard deviation) of the optimum threshold.

[f] A is the width of the Gaussian of the slopes about the three values -0.01, 0, 0.01 (see eq. S1). From the simulations, the uncertainties with respect to the process parameters A is ±0.000076 and for P_0, ±0.016.

[g] This is the fraction of untrended series (Bernoulli trend process $b(t) = 0$).

[h] This is the standard deviation of the raw trends.

[i] The bootstrap parameters were estimated as discussed in the text; they are probably optimal for Keenan's series.

As expected, the estimate from the ensemble trend threshold yields slightly less than use of the "optimum" threshold indicated in the row above. Although the model parameter space is big (H, s, P_0, σ_{sproc}), as explained above, the parameters are in fact constrained by a number of statistics: recall that H, s were chosen to reproduce the fBm maximum likelihood estimates, P_0, the empirical trend distribution, σ_{sproc} the width A and the overall standard deviation of the trends σ_{trend}. Overall it seems that we can get tantalizingly close, to 900 correct, but not quite get there.

Appendix: Some theory on randomly classifying trends

In the climate contest problem there are trended and untrended series. We saw that we could get to within one standard deviation of a winning technique by using the fBm (strong correlation) based trend estimates and a bootstrap procedure for determining the optimum classification threshold x_c. It is therefore tempting to "shuffle" or randomize the classification method so as to obtain many possible guesses so that at least one will be a winner. Unfortunately, the theory below shows that this is not a very fruitful approach.

Denote the (unnormalized, relative) probability density of an absolute trend x from the untrended population as $p_u(x)$, from the trended population, $p_t(x)$. Let here be a total number of N draws (i.e. series, here $N = 1000$).

The expectation for the number of untrended and trended series is:

$$\langle N_u \rangle = N \int_0^\infty p_u(x')dx'; \quad \langle N_t \rangle = N \int_0^\infty p_t(x')dx'$$

$$\int_0^\infty \left(p_u(x') + p_t(x') \right) dx' = 1; \quad N = \langle N_u \rangle + \langle N_t \rangle \tag{A1}$$

If we now use the rule that any trend $x' \leq x$ is classified as untrended, and any with trend $x' > x$ is trended, then the mean number of errors is:

$$\langle E_u(x) \rangle = N \int_x^\infty p_u(x')dx'; \quad \langle E_t(x) \rangle = N \int_0^x p_t(x')dx' \tag{A2}$$

The condition to minimize the total mean error is:

$$\frac{\partial \langle E \rangle}{\partial x} = 0; \quad \langle E \rangle = \langle E_u \rangle + \langle E_u \rangle \tag{A3}$$

Which implies that the critical x_c satisfies:

$$p_u(x_c) = p_t(x_c) \tag{A4}$$

Fluctuations:

For a large number of series, the probabilities will approach a Gaussian, we need to calculate the variance of the total error $\langle E^2 \rangle$. From a random sample of N trends, the probability that the above procedure will misclassify a given series is P_E:

$$P_E(x) = \int_x^\infty p_u(x')dx' + \int_0^x p_t(x')dx' \tag{A5}$$

We seek the variance of the Bernoulli process for which for a single series, E = 1 with probability P_E, 0 otherwise. The overall variance for N series is then N times larger:

$$\left\langle E^2 \right\rangle = NP_E(x)\left(1 - P_E(x)\right) \tag{A6}$$

Thus, the random variable E should be a Gaussian with:

$$E = NP_E(x) \pm \left[NP_E(x)\left(1 - P_E(x)\right)\right]^{1/2} \tag{A7}$$

(the notation indicates a Gaussian with mean NP_E and standard deviation after the "±"). In our case, $E \approx 100$, N = 1000, $P_E \approx 0.1$ so that: $E \approx 100 \pm 9$ as confirmed by the numerics.

Randomization by a Flipping process:

We can now see the effect of randomizing the variable by flipping each choice with a probability density $p_f(x)$; this means that we first choose as before: we classify a series as untrended when $x<x_c$, and as trended when $x>x_c$, but then we reverse the choice with probability $p_f(x)$.

By similar reasoning to above (calculation of the mean) we find that there will be a shift δE in the number of errors from $E(x)$ to $E'(x)$:

$$\delta E(x) = E'(x) - E(x) \tag{A8}$$

Mean shift δE:

$$\delta E = N\left[\int_0^{x_c} p_f(x')\left(p_u(x') - p_t(x')\right)dx' + \int_{x_c}^{\infty} p_f(x')\left(p_t(x') - p_u(x')\right)dx'\right] \tag{A9}$$

since from the above condition for the critical x_c we have:

$$\begin{aligned} p_u(x) &> p_t(x); \quad x < x_c \\ p_u(x) &< p_t(x); \quad x > x_c \end{aligned} \tag{A10}$$

and from the fact that $p_f \geq 0$, we can see that $\delta E \geq 0$, so that no choice of $p_f(x)$ can improve the initial choice x_c. But what about fluctuations?

Variance of shift:

Repeating the same Bernoulli argument for the variances, we obtain:

$$\sigma_{\delta E} = \delta E^{1/2}\left(1 - \frac{\delta E}{N}\right)^{1/2} \approx \delta E^{1/2} \tag{A11}$$

Thus, the overall result is that "flipping" increases the error by:

$$\delta E \pm \delta E^{1/2} \tag{A12}$$

Since $\delta E \approx 9$, $\sigma_{\delta E} \approx 3$ (also confirmed by numerics). We see that flipping will only decrease the error at the 3 standard deviation level i.e. only 0.13% of the time, and that the method of flipping ($p_f(x)$) is (surprisingly) irrelevant! If we are 1 standard deviation (9 counts) below 900, then we need a 6 standard deviation event in order to win, i.e. we would need about a billion entries to the contest.

References:

Lovejoy, S. (2014a), Scaling fluctuation analysis and statistical hypothesis testing of anthropogenic warming, *Climate Dynamics*, *42*, 2339-2351 doi: 10.1007/s00382-014-2128-2.

Lovejoy, S. (2014b), Return periods of global climate fluctuations and the pause, *Geophys. Res. Lett.*, *41*, 4704-4710 doi: doi: 10.1002/2014GL060478.

Lovejoy, S. (2015), Using scaling for macroweather forecasting including the pause, *Geophys. Res. Lett.*, *42*, 7148–7155 doi: DOI: 10.1002/2015GL065665.

Lovejoy, S., and D. Schertzer (1986), Scale invariance in climatological temperatures and the local spectral plateau, *Annales Geophysicae*, *4B*, 401-410.

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