Manuscript prepared for Earth Syst. Dynam. Discuss.
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 28 September 2015

Response to reviewer Shaun Lovejoy

T. Nilsen[1], **K. Rypdal**[1], **and H.-B. Fredriksen**[1]

[1]Department of Mathematics and Statistics, UiT The Arctic University of Norway, Norway

Correspondence to: Tine Nilsen (tine.nilsen@uit.no)

Abstract

Shaun Lovejoy's report is not a conventional review of our discussion paper, but rather an essay which is a critique of our paper, but also a review and defence of his own work. A point-to-point response is almost impossible, and we feel confident that this is not the kind of discussion that the reviewer wants. We will therefore address his comments by following the logical structure of the review. Our response contains a number of new analyses and figures. Some of the text and figures may be incorporated in the revised paper and Supplementary Material. Substantial revision of our paper will follow as a result of both reviews, but an outline of this revision will not appear in this response, but at the end of the reponse to reviewer #2.

1 Clarifying the basic issues

1.1 Scaling analysis of evolving systems

On geological time scales the Earth is an evolving system. There are cycles, but the Earth rarely repeats itself. The Eemian was similar to the Holocene, but also very different, the most striking difference being the evolution of human civilizations. Thus, the dynamics of the Earth is non-stationary in a very fundamental sense. This makes scaling analysis, and modelling of Earth processes based on such analysis, a quite problematic issue. It has little meaning to talk about a universal scaling in Earths climate since the scaling characteristic on a given range of scales up to a chosen maximal scale τ_{max} will depend on the eon, era, period, epoch, or age the analysis is done. In other words, the result will depend on the time t around which the time range τ_{max} is centred. A scaling analysis of a given Earth-system variable must therefore be be conditioned by two essential parameters; the range τ_{max} of scales considered, and the positioning t of this range in time. One obvious mathematical and conceptual tool for handling such non-stationary data is the wavelet transform, but more about that later.

The controversy we have with Shaun Lovejoy has its root in different understanding of how this issue of non-stationarity could be handled. Ice cores restrict the information we can obtain to somewhat less than $\tau_{\mathsf{max}} = 1$ Myr BP. This is the range of time scales considered in Lovejoy's work, and the period is the Quaternary (2.5 Myr to present) in which the Earth's climate has been in a bistable state shifting between glacials and interglacials. Lovejoy's methodology and interpretations are based on this choice of the parameters (τ_{max}, t). We don't see anything wrong with that, as long as one is mindful on that this is a choice, and recognises that there are other, equally valid, choices. It is here our views collide, because Lovejoy tells us that his choice is the only one worth to pursue.

1.2 Climate prediction in the Holocene based on scaling analysis

It seems unclear whether scaling analysis per se can be helpful in understanding glacial-interglacial transitions, but if this is the issue, Lovejoy's choice is certainly a reasonable one. If there are strong fluctuations on time scales of millennia they can contain the seed for a flip from an interglacial to a glacial and vice versa. On the other hand, if the issue is understanding of the present and future climate in our present interglacial state, we don't believe this choice is useful, simply because it ignores the knowledge that the Earth at present resides in an interglacial state and probably will continue to do so as long as there is human civilisation and anthropogenic forcing on this planet.

The time series and the wavelet scalogram of the GRIP temperature series for the past $t_{\mathsf{max}} = 90$ kyr illustrates the issue, and shown in Figure 1 of this comment. The central time parameter t is along the horisontal axis and the scale τ along the vertical. We have no data for the future, which means that the transform cannot be computed correctly above the upper white line in the figure. Likewise, the area below the lower white curve is influenced by the interpolation made due to uneven sampling of the time series. It is apparent that the scalogram is different in the first 11.5 kyr (the Holocene) from the remaining 80 kyr (the last glacial). There is generally lower power on all scales in the Holocene, and the increase in power with increasing scale as t is kept constant is lower. This is a signature of the different scaling exponents characterising the interglacial and glacial temperature

fluctuations in Greenland. The difference between glacial and Holocene fluctuations is also illustrated very clearly in Fig. 4a of Lovejoy's report. Unfortunately, the Holocene so far has lasted only about 11.5 kyr, which limits the scales accessible by the wavelet transform (or any other method of fluctuation analysis) to those confined by the upper white line. This means that we cannot say anything certain about Holocene scaling for scales beyond a few kyr.

It is not unlikely that the anthropogenic perturbation has overridden the orbital forcing for several millennia already (early deforestation) and that the Earth will remain in the interglacial state for sufficiently long time to make our remote successors able to establish accurately the scaling properties of Holocene climate up to scales of tens of kyr. It can be argued that the anthropogenic perturbation will change the scaling, and therefore that our successors will find scaling characteristics different from those that has ruled the Holocene until now. But, if we want to use a statistical model based on empirical scaling properties as a tool for prediction, we have to base it on observations in the past. The central issue is then whether we should use the scaling obtained from time series dominated by the glacial state ($\beta \approx 1.7$ for the GRIP record) or a model based on observations from the Holocene only ($\beta \approx 0.5$ for GRIP). Our position is that it is unreasonable to use data from the glacial state to make centennial-to-millennial predictions for the Holocene.

1.3 The issue of uncertainty and hypothesis testing

If we use only Holocene observations, the limited length of the time series confronts us with the issue of uncertainty. This issue is largely ignored in Lovejoy's work, and in section (c) of his review he makes an attempt to justify it. He also argues in section (a) that it is unnecessary to use statistical models or "any assumptions about the scaling or otherwise of the temperatures. Conclusions can be verified using straightforward fluctuation analyses."

In our ears this sounds like a rejection of the scientific method. Without models it is not possible to perform statistical hypothesis testing. Lovejoy's preferred methodology is to perform estimates on some samples and then draw positive conclusions regarding the validity of hypotheses, running a great risk of committing type-I statistical errors (false positives).

The widely accepted approach is to test a hypothesis (e.g., the hypothesis of a scaling break at 100 yr) against a simpler null hypothesis. This null hypothesis *must* take the form of a statistical model, otherwise one cannot assess the probability of obtaining the observed sample under the null hypothesis. The natural choice of a null hypothesis is the simplest possible noise model which seems compatible with the observed data, but still is distinguishable from the alternative hypothesis that we want to test. In section 1.8 we shall demonstrate that a fractional Gaussian noise (fGn) is the proper choice for the Holocene surface temperatures by showing that intermittency is negligible.

One result of our paper is that the fGn (single scaling regime) null hypothesis cannot be rejected for scales up to a millennium by available observation data from the Holocene. A stronger, and more important result, is that considering the scale regime from $10 - 10^4$ yr even a white-noise null model cannot be rejected by most proxies (see Fig. 2c and Table 1 of our paper). Lovejoy quite correctly points out that these findings do not reject the scaling-break hypothesis, but we never claim that. Our claim is that observations are consistent with the fGn null hypothesis, and because of its simplicity, this should be the preferred model for predictions on time scales up to several centuries. Lovejoy et al. have recently published a paper where in effect the fGn-model ($\beta < 1$) is used for prediction on time scales up to a few decades. Following his logic, prediction on longer time scales should employ an fBm-model ($\beta \approx 1.7$), or even some multifractal version.

1.4 The notion of a macroweather-climate scale break

The occurrence of a scale-break around time scales of $\tau_c \sim 100$ yr is used by Lovejoy to justify his notion of “macroweather", as opposed to “climate." GCM experiments with and without full ocean circulation suggest that the main mechanism for the creation of scaling and long-range memory in the climate system is the energy exchange between subsystems with a wide range of response times, and that there are response times both smaller and larger than the “magic” scale $\tau_c \sim 100$ yr of the macroweather-climate transition. Our perception is that when such a transition scale τ_c appears when non-detrending scale estimators are applied to data, it is always explicable as a result of a particular external forcing

or a distinct oscillatory mode. This is why τ_c in Lovejoy's work varies from 10 to 100 yr depending on which part of the Holocene his data cover, and for the Holocene Greenland data $\tau_c \sim 1000$ yr. We find no evidence supporting the notion that there is a universal transition time scale, and that strong large scale variability signifies a new scaling regime characterised by a scaling exponent. When scaling appears to be broken, it brings more useful insight to investigate the particular events that are causing it, rather than taking it as a natural transition to a new scaling regime that is called "climate."

Another point is that scaling is also broken at scales shorter than decadal. One striking example is ENSO, which destroys the scaling properties in the Pacific tropics and subtropics, and even affects the scaling of global temperatures. According to Lovejoy, the timescales of ENSO classifies it as a macroweather fluctuation. We, on the other hand, don't find any compelling reason to classify e.g., Dansgaard-Oesgher events as climate variability and ENSO as macroweather.

For these reasons we are not so enthusiastic about the macroweather-climate notion, but that is not the subject of our paper. The issue is the support one can find in the data, and here the main evidence is presented in fluctuation measures or spectra derived from composites of proxies representing different time intervals and scales. Examples of such composites are presented Figs. 1, 3a,b and 4b in Lovejoy's review report. We will comment on those in the next section.

1.5 The potential fallacies of composite spectra

Some composites shown by Lovejoy are intended to demonstrate the existence of regimes of different scaling, supposed to represent scale regimes at which different physics dominate the fluctuations. As mentioned earlier, fluctuations on long time scales can only be investigated by probing far back in time, possibly into climate states different from the present. In other words, it can be difficult to distinguish a scaling break from an evolutionary change (nonstationarity) of the climate system, i.e., from a change in time of the scaling exponents. An illustrating example is Fig. 1 in Lovejoy's report. The figure shows the square-root of the second-order structure function for 5 different proxies. For the scales up to 400 yr it

shows the fluctuations of the 400 yr long Central England Temperature record (CET), and for a Northern Hemisphere mean temperature for 1850-1969 (*Budyko*, 1969). For a process where fluctuations decrease with incrasing scale ($\beta < 1$), this fluctuation measure gives a flat curve in a log-log plot of fluctuations vs. scale. Hence, for the CET record it is flat because the trend is weak. The NH-mean series, however, covers only the industrial period and is dominated by the anthropogenic warming trend. For a trend-dominated signal the characteristic exponent estimate H for this structure function is positive (actually close to unity). Similar results are obtained on scales 10^2 – 10^5 yr obtained from ice cores, but the nature of the signal causing these similar exponents is profoundly different. It illustrates that scaling analysis based on one single estimator and without careful physical interpretation can be very misleading.

In this plot all observations used for scales < 400 yr are made in the Holocene, while almost all observation for longer scales are based on observations during glacial periods. Hence, what appears as a scaling break around 400 yr when the CET record is used for the short scales, could just as well be due to different scalings in the glacial and interglacial states.

In Fig. 3b in Lovejoy's report the Haar fluctuation of some short multiproxy reconstructions (1500-1979 AD) of Fig. 3a are combined with some very long proxy records. The EPICA record is not so interesting in the present context, since it is necessarily dominated by the glacial state. The *Marcott* (2013) reconstruction, however, covering the entire Holocene, might have the potential to be the answer to our prayers about reconstructions with global distribution that covers the entire Holocene. Lovejoy's analysis, however, suffers from a serious flaw due to uncritical use of these paleoseries. In Fig. 3b he shows the Haar fluctuation for four different series, two for global temperature and two for 30-90°N. Each pair consists of a long series covering the entire Holocene and a short series covering 1500-1900 AD. *Marcott* (2013) write that the long series recover no variability for scales less than 300 yr, 50% of the variability on scales of 1000 yr, and all variability on scales greater than 2000 yr. The short series presumably recover virtually all variability on the (short) scales they cover. This low-pass filtering of the long series by the reconstruction method alters the scaling

and creates a spurious increase of the scaling exponent. This fact is very apparent from Lovejoy's plots, since the fluctuation level derived from the short series is almost one order of magnitude higher than derived from the long series on the scales where the two series overlap. They cannot both be correct, and it is reasonable to assume that the short series give the correct fluctuation on the short time scales around 10^2 yr where the long series is filtered. If we compare this fluctuation level with the level on scales around 10^3 yr derived from the long records, we observe that they are of similar magnitude. In other words, a critical assessment of these data show that the real fluctuation level on centennial and millennial scales are of similar magnitude, consistent with $H \approx 0$, $\beta \approx 1$.

In Figure 2 we demonstrate this by analyzing the reconstruction by *Marcott* (2013) with the periodogram in a particular way to overcome the gradual smoothing as one goes back in time. For the full record, the variability should be trusted only for time scales longer than 2000 years. On shorter time scales, the power is artificially low due to the smoothing. To overcome the smoothing problem, the time series was divided into segments of 2^n * 400 years, with n=0, 1, 2....5 and starting with the most recent period. Segment number: 1=50-450 yr BP, 2=50-850 yr BP, 3= 50-1650 yr BP, 4=50-3250 yr BP, 5=50-6450 yr BP, 6=50-11290 yr BP (longest possible record, shorter than $2^5 * 400$). The periodogram was estimated for each segment, and then a new power spectrum was created using only parts of each segment assumed to be trustworthy with regard to preserved variability. All parts of segment 1 were included, while for segment 2-6 only the low-frequency parts were included (none overlapping). By this composition, the resulting power spectrum represents the variability on all time scales more correctly. The estimated spectrum displays only one scaling regime with $\beta \approx 1.3$, while the spectrum of the full, raw time series exhibits a scale break and $\beta > 2$ in the regime of scales longer than a century, similar to what Lovejoy finds using the Haar fluctuation function.

The point of including Fig. 4b in the review report is hard to understand. The only curve shown here that exhibits a break around 10^2 yr is the mean of a number of ice core records for 10-90 kyr BP. We have never disputed the existence of such a break in the glacial state.

1.6 The issue of scaling – the effect of trends and oscillatory modes

The Haar fluctuation has the advantage with respect to the standard structure function of not going flat for processes with decreasing fluctuations versus scale, and hence works for β both larger and smaller than unity. In Fig. 3a of Lovejoy's report the Haar fluctuation is plotted for the instrumental temperatures and for two multiproxies for the period 1500-1979 AD. All the curves show a break, but at different time scales. The instrumental curve breaks close to 10 yr, while the multiproxies break closer to 50 yr. The slopes after the break are close to those typical for a signal dominated by a trend. For the instrumental record the fluctuation function for the linearly detrended record shows an oscillation after the 10-yr break. This is easy to interpret by noting that the trend in the anthropocene is not linear – a quadratic trend is much more representative of the anthropogenic forcing.

In Figure 3a we show the Haar fluctuation for the full instrumental record, and for the linearly and quadratic detrended record. For the latter, the scale break disappears completely, suggesting that the internal variability follows the same scaling as on the smaller time scales. In Figure 3b we demonstrate the same feature for the Central England Temperature, but for this 350 yr record a linear detrending is suffcient to restore scaling in the Haar structure function for all scales from months to the length of the record.

We agree with Lovejoy that the anthropogenic warming destroys the scaling, but our perception is that it is misleading to think about that as a new scaling regime characterised by a scaling exponent. In the revised manuscript we elaborate on this. We demonstrate that the observations are consistent with both a two-scaling regime model and a one-scaling regime + trend model, but that the latter constitutes a "better" statistical model because it yields much lower errors (uncertainty) on the long time scales. It also has the advantage that part of the model (the trend) makes use of existing knowledge and solid physical understanding. In our opinion statistical modelling should be reserved for those phenomena that cannot be described by simple deterministic models. This is, for example, the idea behind regression analysis; those parts of the variability which can be "explained" by deterministic predictors should be described as such, and in traditional regression the residual is often

assumed to be a Gaussian white noise. In our opinion the main goal of scaling analysis of climatic time series is to establish the true nature of this residual. As will be shown below is that the fractional Gaussian noise is a good model for this residual for Holocene surface temperatures.

In Figure 4 we compute structure functions (empirical moments of order $q = 1, 2, \ldots$) for the global mean surface temperature (GMST) and for its cumulative sum. Structure functions (SF) for the signal itself should be flat for a process where the fluctuations do not increase with increasing scale ($\beta < 1$ if the process is monofractal). For the GMST the high-order SFs are not straight lines in a log-log plot, but curve upwards for scales larger than 10 yr. The reason for this is the trend, which also causes the break in the Haar fluctuation at this scale. However, the collection of SFs for different q contain additional information to the second order statistics expressed by the Haar fluctuation curve, and is more clearly exposed for the detrended signal, as shown in Figure 4b. These SFs are almost flat, but exhibit two peaks corresponding to the annual cycle and another cycle of period of about 20 yr. If there had been an underlying scaling with $\beta > 1$ ($H > 0$) this scaling would have dominated the structure functions and given straight lines with positive slopes $\zeta(q) \equiv \log S_q(\tau)/\log(\tau) = Hq$.

In order to test if the GMST is a monofractal noise process ($H < 0$, $\beta < 1$), we form its cumulative sum (cumsum), which then should be a self-similar process with $H \to H + 1$, $\beta \to \beta + 2$. For scales < 10 yr the structure functions behave as for a monofractal, while for larger scales they bend over. If the scaling function $\zeta(q)$ is computed from the slopes of the SFs up to the 10-yr scale, it is a straight line with a slope (Hurst exponent) $H_u = H + 1 = 0.97$, corresponding to $H = 0.97 - 1 = -0.03$ ($\beta = 0.94$). This is shown in Figure 5a. However, the value of H_u and β close to unity is a typical signature of a trend-dominated process. If we subtract the quadratic trend, and repeat the analysis, the cumsum SFs are still straight and so is the scaling function. But now the Hurst exponent is reduced to $H_u \approx 0.85$ ($\beta \approx 0.70$), as shown in Figure 5b. This is the "true" scaling exponent of the natural variability.

The curving of the SFs for large scales is probably a consequence of the 20 yr oscillation and another oscillation of period around 70 yr. In the second-order Haar fluctuation analysis these oscillations are indistinguishable from true monofractal scaling. Note that it is not the Haar wavelet itself that is the limitation, but the restriction to using only the *second-order* Haar structure function. By using higher-order statistics we reveal the non-scaling nature of some of the fluctuations. By employing the same methodology to the CET and the proxy data, we will observe that the scale-breaks that appear in the Haar fluctuation are due to such non-scaling fluctuations. We will include such evidence in the revised Supplement.

1.7 The issue of scaling – forced versus internal variability

Similar reasoning as presented above pertains also to the multiproxy fluctuations. The periods 1500-1979 AD and 1500-1900 AD are dominated by the warming as the Earth came out of the Little Ice Age (LIA), and the scale break will disappear with detrending. It may be argued that removal of fluctuations by detrending is unjustified, but that depends on whether we prefer to consider scaling as a property of internal climate variability or as a property of the total forced climate signal.

In modern climate science the various natural and anthropogenic drivers of climate variability have been quantified in the form of time series and makes it possible to separate the internal from the forced climate signal based on linear models for the forced response (*Rypdal and Rypdal*, 2014; *Rypdal et al.*, 2015). In General Circulation Models (GCMs) the internal variability can be studied in unforced control simulations, and the scaling of internal and forced temperature fluctuations can be compared (*Østvand et al.*, 2014). The cited studies conclude that the low temperatures during the LIA can be attributed to a combination of volcanic and solar forcing, and hence a detrending of the multiproxy signals is justified if one wants to get closer to the signal representing the internal variability. We conclude that the scale breaks observed in Lovejoy's Fig. 3a arise from the forcing, and is not a property of the internal climate variability. The scale break in the instrumental series is obviously associated with a unique event, the industrial revolution, and this may also be the case with the multiproxy records and the LIA. In order to clarify how unique the fluctuations

like the LIA is, we need to analyse considerably longer time series. This is what we have done in the paper. In some records we have found indications of higher power on low frequencies than consistent with an fGn null hypothesis, but this has disappeared when the forced variability has been separated out.

1.8 The issue of scaling – no intermittency in Holocene temperatures!

Let us first agree with the reviewer that the use of the term "monoscaling" and "multiscaling" towards the end of the paper were misnomers, although it should be clear from the context that what we mean here is "single scaling regime" and "multiple scaling regimes." In his section on "the issue of scaling" Lovejoy presents a lengthy introduction to intermittency and multifractals and end up with the claim that our "restriction to nonintermittent models is unnecessary and unrealistic." And further: "The monofractality - or lack of intermittency - must be quantitatively established not simply assumed a priori."

Since our focus is on Holocene data, we shall establish monofractality of such data here. The evidence presented by Lovejoy for his claim of intermittency is non-Gaussian tails of some PDFs for differences ΔT for time lags $\Delta t = 1, 4, 16, 64$ derived from a multiproxy record. The tails allegedly have the power-law form ΔT^{-5}. This "heavy" tail is in practice indistinguishable from an exponential, and the existence of such tails alone is not a signature of intermittency/multifractality.

For the instrumental global mean temperature the Gaussianity and monofractal scaling was established by *Rypdal and Rypdal* (2010). We refer to this paper for details of the analysis. Since the instrumental data covers a quite limited range of time scales we shall show a similar analysis for the the GRIP ice core δ^{18}O series for the Holocene and the Moberg Northern Hemisphere multiproxy temperature reconstruction. In Figure 6a we show the PDF of the GRIP δ^{18}O anomaly. It is slightly skewed, and a negative tail slightly heavier than a Gaussian. Panel (b) shows a so-called Quantile-Quantile plot, where the quantiles of the data is plotted against those of a normal distribution. The non-Gaussian negative tail shows up as the deviation from the dashed line in the left part of the plot. In panel (c) and (d) we make similar plot as made by Lovejoy, i.e., we plot the probability that $|\Delta(\delta^{18}\mathrm{O})|$ exceeds

a threshold $(\delta^{18}O)_{th}$. Panel (c) is a log-plot and shows that the tail is close to exponential. Panel (d) is a log-log plot and shows that the tail is not a power-law.

A multifractal analysis and test of intermittency can be done by computing structure functions to high order and plot the associated scaling function. Figure 7a shows structure functions and Figure 7b the corresponding scaling function for the GRIP data. The structure functions are straight lines in a log-log plot up to scales of 2000 kyr. The depletion for the highest structure functions beyond that scale is a signature of insufficient statistics on these scales. This is about 1/5 of the total length of the data record, and illustrates our claim that we cannot faithfully estimate scaling properties on scales longer than this. The record is monofractal if the scaling function is close to a straight line. The scaling function shown here indicates that the intermittency is very weak. The main reason for the heavy negative tail is the 8.2 kyr event discussed in the paper. In Figure 8 we show QQ-plots and tail PDFs for the same time record, but truncated at 7.5 kyr BP. Exclusion of the event creates a PDF very close to Gaussian. The remnant of a negative tail still present is probably caused by the forcing from volcanic eruptions. Figures 9 and 10 show similar results for the Moberg multiproxy record, which turns out to be even less intermittent. Instead of $\delta^{18}O$ it is here referred to $|\Delta T|$ exceeding a threshold T_{th}. The conclusion is that the fractional Gaussian noise is a *very* accurate model for Holocene temperatures up to the time scales where the structure function plots start to deviate from straight lines. For the Moberg record this maximal scale is around 400 yr. Beyond this time scale we cannot conclude anything about the scaling from this record with statistical confidence.

As a contrast we show in Figure 11 and 12 a structure function analysis of the GRIP $\delta^{18}O$ record for the last glacial maximum. For scales longer than a few decades (on smaller scales the record is smooth due to interpolation), the second-order structure function suggests mono-scaling with $h \approx 0.3$ ($\beta \approx 1.6$) as shown in Figure 11. However, the higher-order structure functions shown in Figure 12b are not straight in the log-log plot for scales longer than a few decades. If we ignore that and fit straight lines to the SFs in two scale regimes as shown in the figure, we find scaling functions that look multifractal, but this shouldn't lead us to believe that these data can be modelled as a simple multifractal process. The skewness

of the PDF shown in Figure 12a, and the curved SF suggest that this record requires more complex modelling.

2 Uncritical treatment of paleoseries?

The selection of multiproxy series in our paper is far from uncritical, but was a careful assessment based on Lovejoy's previous works. The Marcott reconstructions are interesting and should be discussed in the paper, but as discussed in section 1.5, we believe that the reviewer's analysis of these series are flawed.

Lovejoy is probably right in that the multicentennial variability of the multiproxy records is not a scientifically settled issue, but that only adds to the statistical uncertainty of low-frequency scaling which is the main proposition of our paper.

2.1 Greenland Holocene ice core records and the Berner SST reconstruction

For the Holocene the Greenland record behaves similar to instrumental records for continental interiors. These are characterised by low persistence ($\beta \approx 0$). The Berner reconstruction is similar to instrumental SST-records, which are strongly persistent and sometimes with $\beta > 1$. This does not mean that the Greenland record is more "exceptional" than Berner SST. Both are local temperature proxies, but since 70% of the Earth surface is ocean, the SST is more representative for the global temperature. The Haar fluctuation analysis of the Berner record for the Holocene shown in Lovejoy's Fig. 4b does not show a clear scale break, but a rather flat Haar fluctuation spectrum, corresponding to $\beta \approx 1$. Hence this record seems to confirm our conclusions, rather than refuting them.

It seems rather odd that Lovejoy now argues so strongly that the Greenland records are useless for scaling assessments, considering that he has used them extensively in the past for this purpose, but then using records that span mostly the glacial period. We conclude from this that he now admits that the scale-break hypothesis has been based on non-representative data, but that he believes more recent data are more representative and support the hypothesis.

3 Additional technical points

3.1 Statistical testing and significance

The reviewer's comments on this issue is consistent with the fact that error bars are absent in the figures he presents in his review report, and generally throughout his papers. Error analysis is very important because, if done properly, it forces us to be precise on (i) which estimator we use for hypothesis testing, (ii) which null hypothesis we choose, and (iii) which (alternative) hypothesis we want to test. Without these elements, the concept of statistical significance has no precise meaning.

(i) *Which estimator do we use, and why?*

In this paper we have chosen to estimate the power spectral density (PSD) by means of the periodogram. We have chosen to estimate the PSD because it is widely used and does not eliminate trends. It also works equally well for processes with fluctuations growing ($\beta > 1$) and decreasing ($\beta < 1$) as scale increases. The Haar wavelet has many of the same properties, but as far as we know, neither the periodogram nor the Haar wavelet has been tested for biases and errors as we did with the DFA and the Mexican hat wavelet in *Rypdal et al.* (2013). This should be done, but there is little reason to believe that biases and errors for either are very different from those we have tested. DFA(n) with $n \geq 2$ is insensitive to a linear trend, which means that for an ensemble of fGns with a linear trend superposed it will give an unbiased estimate of the scaling exponent for the underlying fGn. In this respect it performs similar to the Mexican hat wavelet, which eliminates linear trends because it is a symmetric wavelet. Estimators that eliminate a linear trend effectively uses the shorter time scales for the estimation of the scaling exponent. Therefore they are not suitable for detecting a scale break or a trend. This is why we have not used such an estimator in our analysis, and the reviewer's comments on DFA and trend detection is therefore completely beside the point. Whether we use the periodogram or the Haar wavelet in the analysis is probably unimportant. The important thing is to estimate the error bars for the estimator

under a null hypothesis, as we do for the periodogram by the blue-shaded areas in all figures where we show power spectra.

A principal point is that there is no such thing as a "correct" or "incorrect" estimator, but some are more sensitive than others when it comes to detctection of particular features. If an estimator fails to reject the null hypothesis, it just means that the alternative hypothesis is insignificant with respect to this particular estimator. If we are able to find another estimator that allows us to reject the null hypothesis from the data, while the alternative hypothesis is not rejected, then the alternative is significant under the chosen null.

From these considerations one might be lead to search for estimators that are particularly sensitive to the large scales, such as the Haar wavelet. The problem, however, is that such estimators are also particularly sensitive to the large scales in the random noise that constitutes the null hypothesis, giving rise to large error bars at these scales. This is exactly what is observed is our spectra, where the error bars are much wider at the low frequencies.

(ii) *The choice of null hypothesis, and why?*

The null hypothesis cannot be chosen subjectively to the same extent as the estimator, since the null should represent a plausible explanation of the observed data provided the alternative hypothesis is false. In section 1.8 we demonstrated that a Gaussian monoscaling process with $\beta < 1$ is the appropriate null model for Holocene temperatures.

(ii) *The choice of alternative hypothesis*

One of the most serious problems in Lovejoy's reasoning is that there is no clear hypothesis, i.e., no quantitative model for the large-scale fluctuations. The notion of a "break in scaling" is used, and figures are presented that give the reader the idea that there is a regime for large scales that can be characterised by a scaling exponent ($\beta > 1$) which is larger than for the short scales. This interpretation is underscored by straight lines drawn in the log-log plots of the Haar fluctuation for the large scales. In the text of his review report, however, he

downplays the significance of self-similar scaling on the large scales and reduces the issue to a question of the existence of large fluctuations on these scales, regardless of the cause and nature of these fluctuations. They could just as well be trends caused by specific, well understood forcings, or well established internal oscillations – all fluctuations that do not exhibit scaling.

In our paper we draw a distinction between these two alternative hypotheses. In the figures where we compute periodograms with the blue-shaded error fields (red-shaded for the Moberg record in Fig. 2c) we just compute the 95% error bars corresponding to the fGn null hypothesis, and compare it with the estimated spectrum from the actual observed record. For the Holocene records, the fGn hypothesis for scales up to a millennium cannot clearly be rejected by the observations we have considered. But this is not a very strong result. The difference between the Medieval Warm Anomaly (MWA) and the Little Ice Age (LIA) is so large in the Moberg record that it is not a very likely outcome of an fGn-fluctuation. But it is very explicable as a result of a combination of volcanic and solar forcing (volcanic more important than solar). This was shown in Figures 2b and 3c of *Østvand et al.* (2014), where the residual after subtracting the forced response was clearly within the confidence range of the fGn. Thus, the fGn null for *internal* variability is not rejected by the Moberg record.

Our main focus, however, is on the hypothesis of a low-frequency scaling regime with $\beta > 1$ in the Holocene. We then have to estimate β from the few independent measurements we have for those scales. The result for the Moberg record is shown in Figure 2b, and for the other multiproxies in Figure 3 and Table 1. Supposing that the scale break is at 100 yr, we have effectively $N/100$ independent observations for estimation of the scaling exponent of this regime, if the resolution is annual and N is the length of the record measured in in years. For the CET record we then have 4 independent measurements. For the Moberg record we have 20, and for the Marcott reconstruction 100. If we could trust that the Marcott reconstruction has effective resolution < 100 yr, it would have been the best candidate for estimation of the scaling exponent of the low-frequency regime. But we know that this reconstruction suppresses variability on scales less than 2000 yr.

3.2 The "rule of the thumb"

The "ill-starred rule of the thumb" of not drawing conclusions about scaling properties for scales corresponding to more than one fourth of the length of the sample at hand, is not particular to the DFA estimator. For the large scales (compared to the length of the sample record) the uncertainty is large for any estimator, and if we use these scales when estimating scaling exponents from single realisations in the MC ensemble, we end up with large uncertainties in the estimated exponents. In other words, the rule of the thumb is chosen such that the estimate of the scaling exponent from a realisation of a known monoscaling process is reasonably well confined. If we use all available scales for this estimate, the estimate is so uncertain that it gives no useful information about the scaling exponent of the underlying process, even if it is perfectly monoscaling. The reviewer writes about the significance of "events that are four or five standard deviations from the mean." But that is not what we observe in Holocene climate. In the spectra we hardly find events that are two standard deviations from the mean.

3.3 Mexican Hat and Morlet wavelet vs. Haar wavelet

The reviewer's comments about our use of wavelets are also beside the point. We don't use the wavelets for scaling analysis, only for demonstrating the influence of the 8.2 kyr event on the power spectra. We have decided to use the term "scalogram" instead of "wavelet power spectrum" from now on and in the revised paper. The Mexican hat and Morlet wavelets are similar to local Fourier transforms and hence their scalograms are suitable as local supplements to the periodograms. A known weakness of the Haar wavelet is stronger spectral leakage (see e.g., textbook of Percival and Walden, 2008), which is not so important for scaling analysis, but a drawback if we want to study the effect of local events on spectra. The Mexican hat and Morlet are also more sensitive to oscillations, and can be used to detect those in the noise. In this respect the Haar wavelet (there are different versions) performs more similar to the DFA.

Lovejoy has developed an algorithm for using the Haar wavelet on unevenly spaced data. This algorithm is explained in appendix A of his 2014b paper. We agree that linear interpolation is not optimal, but in our case it is not a problem. Our wavelet scalograms show the lower white curve where interpolation has an effect. We are interested in power increases on centennial time scales, and interpolation does not affect these time scales.

4 Concluding remarks

Lovejoy's review contains no concrete suggestions for revision of our paper. We cannot agree with his assertion in his section (a) that the paper is particularly technical, and that these technicalities obscure the real issue. On the contrary, some of his comments on technical points in his section (c) are largely beside the point and only serve to obscure the issue. Other comments in this section are at odds with sound approaches to statistical hypothesis testing.

Our "uncritical use of paleoseries" claimed in his section (b) is the same use as made by Lovejoy himself in quite recent papers, but we shall happily include a critical analysis of the Holocene multiproxies of *Marcott* (2013) in a revised paper.

In spite of our disagreements, the review has been helpful in bringing the nature of these to the forefront, and we will try to incorporate the essence of this response document into the paper. Some text and a few figures will be included in the main paper, and further text and figures in the Supplementary Material.

Acknowledgements. This work was funded by project no. 229754 under the the Norwegian Research Council KLIMAFORSK programme. We acknowledge in-depth discussions with Martin Rypdal during the work with this response, and also have made the use of routines for generating Monte Carlo ensembles of fractional Gaussian noises and for structure function and scaling functions produced by Martin Rypdal and Ola Løvsletten.

References

Berner, K.S., Koc, N., Divine, D., Godtliebsen, F., and Moros, M.: A decadal-scale Holocene sea surface temperature record from the subpolar North Atlantic constructed using diatoms and statistics and its relation to other climate parameters, Paleoceanography, 23, 10.1029/2006PA001339, 2008.

Budyko, M. I..: The effect of solar radiation variations on the climate of the Earth, Tellus, 21, 611-619, 1969. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, J. Geophys. Res, 111, D12 106, 2006.

Marcott, S. A., Shakun, J. D., Clark, P. U., and Mix, A. C.: A reconstruction of the regional and global temperature for the past 11.300 years, Science, 339, 1198, doi: 10.1126/science.1228026 (2013). Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, J. Geophys. Res, 111, D12 106, 2006.

Rypdal, M., and Rypdal, K.: Testing Hypotheses about Sun-Climate Complexity Linking, Phys. Rev. Lett. 104, 128501, doi: 10.1103/PhysRevLett.104.12850, 2010.

Rypdal, K., Østvand, L., and Rypdal, M.: Long-range memory in Earth's surface temperature on time scales from months to centuries, J. Geophys. Res, 118, 7046-7062, doi:10.1002/jgrd.50399, 2013.

Rypdal, M., and Rypdal, K.: Long-memory effects in linear-response models of Earth's temperature and implications for future global warming, J. Climate, 27, 5240-5258, doi:10.1175/JCLI-D-13-00296.1, 2014.

Rypdal, K., Rypdal, M., and H.-B. Fredriksen: Spatiotemporal Long-Range Persistence in Earth's Temperature Field: Analysis of Stochastic-Diffusive Energy Balance Models, (in press) J. Climate, 2015.

Østvand, L., Nilsen, T., Rypdal, K., Divine, D., and Rypdal, M.: Long-range memory in internal and forced dynamics of millennium-long climate model simulations, Earth. Syst. Dynam., 5, 295-308, doi:10.5194/esd-5-295-2014, 2014.

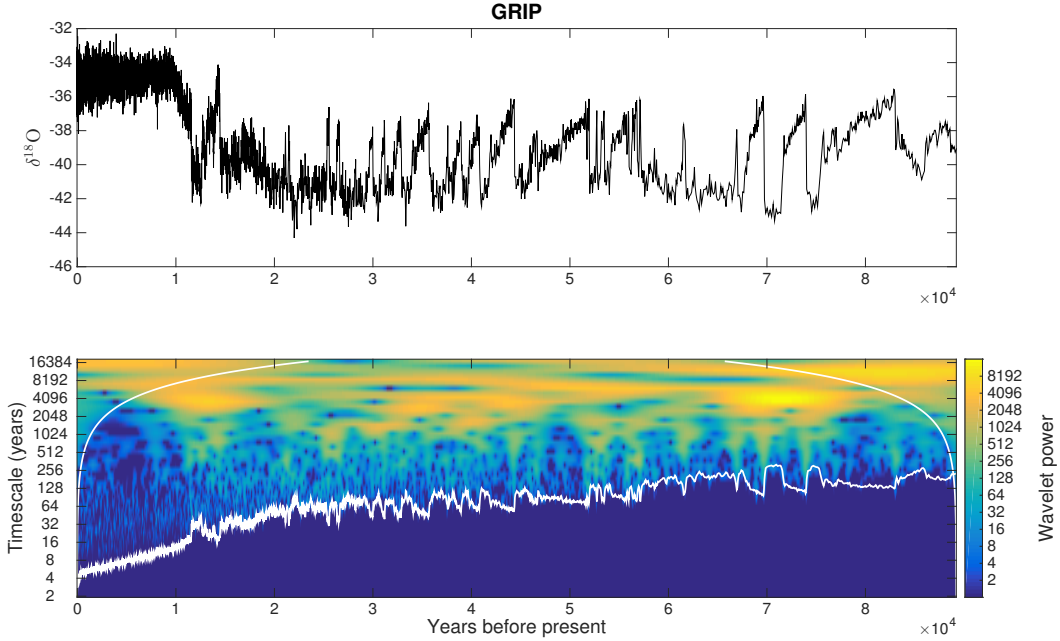

Figure 1. Upper panel: The δ^{18}O proxy rime series for Greenland temperature from the GRIP ice core for the period 0 – 90 kyr BP. Lower panel: The Morlet wavelet scalogram for the signal in the upper panel.

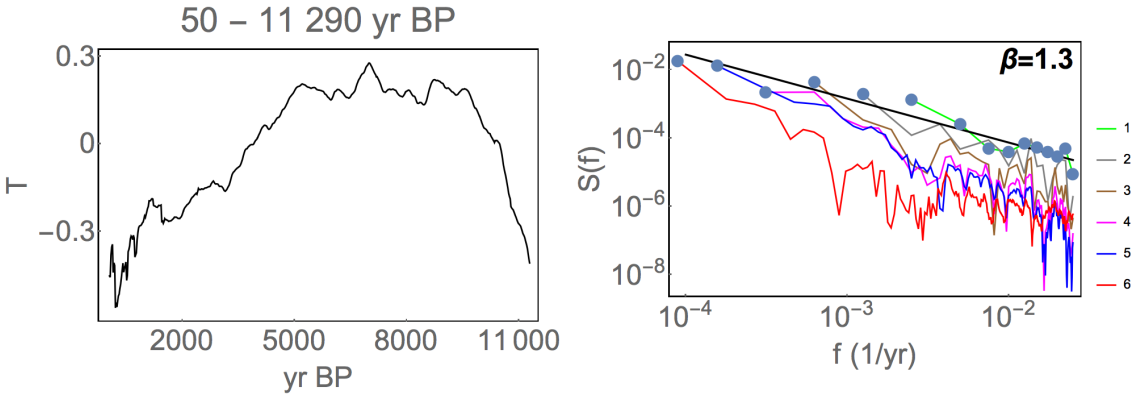

Figure 2. (a): The Marcott global multi proxy reconstruction. (b): Periodograms of time series in (a) by dividing into segments of 2^n * 400 years, with n=0, 1, 2....5 and starting with the most recent period. Segment number: 1=50-450 yr BP, 2=50-850 yr BP, 3= 50-1650 yr BP, 4=50-3250 yr BP, 5=50-6450 yr BP, 6=50-1 290 yr BP (longest possible record, shorter than $2^5 * 400$). The periodogram is estimated for each segment, and then a new power spectrum created using only parts of each segment assumed to be trustworthy with regard to preserved variability (the blue dots). All parts of section 1 are included, while for section 2-6 only the low-frequency parts are included (none overlapping).

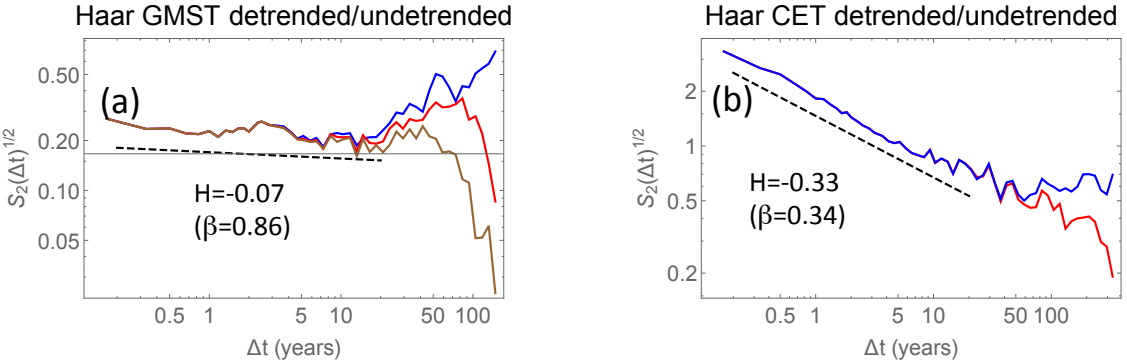

Figure 3. (a): Haar fluctuation for GMST. Undetrended (blue), linearly detrended (red), and quadratically detrended (brown). (b): Haar fluctuation for CET. Undetrended (blue), linearly detrended (red).

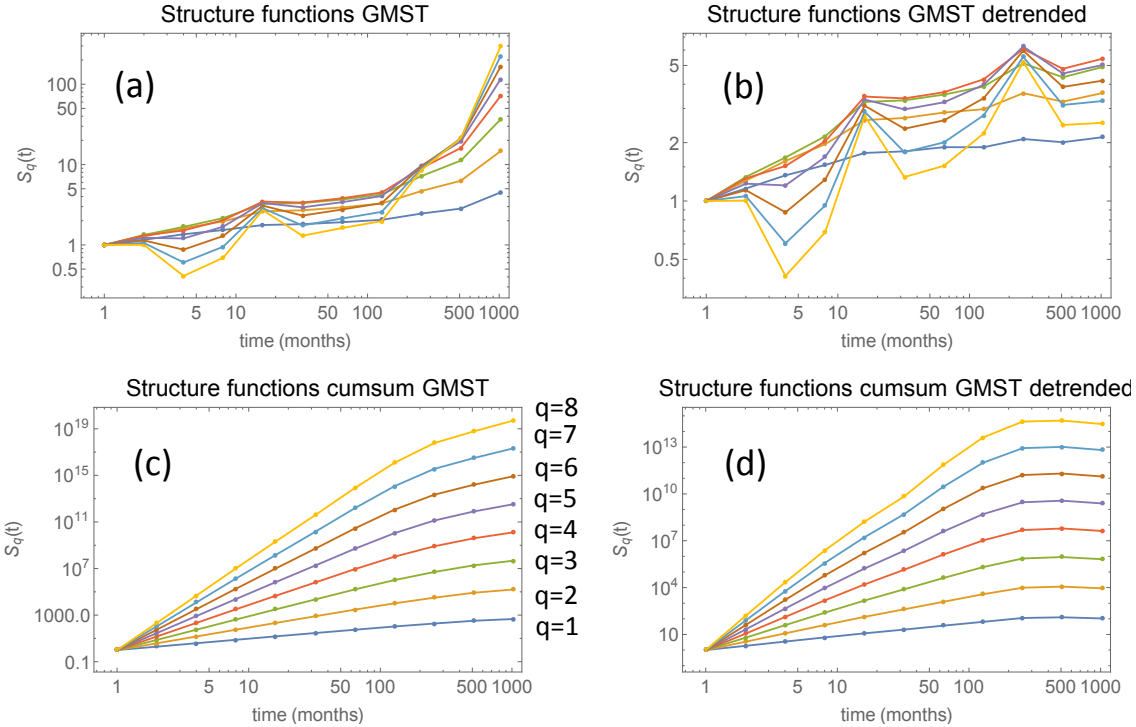

Figure 4. (a): Structure function estimates (empirical moments) $S_q(\tau) = (N-\tau)^{-1} \sum_{i=1}^{N-\tau} |T(t_i + \tau) - T(t_i)|^q$ for the GMST (HadCrut3) monthly record 1880-2010; $T(t_i)$; $i = 1, \ldots, N$. (b): Structure function for the quadratically detrended GMST (residual after subtraction of a second-order polynomial fit). (c): Structure function for the cumulative sum $y_{t_i} = \sum_{j=1}^{i} T(t_i)$. (d): Structure function for the cumulative sum of the quadratically detrended GMST.

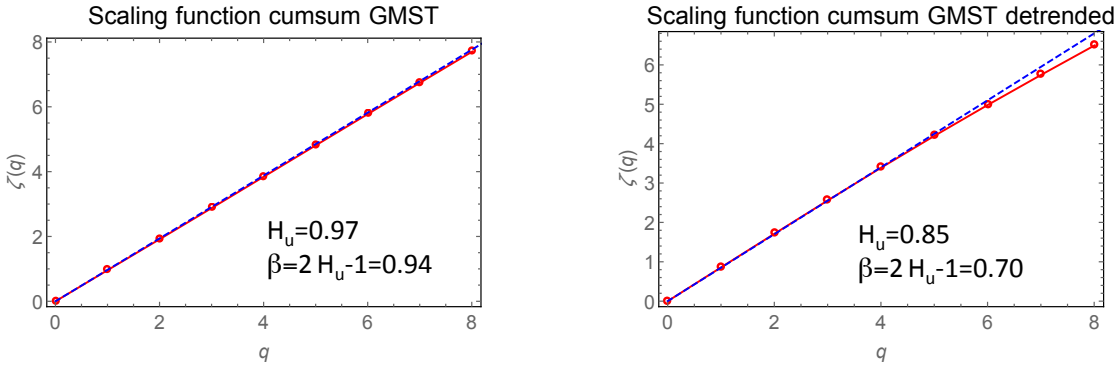

Figure 5. (a): The scaling function $\zeta(q)$ defined through the relation $S_q(\tau) = \tau^{\zeta(q)}$ for the cumulative sum (i.e., as the slope of the $\log S_q(\tau)$ vs. $\log \tau$). The slopes have been computed from the structure functions in Fig. 4c. (b): The same as in (a), but for the cumulative sum of the quadratically detrended GMST.

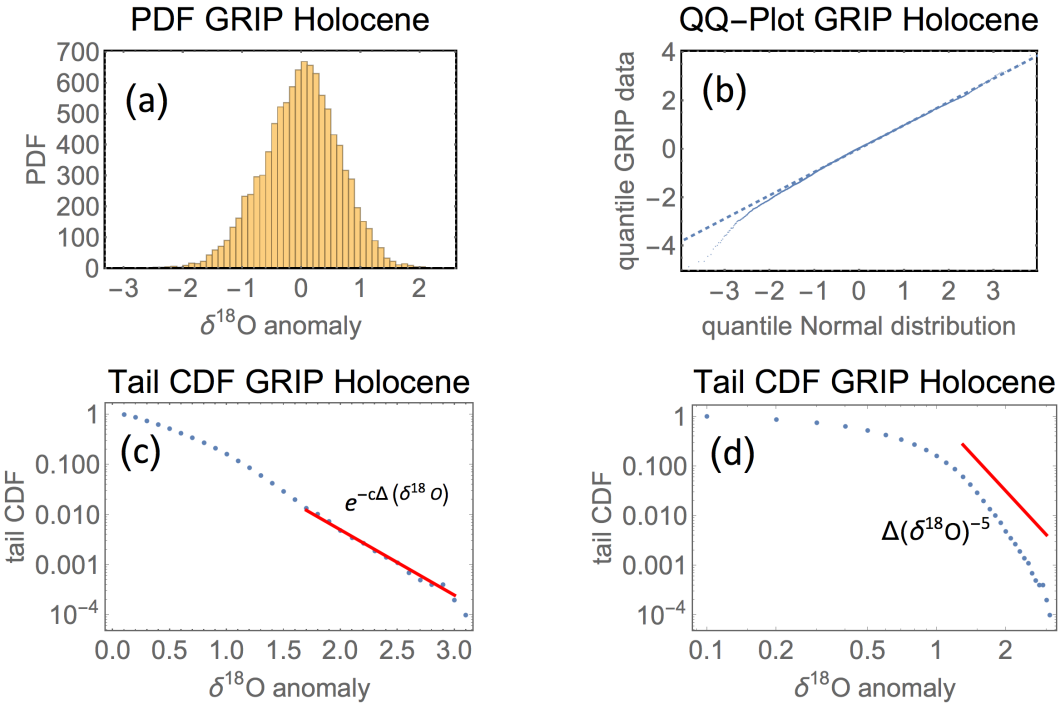

Figure 6. (a): Probability Density Function (PDF) of GRIP temperature anomaly for Holocene data 0 – 10500 yr BP. QQ-plot for these data. (c): Tail Cumulative Probability Function (probability for $\Delta(\delta^{18}O)$> threshold) for the data in log-plot. (d): Same as (c) in a log-log plot.

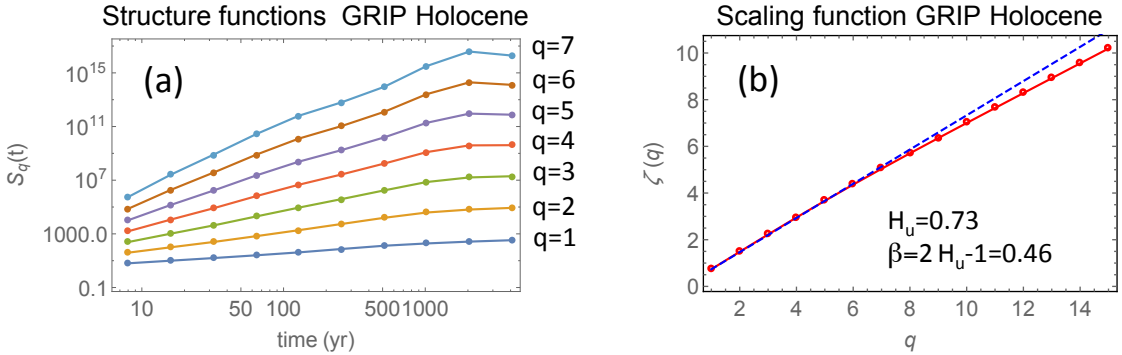

Figure 7. (a): Empirical moments (structure function estimates) for the GRIP Holocene data for moment orders $q = 1, \ldots, 7$. (b): The corresponding scaling function estimate $\zeta(q)$ for $q = 1, \ldots, 15$ estimated as the slope of the empirical moment curves in the range 8-1024 yr. The dashed line has slope $H_u = 0.73$ (Hurst exponent), which corresponds to the spectral index $\beta = 2H_u - 1 \approx 0.46$.

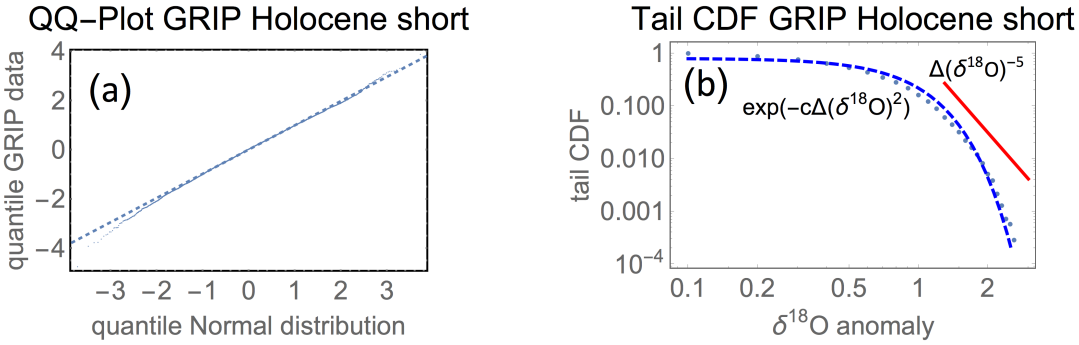

Figure 8. Analysis of the GRIP Holocene data in the range 0 – 7500 yr, i.e., without the 8.2 kyr event. (a): The QQ-plot. (b): The tail CDF (bullets) and the fitted Gaussian for these data (dashed).

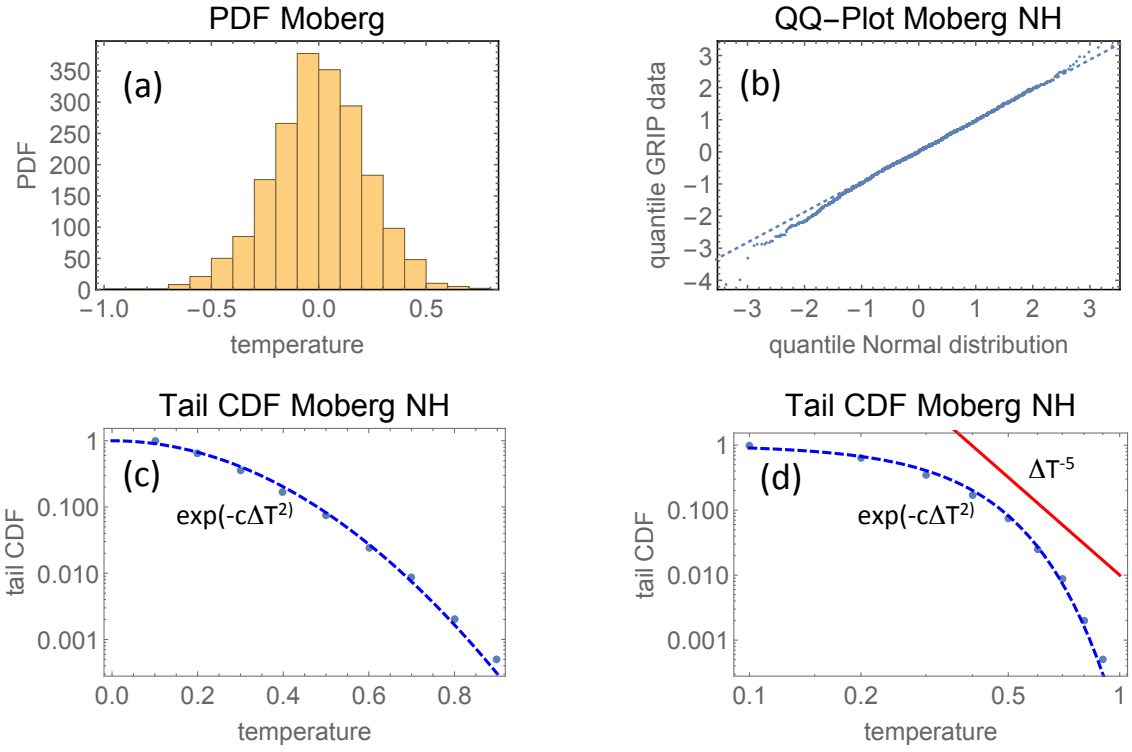

Figure 9. (a): PDF for the Moberg Northern Hemisphere multiproxy data. (b) QQ-plot for these data. (c): The tail CDF (bullets) and the fitted Gaussian (dashed) in a log-plot. (d): The same as in (c) in a log-log plot. The red line is a plot of the power-law function ΔT^{-5}.

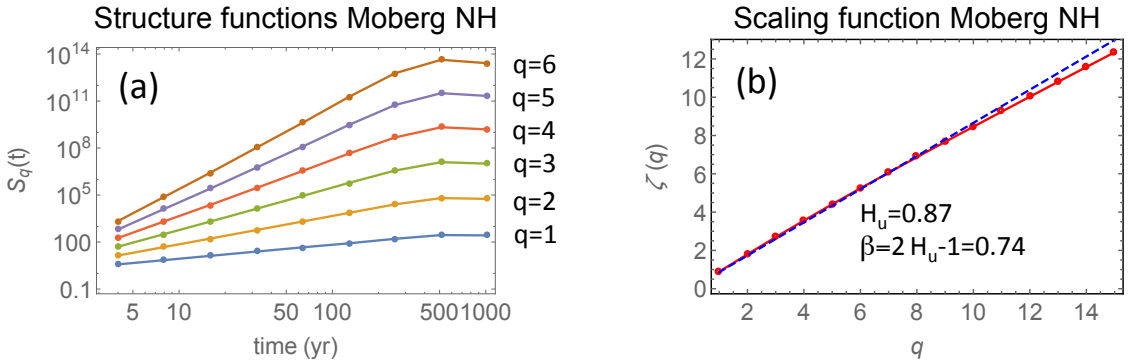

Figure 10. (a): Empirical moments for the Moberg Northern Henisphere data for moment orders $q = 1, \ldots, 6$. (b): The corresponding $\zeta(q)$ for $q = 1, \ldots, 15$ estimated as the slope of the empirical moment curves in the range 4-256 yr. The dashed line has slope $H_u = 0.87$, which corresponds to $\beta = 2H_u - 1 \approx 0.74$.

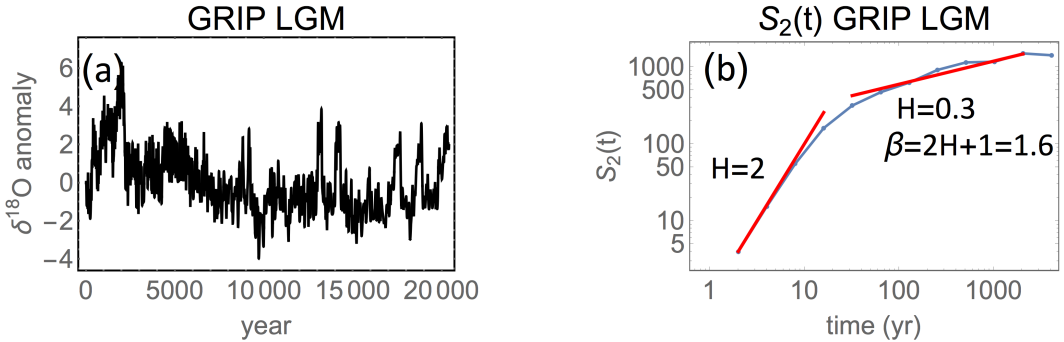

Figure 11. (a): δ^{18}O anomaly time series for 20 kyr of the Last Glacial Maximum (LGM) (b): The second order structure function $S_2(t)$ for this time series (not of its cumulative sum). The steep slope on short time scales up to ≈ 30 yr is due to the smoothness from interpolation, i.e., the actual time resolution is not better than a few decades. The curve is not completely straight in the range above 30 yr, but a slope computed on the scales $32 < t < 256$ yr is $H \approx 0.3$. The corresponding Hurst exponent is $H_u = H + 1 = 1.3$, and the spectral exponent is $\beta = 2H_u - 1 = 1.6$.

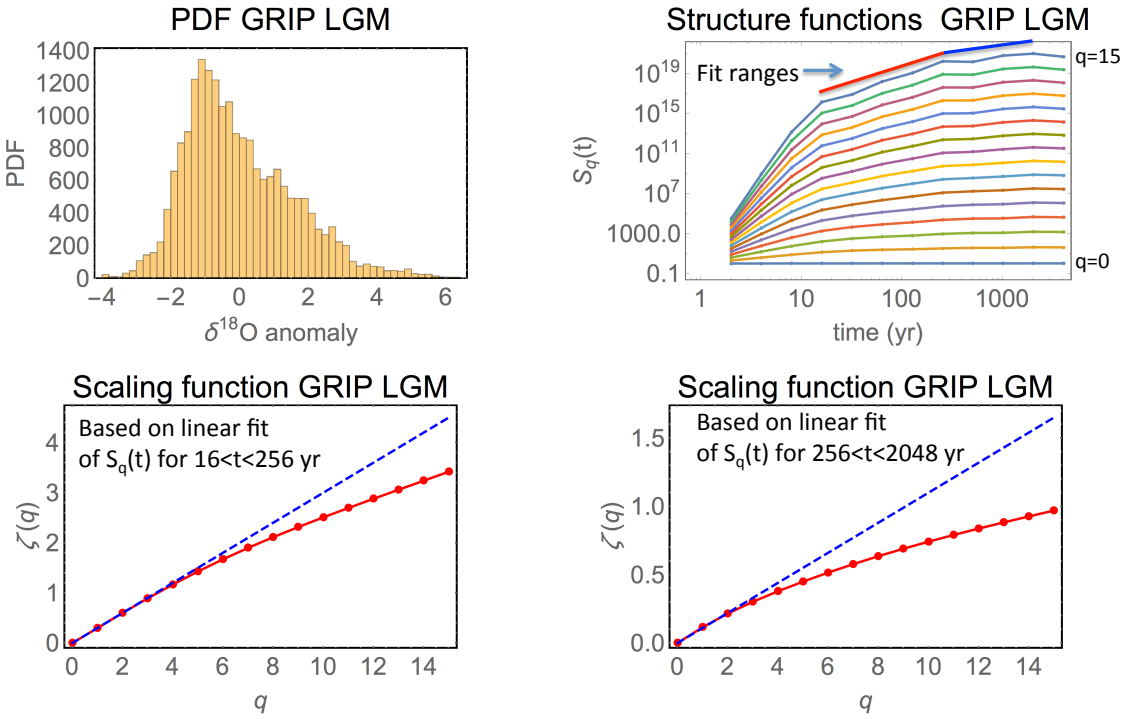

Figure 12. (a): PDF for the LGM data. (b) Empirical moments for the LGM data (not for the cumulative sum) for moment orders $q = 1, \ldots, 15$. (c): The corresponding scaling function $\zeta(q)$ for $q = 1, \ldots, 15$ estimated as the slope of the empirical moment curves in the range 16-256 yr. The dashed line has slope $H = 0.3$, which corresponds to $H_u = H + 1 = 1.3$ and $\beta = (1 + H_u)/2 \approx 1.15$. (d): The $\zeta(q)$ for $q = 1, \ldots, 15$ estimated as the slope of the empirical moment curves in the range 256-2048 yr. The dashed line has slope $H = 0.11$, which corresponds to $H_u = H + 1 = 1.11$ and $\beta = 2H_u - 1 \approx 1.22$.