International Journal of Climatology Decision Letter regarding McIntyre and McKitrick, “An updated comparison of model ensemble and observed temperature trends in the tropical troposphere”, submitted January 26, 2009

01-May-2009

Dear Mr McIntyre

Manuscript # JOC-09-0031 entitled "An updated comparison of model ensemble and observed temperature trends" which you submitted to the International Journal of Climatology, has been reviewed.

Below I have provided the comments from the two reviewers plus my decision at the end of this mail.

Authors Comments

Reviewer 1 has made the following points in relation to technical issues.

1. The reviewer notes that S08 stopped in 1999 because most of the model runs stopped in 1999 and that the method followed in your paper relies on the assumption that the trend in the model data would be the same if extrapolated up to 2008. Even if the idea was first suggested by S08, one needs to ask whether that's a valid assumption. Clearly it is possible that if model runs could be continued up to 2008, one would observe changes in trends of both series. The reviewer suggests that this needs to be acknowledged as a possible alternative explanation of the findings.

2. In relation to the one-sided or two-sided tests (page 3) Reviewer 1 is not convinced that there is any a priori reason to assume that models will exhibit a stronger trend than the real data. In other words, he thinks using two-sided tests is reasonable. He notes that for testing against zero trend, the argument is stronger than one would expect the trend to be positive, based on simple physical interpretations of the greenhouse effect.

3. Reviewer 1 finds the interpretation of Fig 1 questionable, especially the RSS results. He suggests that there is no real evidence of a trend towards a statistically significant result, and (in the case of RSS-T2), notes that if only one out of 10 tests conducted at the 10% level is significant, why is that even worth mentioning at all? Further he states “as a general matter, as the author surely understands, performing a sequence of hypothesis tests on the same dataset and then looking for a "trend" in the results is not a standard interpretation of hypothesis tests”.

4. Reviewer 1 has raised the issue of the impact of autocorrelation calculation, especially the possibility of using alternatives besides AR1 (page 7). He has questioned why not do this? He notes that the arima command within R has been used, which allows for any ARIMA(p,d,q) process, and highlights that only the case where (p,d,q)=(1,0,0) has been used in the paper. The reviewer also notes that the xreg option within the arima procedure has been used which would allow you to estimate the trend and standard error directly, for ARMA models of any order, without getting into the effective degrees of freedom mess at all.

5. The Reviewer notes that despite claims made in the paper, a test of the UAH trend against the RSS trend has not been made. What is shown is that up to 2008, the RSS trend is statistically significant and the UAH trend is not. The reviewer notes that this does not prove that the RSS trend is statistically significantly different from the UAH trend. He considers this a hypothesis worthy of testing.

Further to the above Reviewer 1 comments on the major non-technical issue - the apparent refusal of Santer to release the datasets on which the S08 paper was based. He notes that while those who believe in the open dissemination of climate data could consider Santer's stance unfortunate, Santer did not in fact violate any legal or professional obligation.

Lastly the Reviewer notes that the intention of your paper should be a full independent "audit" of the results of S08. To achieve this he makes the following suggestions and comments.

1. A large repository of data is available from pcmdi (http://www-pcmdi.llnl.gov/). Although the reviewer has some reservations about whether ALL the S08 model runs are there (in his experience, modelling groups put some of their data in the public data depository, but not all of it) he notes that it should still be possible to find enough data to do some independent reconstruction of the model dataset. Further if you found the trends to be different because you used not exactly the same data sets that would be a worthwhile conclusion in itself.

2. Processing the raw data is not actually so hard. He notes that there is a downloadable package "ncdf" to read netcdf files and process the data in R, and the NCAR group maintains a webpage explaining how to do it (http://www.image.ucar.edu/Software/Netcdf/). Consequently it should be possible to calculate time series of observations averaged over specific latitude-longitude bands using these tools.

3. If there is indeed a published formula showing how to calculate synthetic MSU temperatures from archived climate model data, it should not be too hard to apply it.

Overall Reviewer 1 notes that some interesting points have been made in the paper but believes that not enough has been done for the paper to stand as an independent research paper. The reviewer believes that even with the data available at the time of writing more could have been done (two specific suggestions are pertinent in this regard: investigate alternative time series models; and perform a direct test of whether UAH and RSS trends are the same). He suggests that if you could independently reconstruct the model data from archived sources, or document why this cannot be done, it would be an even more worthwhile exercise.

Reviewer 2 at the outset of his review notes that the major point, although obscured in many aspects by presentation style choices and non-essential cul-de-sacs, is that end-point effects are important to ascertaining whether a result is significant or not. While he does not disagree with the pertinence of this point he feels that it is hardly a novel point to make. He notes that compared to what is presented in your paper this is an issue that has been addressed more fully and analysed more robustly in the recent literature. Further he is rather concerned that the considerable volume of relevant literature relating to trend determination viz a viz end-points has not been covered at all.

Having said that Reviewer 2 has some sympathy with the viewpoint that S08's choice of 1999 as an end-point may be sub-optimal. However, he notes that 2008 is very likely an end-point that will have an even more substantial impact on any analysis and makes the point that the choice to consider only sensitivity to 1999-2008 and not 1979-1988 start years belies a significant lack of understanding of the satellite data issues. He highlights the fact that it is the early part of the satellite record that is more substantially uncertain.

Further to these overall comments Reviewer 2 makes the following major points in his review:

1. The author chooses to focus solely upon the statistics of the time series behaviour, which would be fine if we believed those time series to be well defined. But we don't in the observational record. We know that there are substantial issues regarding the homogeneity of the long-term record from satellites (and arguably all observing types). The complete lack of referencing that literature or accounting for structural uncertainty other than through taking 2 points from an infinite sample of plausible MSU datasets means that the necessary context to analyse their work within

is missing. Worse, these points are taken absent of the quoted dataset construction uncertainties (+/-0.09K/decade for RSS, +/-0.03 K/decade for UAH (who consider only a subset of the sources considered by RSS)) which surely are important in the context of answering such a question in a statistically robust manner. The author would need to read, understand and then quote appropriately the wealth of literature on dataset uncertainty and published dataset construction uncertainty estimates in any resubmission. It would be necessary to include the radiosonde datasets used in S08, in part to provide a longer-term context, and in part to sample the plausible dataset structural uncertainty space better.

2. I am unclear as to whether the authors are adequately accounting for the fact that the model mean trend is greater than the reported observed trend at the surface. This is an issue that I have with the choice of metric in the S08 and Douglass et al. also. What is tightly constrained in the models is the differential rates of change (or amplification). So, arguably, we should be undertaking this kind of consistency analysis in the style of Santer et al., 2005 / CCSP by looking at the observed and modelled amplification behaviour. As noted in S08 with regards to the discussion of their radiosonde trends figure comparing multi-model TLT to observational TLT may well be a conservative consistency test given that the model surface trend is greater than that observed. This may well have a huge impact on marginal consistency / inconsistency tests as is the case here. I would need to see this issue being thoroughly addressed in any resubmission to ensure that the author's findings stand up to this uncertainty. One way to do this would be to apply a transform to the model data to account for the mis-match. Another would be to consider the ratio of changes.

3. I find the analysis over-simplistic and too focussed on one end-point to the detriment of our greater understanding. Papers cited in SO8 present comprehensive assessments of uncertainty in amplification behaviour to timeseries start and end-point effects and find these to be large even at the 30 year timescale both in climate model runs and in the observations (satellites and radiosondes). This is a weakness. That said, the author could improve upon such analyses by considering the full range of observational and model datasets available and performing a formal S08 style statistical analysis. The author would also need to use a more appropriate linear trend estimator than OLS which will be super-sensitive to end-point effects. There are several such robust estimators in the literature and a robust sensitivity study to this choice would be entirely appropriate given the borderline consistent / inconsistent nature of the issue.

4. I find the inclusion of "is the AR(1) model valid" discussion to be disingenuous. If the author feels this is important then they should do the actual hard yards of implementing such analyses to ascertain whether it makes any difference. This would actually move the science forwards and constitute a useful contribution. Simply arm waving isn't good enough.

5. The abstract is disingenuous in its assertion in the final sentence. As the author points out RSS is still consistent with models yet reading the abstract the strong implication is that all observations disagree with the models. This simply isn't true. I therefore fail to see how their work fundamentally disagrees either with Santer et al. 2008 or the earlier Santer et al., 2005 or CCSP analysis. Or for that matter with an assertion of "partial resolution" of the issue. Or does the author want to wish RSS away as a plausible dataset?

6. The discussion of CCSP is plain wrong. Chapter 5 explicitly considered whether models and observations were consistent in their amplification behaviour. It is also stated that Chapter 5 had an appendix; there was no such appendix. The whole report had an appendix. This is but one glaring example of referencing on the fly with at best half truths and leaves the knowledgeable reader with the impression that the author has referenced the material without actually reading it first. This may be a mis-impression. In which case the author needs to work very hard on greater clarity in their referencing style.

7. The RSS dataset version is different and is why the analysis of that series differs. The readme file on the RSS site documents this and is easily available. This and a lack of referencing

conspire to rightly or wrongly give an impression of lack of due diligence in understanding the data before its use.

8. The question of trend significance from zero was not covered by S08 and is a side show. I see no reason why this should remain in any published submission. This is not the question that was being posed in S08. It simply distracts from what could be scientifically useful.

9. I tend to side with Dr. Santer's argument that there is value to full replication. The data sizes are not as huge as stated and the problem is easily solved computationally. However, it is as likely that an error was made in creating these as in any other step so simply grabbing value added data will deny an important opportunity to assess the overall uncertainty. The author may well have been able to replicate the final step of S08 if the data had been shared but both analyses may well have been based upon value-added data that was flawed. Short-cuts are dangerous. Anyway, this discussion should be cut.

Finally Reviewer 2 makes it very clear that some of the "political comments" in the paper do not have a place in the scientific literature. His preference would be for the paper to focus on the scientific issue of model ensemble and observed temperature trend comparison.

Decision
The issue addressed in the paper is an important one and the paper does contain some valuable scientific points in places. However in its present form the paper is unacceptable for publication and my decision is reject. This is because not enough has been done for the paper to stand as an independent paper. As comes through in the reviewer's comments there are major concerns relating to understanding of the literature and the nature of the satellite data, the methodology, missed opportunities in terms of what could have been done with the data available at the time of writing and inappropriate excursions from the main scientific point of the paper.

My recommendation is that you consider ways in which you could improve the scientific focus of the paper based on the data that was available to you at the time of writing. In this regard following up on the suggestions made by the reviewers would be an extremely useful starting point. In the absence of non-scientific issues, a paper that adequately reviewed the literature on the general subject matter of observational and model trends, implemented a methodology appropriate to the problem and objectively considered the analysis results and drew conclusions accordingly could then be reconsidered for submission and review.

Thank you for considering the International Journal of Climatology for the publication of your research.

Sincerely,

Prof. Glenn McGregor
Editor, International Journal of Climatology