Earth Syst. Dynam. Discuss., 5, C348–C354, 2014
www.earth-syst-dynam-discuss.net/5/C348/2014/
© Author(s) 2014. This work is distributed under
the Creative Commons Attribute 3.0 License.

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.

R. Chandler (Referee)

r.chandler@ucl.ac.uk

Received and published: 15 August 2014

The stated objective of this paper is "to establish beyond doubt the significance of the global warming signal, and if possible also the multidecadal oscillation" (page 334, lines 9-10). To accomplish this, the authors fit a sequence of statistical models to two monthly temperature anomaly time series: the HadCRUT3 global land temperature record from 1850 to date, and "a global ocean record" (is this from the ICOADS dataset?). The staatistical models considered are: a "noise-only" model which can be either short-memory (AR(1)), long-memory (fGn) or nonstationary (fBm); a "noise plus linear trend" model; and a "noise plus linear trend plus cycle" model. The authors

use what is essentially a parametric bootstrap technique (although they indicate that it is Bayesian) to compare the models and to draw conclusions about the statistical significance of the trend components. They conclude that there is "no doubt about the significance of a global warming signal over the last 160 years even under null hypotheses presuming strong long-range persistence of the climate noise" (page 346, lines 17-18), and that "we are able to establish statistical significance of the oscillatory trend in the land data record" (page 347, bottom).

At its heart, this paper revolves around the interpretation of the statistical analyses that it contains. Most of my comments focus on the statistical aspects, therefore. In fact, I presume this is why I have been asked to review the paper.

Given that the paper is fundamentally statistical in nature, I am surprised that the bibliography contains only two statistical references - the books by Beran and by Gelman et al - and the authors really struggle with the clear presentation and conceptualisation of statistical arguments. I think they have tried to engage carefully with statistical subtleties and pitfalls, and some of the points that they make (e.g. regarding the precise definition of a trend, and the need to be careful about the interpretation of hypothesis tests) are sort of correct. Unfortunately however, I have very strong reservations about the way that the statistics are used and interpreted, as well as more general concerns about the paper; and I am sorry to report that I do not think that it should be published.

By way of justification for this viewpoint, here are my most serious concerns:

- The authors do not make any clear case that the analysis is necessary. The overwhelming consensus within the scientific community (and wider society) is that there is a genuine warming signal in the climate records of the last two centuries, and there are plenty of analyses around that test for this against a variety of null hypotheses just as the authors do here. The authors themselves cite the Bloomfield and Nychka (1992) paper that first tested for a linear trend against a long-memory null hypothesis; and there is really far too much literature around on the detection of linear trends in climate time series (I consider myself to be reasonably informed in this area). Within the context of this literature, this is just another paper that comes up with more or less the same result using more or less the same kinds of techniques - differing in the details of implementation, but not in the fundamentals. It is naive to consider that this kind of analysis will establish the significance of the global warming signal "beyond doubt" as claimed: this paper is not going to convince anyone who is not already convinced by the enormous body of literature that all points in the same direction. It is possible that the inclusion of the oscillation is new here, but my personal view is that this particular interpretation of the record by some sectors of climate community is fundamentally misinformed (more on this below), and I don't think the present paper makes a convincing case.

- I am very concerned about the statement of the main objective of the paper (quoted in the first sentence of my comments above), which implies that you already know what answer you're looking for. I initially thought that this was just an instance of careless writing, but there are several other instances that give the strong impression that the authors have a - possibly unconscious - bias in the way that they approach their analysis. Thus, on page 334 lines 13-14: why is HadCRUT3 "optimal for trend detection"? As written, it sounds as though you have chosen to use this dataset because it maximises your chance of finding a significant trend. And Section 3 reads like a blatant attempt to make the oscillation seem significant after the first analysis failed to do this. Actually, I think the results in this section probably *are* broadly OK, but not for the reasons that the authors give. Overall, this impression of a biased approach to the analysis is very disturbing: it is poor statistical practice, to say the least, and - more importantly - when this kind of material appears in the peer-reviewed literature then it undermines the credibility of climate science.

- There is a major disconnect between some of the rather discursive material in the introduction, and what was actually done. Thus I fail to see the connection between the data analysis and the introductory discussion about different interpretations of trends,

or about the separation of forcing into different components (Fig 1a is not necessary, nor is it helpful to call it a Venn diagram - it just makes the setup seem more complicated than it really is). As another example, there are strong implications (p331, lines 25 onwards) that somehow the analysis gets round the lack of data of sufficient quality to carry out an unambiguous empirical assessment of the role of anthropogenic forcing - but this paragraph is left hanging, without any clear explanation of *how* the proposed methodology achieves this. Indeed, it is fundamentally impossible without some heroic assumptions about how to interpret the results of the hypothesis tests.

- On the subject of hypothesis tests, the paper reads as though the authors do not quite understand how to interpret them - or, at least, as though they would really like to be able to interpret them in a way that isn't quite justified, but that they are aware that it isn't quite justified, but they're going to do it anyway. Thus, on page 336 lines 19-27, there is some rather convoluted attempt to say that if you reject the null hypothesis then you accept (but don't verify) the alternative; and perhaps something similar at the bottom of page 338 (I don't really understand what the authors are trying to say in this paragraph); but then the presentation of the results is very much couched in terms that appear to say "the linear trend and oscillation exist". Hypothesis testing is, of course, one of the most misunderstood and abused techniques in statistics: the correct interpretation is that if you reject the null then you have learned that the null seems inconsistent with the data. That's all. You have not learned *anything* about the alternative, except that it compensates for some of the deficiencies of the null according to the measure that you used to perform the test (at the risk of degenerating into a lecture: hypothesis testing was originally designed for use in situations where data were gathered from experiments, carried out in such a way that the alternative was the only plausible explanation if the null was rejected - but this is certainly not the case here). Thus, for example, the significance of the 70-year oscillation should strictly be interpreted merely as an indication that the linear model doesn't fit the data - but at the top of p348 we find "oscillation cannot be dismissed as spontaneous random fluctuation" which, to the reader, suggests that therefore the oscillation is genuine. The authors may not have intended the reader to draw this conclusion, but if so then they have written up the results rather carelessly. I might add that if you want to make a convincing case that, say, your full model is adequate but that the reduced models are not, then some carefully chosen residual diagnostics (i.e. analyses of the estimated innovations from the fitted models, which should behave more or less as white noise) would not do any harm.

- More generally, the presentation of statistical methodology (particularly - but not exclusively - on pages 336 to 338) is very confused, unclear and non-standard. I think it will be very difficult for a non-specialist fully to appreciate what is being done (I struggled myself in places, and as a Joint Editor of the Journal of the Royal Statistical Society I can claim a reasonable degree of familiarity with the subject). The methods used here are perfectly standard and can be found in any good advanced statistics text. Some of the definitions and vocabulary introduced in the paper are not necessary; and some of the descriptions of (for example) Bayesian inference are very imprecise. I think the authors could benefit from familiarising themselves with a much wider range of modern statistical literature; or with some texts by experienced statisticians that are aimed at environmental scientists. The book "Statistical Models" by Anthony Davison (Cambridge University Press, 2003) is an excellent place to start. For example, the "nested model comparison" procedure that the authors attempt can be made much more rigorous and transparent than the procedure that they describe - particularly with respect to such issues as fixing the values of unknown parameters when carrying out tests.

A few additional, more minor comments are as follows:

- P332, line 21: if the process is nonstationary then neither the ACF nor the power spectral density is defined, although there is no problem with the underlying stochastic process constructions (I mention this because beta has been introduced in the context of the ACF and spectral density, which is a bit limiting). Indeed, line 16 asserts that if beta>1 then you can have (auto)correlations greater than 1! Much more care is required to ensure that the presentation is precise and correct.

- P334, lines 5-7 "This is an important message ...": the tone of this sentence is not really appropriate for a scientific article.

- Section 2.4: the existence of a pure deterministic sinusoidal cycle is very dubious, and this kind of work harks back to the early years of the 20th century when there was a major "cycle-finding" industry in economics in particular - until it was realised that an AR(2) process could produce quasi-cyclical behaviour. I simply don't believe this model. Lines 10-12: I don't understand the argument that f should be fixed rather than estimated, either - it runs contrary to almost all statistical practice. If you had good a priori grounds for fixing it (as in the case of the annual cycle, for example), then this would be a different matter. If you really believe that there is a sinusoid here, you should estimate the frequency and carry out a nested model comparison that accounts for the estimation uncertainty (the author's claim that they can take the frequency as fixed by hypothesis is incorrect, because they used the data to determine its value in the first place). As noted above, nested model comparison in this kind of situation, taking account of the fact that the data have been used to estimate any unknown quantities including those that are not of direct interest, is an absolutely standard procedure.

- P341 lines 16-19: to take the linear trend as an established fact is not at all in a Bayesian spirit - it isn't clear to me what the authors mean by this, but it certainly is not correct. Scientifically in fact, this is completely the wrong thing to do since it treats something as known when in fact there are uncertainties associated with it. This is really poor, and is another instance where it feels as though (consciously or not) the authors are biasing their analyses to get the results they want. The question *can* be addressed using properly conducted nested model comparisons, but it isn't clear to me that the comparisons have been done correctly here (the description isn't clear enough for me to follow precisely, but it suggests that the estimated trend slope has been kept fixed throughout, rather than re-estimated for each ensemble member - if you do this then you fail to account for the uncertainty in this estimate).

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