Clim. Past Discuss., 9, C2992–C2997, 2013
www.clim-past-discuss.net/9/C2992/2013/

Climate of the Past Discussions

Interactive comment on "Discrepancies of surface temperature trends in the CMIP5 simulations and observations on the global and regional scales" *by* L. Zhao et al.

Anonymous Referee #3

Received and published: 26 December 2013

Recommendation: Reject and resubmit. The approach here has some original aspects in that the authors analyze 10-yr running mean trends over the historical record to assess the consistency of models. Unfortunately, the methodology is too simplistic and the authors seem unaware of published work that points toward ways of making their results more robust and relevant. The key shortcoming is in construction of the shaded regions that show the model "range" and in the interpretation of differences between the observations and this "range" as an inconsistency. I go into full detail below (Major comment 2) on how a more appropriate comparison can be done, drawing on previously published works. Through such methods the authors would more clearly recognize the important roles of internal climate variability (vs. just difference in means

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

responses of particular models) in creating differences they see between models and observations.

Beyond this key point there are several major apparent flaws, such as an apparent failure to mask the model data where observations are unavailable (I assume this was not done since it was not mentioned).

The current manuscript is unacceptable and should be rejected. However, with a major upgrade of the methodology along the lines discussed below, the authors could put together and resubmit a much better, and probably publishable, contribution.

MAJOR COMMENTS

1. The authors have not referenced several key works on comparison of multi-model ensembles of trends with observations. These papers give more of an idea of several sounder, more robust approaches to the comparison problem than they used here. Among the relevant papers (also reviewed in Ch. 11, Box 11.2 of IPCC AR5 "Ability of Climate Models to Simulate Observed Regional Trends") are:

Bhend and Whetton (2013), Climatic Change They use a standardized difference approach to compare simulated and observed changes (per degree of global warming) accounting for autocorrelated residuals and the problem of multiple testing.

Knutson, Wittenberg, and Zeng (2013), J. Climate They test consistency of multi-model ensemble and observed trends, including a "sliding trend" test analogous to that used here but with a fixed end point and variable length, rather than a fixed trend length and variable end point. They illustrate a method for finding a multi-model 5-95th percentile range to compare with observations, which could be used here (see major point 2 below).

van Oldenborgh, Doblas Reyes, Drijfhout, and Hawkins (2013) Envir. Res. Lett. They consider whether climate model 'hindcasts' are reliable in the sense that there is an equal probability of the observed trend being at any particular percentile of the multi-

CPD

9, C2992–C2997, 2013

model distribution (i.e., using a ranked histogram approach). This can show whether the model results are under-dispersed (models are over-confident) or over-dispersed.

Yokohata et al., Clim. Dyn., 2013. Another example of using the ranked histogram approach used by van Oldenborgh et al.

The techniques in the above papers should be considered, and then the authors should redo their analysis, upgrading it based on the methods in one of more of these papers. While any of the above approaches would be a great improvement over what the authors have submitted, the Knutson et al. approach is probably closest to what they are attempting to do in this particular manuscript, and so is outlined and contrasted with the present approach in more detail below.

2. How were the mix and max range for the models constructed e.g., Fig. 1? This is not clearly spelled out in the paper, but a proper methodology is absolutely crucial to make a reasonable comparison of consistency between models and observations. Based on the statement near the top of p. 6165, I am assuming that the authors compute an ensemble mean trend for each individual model and the min and max of this set of ensemble trends forms the min and max for Fig. 1. Since the authors are not clear on their my interpretation could be wrong, but they almost certainly do not implement a method such as the one discussed below.

As mentioned, the issue of how to form a shaded "consistency range" on a figure like Fig. 1 in the context of checking model consistency is the central problem the authors must deal with as it now stands, and their method must be upgraded.

For example if their current approach, as described above, is used, then the min/max range is being constructed from a "mixed bag" of ensemble means of various sizes. For models with larger numbers of ensembles, like n=10, the ensemble mean will begin to approach the forced response, with interannual variability largely filtered out (especially for large regions like global means). For models with only small ensembles, like n=1 as a limiting case, the ensemble mean contains a much larger relative contribution of in-

CPD

9, C2992–C2997, 2013

ternal variability (being a single sample of such variability), with the relative contribution also depending on the size of the region averaged, etc. The range so constructed conflates difference in forced response with some (limited) measure of internal variability that is inconsistent across the models as it is based only on the limited (and variable) number of individual runs available for each model.

In any case, using the control run variability can provide a more robust (much better-sampled) estimate of the internal variability potential contribution than does the limited number of individual ensemble members. Making use of the control run can be done in the following way. For each model, the ensemble mean trend across the different ensemble members is calculated. A 5th to 95th percentile spread about this mean can be estimated by drawing (with replacement) a large number (n=100 to 1000+) of random 10-yr trend samples from that model's control run. The observed trend can be compared with this distribution: if inside the 5th to 95th percentile range that particular model is consistent, etc. For multi-model comparison, the set of trends from each model (just constructed) can be combined into a single multi-model distribution of trends. This distribution will have a spread due to the different ensemble mean trends from the different models and the internal (control run variability) from the different models. The observed trend can be compared with this "grand" distribution and if inside the 5th to 95th percentile range then the multi-model ensemble and the observed are consistent, etc.

3. The authors did not state how they treated missing observations in the analysis. The proper way would be to mask the model data by removing any model data that corresponds to missing data in the observations. Otherwise comparisons for data poor regions such as the Arctic and Antarctic are highly questionable. As an example, the assessment on p. 6169, line 26 for Arctic temperatures will depend on how missing data is handled (model vs. observation). A reference from the observed perspective is Cowtan and Way, QJRMS, 2013, doi 10.1002/qj.2297. Similar issues arise for p. 6170, line 5.

CPD
9, C2992–C2997, 2013

4. The treatment of internal climate variability as a potential reason for discrepancies is poorly handled. For example, p. 6166, lines 14-18, the differences in models interpreted treated as flawed estimations whereas internal climate variability in the real world (*which we do not expect models to reproduce in terms of timing*) could be part of the explanation, and the model and observations could in fact be consistent, but this is not even mentioned. Similar comments apply for p. 6167, line 14-16.

5. p. 6167, line 9. How is it determined whether the difference is statistically significant or not? Much more detail is needed here. If a statistical test is not done, this should be implemented (see previous references). Such tests will depend critically on how the control run is used for sampling internal variability, etc.

6. What is assumed about degrees of freedom for the assessment that a correlation of 0.4 is significant at 0.01 level? Is this assumption valid? See Wilks et al. text on Statistical analysis in atmospheric sciences, for example.

7. Discussion of tropospheric temperature trends is mostly tangential to the study except p. 6163 lines 27-29. The rest (lines 17-26) can be deleted. The authors also do not cite an important rebuttal of Douglas et al., but this is all simply tangential to this study. They need to focus on the aforementioned previous work that was not cited or used in the design of this study.

8. On what basis were 6 of 22 models excluded from the analysis?

9. The authors should be aware of several issues with control run drift. First, if the control run has a long-term (multi-century) drift, then this drift component should be removed from the historical runs, which will inherit this drift as a nonphysical trend bias. Second, in sampling the control run for samples of internal variability, the long-term drift should first be removed to avoid adding a (small) bias to the trend samples.

MINOR POINTS

10. p. 6166, line 27-29 the wording here implies that the HadCRUT4 is "truth" and

CPD

9, C2992–C2997, 2013

that GISS observations, if they do not agree with HadCRUT4, are wrong. The authors should rewrite to note that this is not necessarily the case, otherwise substantiate the claim.

11. p. 6167, line 7-14. These are very subtle features that are hard for me to even see on the plots provided and using the methodology that they employ.

12. p. 6168, line 10-11. Part of the reason may be the enhanced internal variability relative to the longer-term trends in this region. Also as averaging regions get smaller, the noise level often rises. Finally there is the issue of poor data coverage over Antarctica which can also lead to inflated variability.

13. p. 6168, line 14-16. This sentence as written does not make sense. Please rewrite.

14. p. 6170, line 1. I don't see the big issue near 1905 that the authors are referring to.

15. Fig. 1 The biggest discrepancy I see in the global mean comparison is around 1945, but this is not mentioned by the authors, who focus on less prominent features.

END OF REVIEW

Interactive comment on Clim. Past Discuss., 9, 6161, 2013.