Earth Syst. Dynam. Discuss., 5, C382–C394, 2014
www.earth-syst-dynam-discuss.net/5/C382/2014/

Earth System Dynamics
Discussions

Interactive comment on "Statistical significance of rising and oscillatory trends in global ocean and land temperature in the past 160 years" *by* L. Østvand et al.

L. Østvand et al.

lene.oestvand@gmail.com

Received and published: 19 August 2014

The review process and how to read this response
Dr. Chandler has had access both to the other reviewer's comment and K. Rypdal's response. Many of the issues are common in the two reviews, hence the reviewers should consider both responses.

The objective of the paper
Since it appears from the reviews that we haven't been able to state the objectives clearly enough, let us state it here:

> The main objective is to assess the significance of a multidecadal oscillation-like variability (MOLV), which appears to have larger amplitude than one can expect from a coloured noise whose parameters are determined from the short-time scale statistics of the observed record. A secondary objective is to quantify the linear trend significance of the global land and ocean data sets under short-range as well as long-range correlated null models. The significance assessment of the linear trend is the motivation for introducing the linear+oscillatory trend model in section 2. Acknowledging the existing overwhelming evidence for an anthropogenic rising trend from physical as well as statistical evaluation of observation data, a model that takes this trend as given is chosen in Section 3 for a more informed assessment of significance of the MOLV.

Authors lack statistics competence and do not cite statistics literature?
Both reviewers have expressed problems with understanding our objectives, but still they claim that these objectives could have achieved by application of standard textbook methods. Chandler even suggests that we are unfamiliar with standard statistics and suggests a textbook for us to read. He claims that the bibliography contains only two statistical references, which is a grave distortion of the facts. Our main statistics textbook reference is von Storch and Zwiers, "Statistical Analysis in Climate Research," which should be an appropriate general reference in ESDD. Furthermore, at least 70% of the references are papers whose content is mainly statistics. We cannot take the blame for the fact that most papers relevant to our objective are published in geoscience journals, not in dedicated statistics journals?

Just another paper reporting the same results with the same methods?
Multiproxy observational records combined with instrumental (the hockey-stick graph) are very compelling when it comes to the rising trend, even without any statistical analysis. The most vocal criticism of this conclusion comes from some dissidents who do not believe in the proxy reconstructions. But there are also a considerable number of

papers that we have cited, who question the statistical significance of this trend in many instrumental data records based on statistical analysis with LRM null models, although most of them do not dispute the reality of the global warming signal. *The physical reality of the signal or trend must not be confused with its statistical significance related to a given data set and a given null-model*. Hence, our paper is *not* "just another paper that comes up with more or less the same result using more or less the same techniques." The facts are that results in the literature diverge significantly and that there is considerable disagreement (see for instance the dispute between Bunde et al. and Bromwhich et al. in Nature Geoscience, **7**, 247 (2014)). Contrary to his own claims, dr. Chandler does not seem well-informed about the full body of literature on long-range memory and climate trends. And if our paper uses "more or less the same techniques" as everybody else, why do the reviewers complain that the techniques we apply are so disturbingly "non-standard?"

The analysis is not necessary since it only tells us what we know already?
Reviewer #3 points out that climate models also present very compelling evidence of this rising trend, and we agree. The ensemble mean signal of multimodel ensembles could justify to treat the rising trend as an established fact and just treat the significance of the "oscillatory trend," as we do in Section 3. Since the reality and magnitude of the global warming signal is a well-established fact the objective of our analysis is *not* to establish the reality of the global warming signal. When it comes to linear trends the main purpose of our analysis is to establish the difference in significance between land and ocean records, and to point out that significance increases when going from local to global records. These results are not obvious and has to our knowledge not been reported previously. They are not obvious because two features pull in opposite direction: higher short-term variability over land and in local records reduces trend significance, but also contribute to lower β, which increases significance. The vast majority of records studied in the literature so far are local. The reviewer has the impression that we have chosen the global datasets to "maximise the chance of finding a significant trend," and views it as a biased approach that "undermines the credibility of

climate science." Such statements demonstrate a lack of understanding of the nature of the different datasets available for analysis, but also an uncritical use of the "bias" notion. An analysis would be biased if comparable data sets are selected or sampled according to specific criteria. But the data sets studied in this and other papers are not comparable. Natural variability in single local temperature records arises from different physics than regionally averaged and global records, and surface temperature over continents are different from sea-surface temperature records. The global warming signal is present in all records, but may be detectable (significant) in some, but not in others. There is absolutely no "poor statistical practice" in exploring the trend significance of the various data sets and point out which data sets that are optimal for trend detection. A benefit of this analysis is that it is perfectly possible to understand a posteriori why the global warming signal is easier to detect in some data sets than in others, although it would be difficult to state this a priori without a quantitative analysis.

The reality and nature of the oscillatory trend
Dr. Chandler writes: "the existence of a purely deterministic sinusoidal cycle is very dubious." We totally agree, and this is perhaps the gravest misconception of our paper behalf of the reviewer(s). We are not aware that we have written anything like that in the paper, but we realise that we need to motivate and explain the model even more carefully. It seems that our choice of a sinusoid to represent the non-monotonic part of the trend has been perceived by the reviewers as a model for a coherent cycle, which it is not (more about that below). The may also have been influenced by the fact that we have made reference, for completeness, to a couple of papers which speculate on climatic cycles mysteriously synchronised with the motion of giant planets. Such coherent cycles is not what we would like to test for; if that were the objective it would be better to test for coherence, and this has already been done (S. Holm, J. Atmosph. Solar-Terr. Phys., **110-111**, 23-27, 2014) with a clear negative result. There is no significant coherence between the solar gyrocenter motion and global temperature.

Maybe it would make things less confusing if we employ the notion "multidecadal

oscillation-like variability (MOLV)." This terminology has recently been introduced in climatology, where it has become more common to talk about the Atlantic Multidecadal Variability (AMV) rather than the Atlantic Multidecadal Oscillation (AMO). The climate dynamics literature is full of “oscillations" that are thought to stand out of the colored-noise continuum, but actually most of these oscillations is what *forms* this continuum spectrum. There are exceptions, such as ENSO, which creates a broad bump on periods 3-5 yr in the SST-spectrum in the equatorial Pacific. This bump clearly stands out of the underlying power-law spectral continuum, and no-one doubts it's statistical significance. We think of the MOLV visually observed in the global instrumental 150 yr records in a similar manner. If we fit a straight line to the log-log periodogram of this record on time scales up to a decade we will observe that there is more power on time scales from a decade up to the length of the record than suggested by this straight line. If we compute the periodogram of the detrended record we end up with a broad bump around a period of 70 yr. This was demonstrated in a recent paper (Rypdal et al., J. Geophys. Res., **118**, 7046, 2013) where we made a heuristic study of the LRM properties of same data sets as in the present paper. Here we showed also that the ACF-estimate of the undetrended record is way outside the confidence region of the fGn-ensemble, and that the third- order polynomially detrended record exhibits an oscillating ACF whose amplitude exceeds the 95% confidence region for the null ensemble (but only by a small margin). The attached Figure 1 shows a figure from this paper which illustrates these features for the global land temperature. Similar features show up in the 350 yr long Central England Temperature (CET) record (also shown in that paper), but here we also observe an additional spectral peak around a period of 25-30 yr. These two periods are described in climatology literature as the characteristic periods of the AMV (se H. Dijkstras monograph cited in the response to Reviewer #3). b) Significance of the 70 yr time scale variability (the MOLV) can also be thought of as significance of the spectral bump at this period in the detrended record. A test that can be considered as the spectral analog to the ACF and to the test described in Section 3 in the present paper (linear detrending means that we take the linear trend for granted)

can be made by creating a null ensemble of fGns with β equal to the slope of the fit-line in the attached Figure 2(a), and then compute periodograms for each realisation. The ensemble of periodograms allow us to compute 95% confidence intervals for the spectrum. If the spectral bump lies outside this confidence region, the null hypothesis is rejected and we may consider the MOLV significant. The result of this analysis is shown in Figure 2(a), and shows that the bump is located at the edge of the 95% confidence interval. Here the periodogram has been log-binned, which is appropriate if it is used to estimate scaling exponents. In Figure 2(b) we perform a similar analysis with another estimator, the DFA2, which is described in the Rypdal et al. JGR, 2013 paper. All three estimators (ACF, periodogram, DFA2) show that the MOLV appears to be at the margin of the 95% confidence interval for the fGn-null ensemble. Without linear or higher order polynomial detrending, the low-frequency variability is way outside the confidence region for the ACF and periodogram, indicating the clear significance of the rising trend. DFA2 removes linear trends, so for this estimator we do not have to perform a detrending to asses the significance of the MOLV.

From our description above it should be clear that the proposed model for slow variability is not taken out of the blue. There is strong physical support and support based on heuristic analysis of the data for proposing this model. However, since the heuristic analysis shown in the attached figures present marginal confidence for the significance of the MOLV, the issue warrants further investigaton.

A key point here is that if the null model is not rejected by one estimator, it can be rejected by another. It is sufficient that the null is rejected by one estimator to conclude that the observed signal has some property that is is not shared by the null model, and hence that this property is significant. If the observed signal exhibits a non-monotonic multidecadal variability that is not shared by the null-model; then this variability is signficant with respect to this null model.

Rejection of null = acceptance of alternative?
We have the same understanding of these matters as both reviewers. It all boils down

to the precise meaning of the word “acceptance." The way we have used this word is that it is synonymous with “rejection of the null." We don’t believe that rejection of the null implies that the alternative is true. Since this wording seems to cause trouble, we will avoid using the ambiguous word.

The reviewer writes that he does not understand the last paragraph on p. 338, so we need to explain this better: We write that by rejecting the null hypothesis we only reject those aspects of the null model that are relevant for the alternative hypothesis. What we mean is that we are testing only the coefficients (A_1, A_2) for the pseudotrends estimated for the null ensemble, i.e., for the slow variability of the realizations of that ensemble. Hence, we do not test for the short-term variability, so rejection of the null does not mean that we reject that the null can describe correctly the short-term variability. This is an important point, because the short-term variability of the observed record is used to estimate the parameters of the null model. Here MLE is the appropriate estimation method, since it weights the short time scales.

Choice of alternative hypothesis model in Section 2
As explained above our objective is to test if the long-term (multidecadal time scales) variability in the instrumental global record falls within the confidence limits of the LRM-processes constituting the null hypothesis. Since the long-term variability appears to have both a monotonic and a non-monotonic component it is reasonable to chose a model for this variability that contains these two components. We have chosen to study a linear + sinusoidal model, but that does not mean that we believe that the anthropogenic warming is linear and that the non-monotonic variability is coherent and sinusoidal. We could, for instance, have performed an empirical mode decomposition (which we have done) and chosen the two slowest EMD components, which turn out to be approximately linear and sinusoidal, respectively. The exact decomposition of the monotonic and non-monotonic slow components is unimportant, as long as the two components together makes a reasonably good fit to the slow variation of the observed signal. For the same reason the exact choice of the frequency f in our

model is unimportant. Why should we spend energy on estimating a frequency when a monochromatic signal is not what we expect to find? A lot of nonsense has been presented in the literature (e.g., by N. Scafetta) by applying the maximum- entropy method (MEM) which is designed to detect coherent spectral lines in a red-noise background. The problem with this method is that it detects such lines which are not there. We have applied the method to the CET record and find sharp spectral lines. The problem is that we find different lines if we divide the record into two equal segments and analyse each of them separately. Dr. Chandler has got the impression that we belong the “cyclomania community." We cannot see that this impression is justified by our paper and have to ascribe it to some unjustified psychological bias.

What are we testing for?
Dr. Chandler writes: “If you reject the null you have learned that the null seems inconsistent with the data. That’s all." Superficially this statement seems right, but it is actually profoundly misleading (and explains why Chandler didn’t understand what we wrote in the last paragraph on p. 238). What we learn from rejection of the null depends on the nature of the test, more precisely, which qualities of the alternative and the null we compare. A model is never identical to a natural phenomenon, and if the alternative and the null are not identical models, it will always be possible to reject the null by looking up a quality that is not identical in the two models. The quality we compare is the fit of the trend model (linear + 70 yr oscillation) to (i) the observation data and to (ii) the the null-ensemble data, respectively. By this procedure we test the significance of the quality we are interested in, ignoring irrelevant qualities. This is not a circular or self-fulfilling procedure; there is every possibility that the test would fail to reject the null in a different data set, e.g., if the oscillation amplitude estimated from the observation data were a bit smaller, or the β estimated from the short-term variability in the data were a bit higher.

We also disagree with the statement: “. . . the significance of the 70-year oscillation

should strictly be interpreted merely as an indication that the linear model doesn't fit the data." Such a lack of fit can be found without reference to any null hypothesis, e.g., by computing the variance of the residual (linearly detrended) record. What our significance test (displayed in Figure 3 in the paper) shows is that both AR(1) and the fGn/fBm null models are very unlikely to produce an ensemble member with similar trend coefficients (A_1, A_2) as produced by the the observation data. The exception is the fBm null model applied to ocean data, where the significance is marginal.

The test presented in Figure 3 allows us to reject the null model by comparing the combination of two qualities of the record, the linear trend and the oscillation. From this rejection we cannot immediately conclude each of these qualities is significant independent of the other. This is what we attempted to do in Figures 4 and 5, but haven't been sufficiently clear when it comes to explaining the limitations of the results in Figs. 4 and 5 in rejecting the null. Since these figures present one-dimensional PDFs obtained by integration over the joint PDF for $(\hat{A}_1, \hat{A}_2)$ we end up with comparing only one quality (e.g., $\hat{A}_{2,obs}$) of the alternative model with one quality $\hat{A}_2$ of the null model, while in Figure 3 we compare both qualities A_1, A_2. Since $\hat{A}_1$ and $\hat{A}_2$ are dependent random variables, Figure 3 represents a stronger test, i.e., there is a greater chance of rejecting the null model. Using Figures 4(b,d,f) and Figure 5(b,d) to test the significance of A_2 of the oscillation therefore represents a test that is weaker than it has to be because it does not take into account the knowledge we have about the linear trend. The corresponding test for the linear trend A_1 is also weaker, but strong enough, because $\hat{A}_{1,obs} \gg \hat{A}_1$ for all $(\hat{A}_1, \hat{A}_2) \in \Omega$ (recall the Ω is the confidence region limited by the dashed curve). In other words, the linear trend is much greater than the linear pseudotrends for virtually all null-ensemble members, while $\hat{A}_{2,obs} \sim \hat{A}_2$ for a significant fraction of the ensemble members.

This fact (and physical evidence) justifies to analyse the residual record after subtracting the estimated linear trend for assessment of the significance of the non-monotonic variability; in other words to analyze the detrended record. We don't know that the

anthropogenic signal is exactly linear (we are pretty sure that it is not) so we cannot attribute a particular physical cause to the estimated linear trend. For the same reason as it does not make sense to use "nested model comparison" procedures to obtain more precise estimates of the frequency f, it is no point in estimating to high accuracy a parameter of a model which gives only a crude representation of the physical phenomenon of interest. Nevertheless, we contend that it does make (common) sense to study the significance of the residual variability around such a linear regression line in relation to a given null model for the undriven variability. This is what we do in Section 3. Figure 6(a) shows that $\hat{A}_{2,obs} > \hat{A}_2$ for all ensemble members, which implies rejection of the null.

Our remark about "the Bayesian spirit" has been discussed in our response to Reviewer #3. Since it seems to be confusing we have no problem with omitting that phrase.

In his last paragraph Chandler writes: "the description$\cdots$ suggests that the estimated trend slope has been kept fixed throughout, rather than estimated for each ensemble member - if you do this then you fail to account for the uncertainty in this estimate." We suggest that the reviewer reads the text once more an take a look at Figure 6. Here we plot a joint distribution for $(\hat{A}_1, \hat{A}_2)$, which means that both slope $\hat{A}_1$ and oscillation amplitude $\hat{A}_2$ is estimated for each ensemble member.

About the definition of β
From a pedagogical viewpoint we think it is appropriate to define LRM in stationary process (fGn) through the ACF. It is of course correct that the ACF, and hence the PSD defined as the Fourier transform of the ACF, does not exist. But the periodogram does, and is used a lot for fBm's ($\beta > 1$). It is always possible to establish a stationary process (for which the ACF exists) from (repeated) differencing of fBm's, which is explained elsewhere in the paper. We will take measures to make this clear in the revision.

What to do in a revision?
If the editor encourages a revision, we will do that. Otherwise we will revise and sub-

mit to another journal. In the revision we will try to incorporate (and condense) the discussions contained in the responses we have given to the reviewers' comments.

Two figures are included.
Figure 1 shows (a): PSD of P_3-detrended monthly GMLT record 1850-2010 A.D. Thick line has slope $-\beta = -0.54$, corresponding to $H = 0.77$. Vertical dashed lines mark the 60-year period (blue), and the 1-year period (red). (b): Variogram of the "profile" $y(t)$ of the GMLT record with variable degree of detrending. Black: after no detrending. Red: after P_1-detrending. Purple: after P_3-detrending. Blue: after P_7-detrending. The slopes for $n = 1, \ldots, 4$ correspond to; black: $H = 0.91$, red: $H = 0.82$, purple: $H = 0.78$, blue: $H = 0.65$. The slopes for $n = 5, \ldots, 8$ are; black: $H = 0.87$, red: $H = 0.70$, purple: $H = 0.77$, blue: $H = 0.23$. Dotted line has slope 0.77. (c): Gray curve: Monthly GMLT anomaly record 1850-2010 in degrees Kelvin (time origin starts 1850 A.D.). Red curve: P_1- fit. Purple: P_3-fit. Blue: P_7 fit. (d): Black: ACF estimate from undetrended GMLT record. Purple: ACF estimate from P_3-detrended record. The shaded areas represent the 95% confidence interval for the ACF computed from ensembles of fGns of the same length as the GMLT record and with $H = 0.75$.

Figure 2 shows (a): Grey curve shows the periodogram of the 350 yr CET record on a log-log scale and black crosses the log-binned version of this periodogram. The red line is the linear fit to the log-binned points in the f- range marked by the line segment. The line has sloe $-\beta = -0.33$. The red shaded area is the 95% confidence region for periodograms computed from an ensemble of fGns with $\beta = 0.33$. The blue vertical lines marks the 70 yr period. (b) Shows the same features for the DFA2 analysis.

Interactive comment on Earth Syst. Dynam. Discuss., 5, 327, 2014.

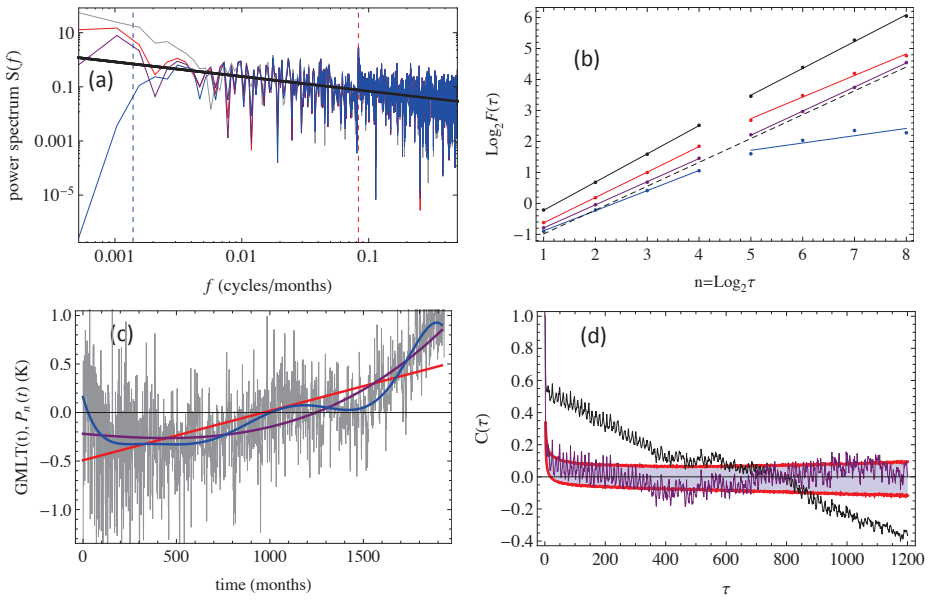

Fig. 1. Analysis of monthly GMLT record 1850-2010 A.D.

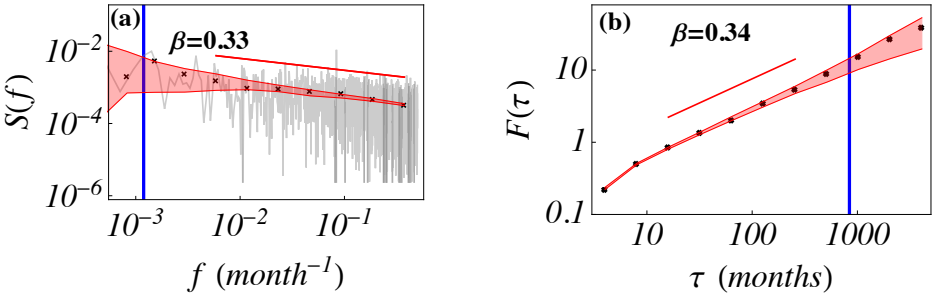

Fig. 2. Analysis of the 350 yr CET record.