Clim. Past Discuss., 8, C1827–C1835, 2012
www.clim-past-discuss.net/8/C1827/2012/

Climate of the Past Discussions

Interactive comment on "Climate of the last millennium: ensemble consistency of simulations and reconstructions" *by* O. Bothe et al.

O. Bothe et al.

ol.bothe@gmail.com

Received and published: 9 October 2012

Dear Dr Edwards, Dear editor,

Hereby we want to acknowledge Dr Edwards comments and address them quickly one by one.

Obviously we do not agree with her final recommendation.

A full response to all three referees will be provided if we are encouraged by the editor to submit a revised version.

Preface: We think that we discussed the relevant pitfalls, but are willing to further extent this in case the editor invites us to submit a revised manuscript.

"General issues":

Comment (a) is in line with the issues raised by referee 1 on page C1364. However, the other referee did not imply the invalidity of the results and the approach. First, we note that the current version of our revised manuscript addresses this concern by including estimates of internal variability. This version considers so far the comments by both anonymous referees. Indeed the over-dispersive character is not as pronounced as before but is still the most common inconsistency since the abstract also refers to the field reconstruction, which according to line five of the review is not considered by Dr Edwards. With respect to the supposedly incorrect inferences, we will answer all specific comments to p2418 below.

However, we agree that it is necessary to (i) justify the use of the ensemble mean and (ii) to discuss here part of the results on page 2418.

In the original manuscript we assume that uncertainties for an ensemble mean which are proportional to the spread of the respective ensemble. In principle, these uncertainties account for the differences between the assumed signal and how the individual ensemble members represent it. These differences can be assumed to represent a combination of the internal variability and the methodological uncertainties. If we randomly sample these uncertainties to inflate members of the ensemble that we wish to verify, this should compensate for the reduced internal, i.e. unforced, variability in the ensemble mean under certain conditions. The main condition is that the forced signals represented by the ensemble mean target and the individual ensemble members are similar. It further implies that these uncertainties are uncorrelated in time. In our understanding the analysis remains valid considering our original discussion of the caveats in sections 2.3 and 4. Obviously the analysis benefits if we address these issues directly by adding estimates of the internal variability to the verification target as we do in reply to referee 1.

With respect to the comment on Figure 1 and p2418/14: the lower panels of Figure 1

are indeed an expression of the use of the ensemble targets, but similarly reduced variability occurs at least for certain periods for a number of temperature reconstructions that are meant to reliably represent interannual variability (not shown). Furthermore, we discussed in the original manuscript: "it is arguable whether an ensemble mean represents unfiltered annually resolved data. A posteriori, our approach seems to be valid for the comparison of the specific simulation ensemble mean with this particular reconstruction ensemble, but the larger variability in the simulations compromises the inverse consideration." This note in part referred to Figure 1, where we see that the proposed centered character of the simulation ensemble mean target relative to the reconstruction ensemble is much less pronounced than for the reconstruction ensemble mean target relative to the simulation ensemble. Figure 1c rather indicates various periods where the ensemble is likely biased compared to the target.

Thus, while agreeing with the notes by referee 1 and 3 an ensemble mean is a priori a valid target. Following Persson (2011), Hargreaves et al. (2011), Johnson and Bowler (2009) and Marzban et al. (2011) we could have inferred a priori that the analysis would result in dome-shaped rank histograms and positively sloped residual quantiles. However, this is not true for the assessment of the bootstrapped intervals.

Less easily a priori inferable are the results obtained from the data if we add an additional estimate of internal variability. This analysis is included in the so far revised version of the manuscript. There we still see over-dispersion for the simulation ensemble even though it is reduced. However, the reconstruction ensemble is found to be probabilistically consistent, but diverse climatological deviations still remain.

The current status of the manuscript follows the comment of referee 1 and discusses this more explicitly. Besides these already implemented changes, we do not think that extending the analysis to one or all (sub-)ensemble members would improve the manuscript since we already discussed this to some extent in section 4 of the original manuscript. However, we will further extend this discussion.

Next, we will first respond to comment (c). We note that for the European data, the maximum inter-ensemble correlation is about 0.17 and it is usually between 0.2 and 0.25 for the northern hemisphere data. The former is notably stronger than the correlation between target and individual simulations. The latter is usually in the range of correlations between simulation members and the ensemble mean reconstruction target.

We quote here the following from Marzban et al. (2011): "...if the correlation between ensemble members and the observation is generally different from that between ensemble members, then the RH cannot correctly assess the climatological component of reliability. For realistic ensembles, where both correlations are large ($\tilde{0}.9$), the RH is then expected to be U shaped *by default*, and so cannot assess the climatological component of forecast reliability. And if the ensemble forecasts are more similar to each other than to observation, then the RHs are still U shaped, even though there exists no climatological overdispersion." Please note that from our understanding the end of the last sentence should read "underdispersion" instead of overdispersion. Please, refer further to Figures 5 and 6 of Marzban et al. (2011). For all possible readers of this reply we note that an author version of Marzban et al. (2011) is available at Caren Marzban's homepage at the University of Washington (http://faculty.washington.edu/marzban/).

From the given numbers, we would assume that our assessment should reasonably approach the uniform outcome and should not show deviations as large as found in our evaluation. If invited to submit a revised version, we are happy to discuss these issues more thoroughly.

Comment (b): The intra-ensemble correlations for the reconstructions (see below for definition) is usually between 0.4 and 0.6, but reaches 0.82. The correlation to the ensemble mean simulation target is usually smaller.

Thus we would expect under-dispersive u-shaped rank-counts following Marzban et al. (2011) for all three area-averaged assessments.

Dr Edwards' comment (d) notes an insufficient discussion of the sensitivity of our results to made choices and temporal correlations.

The bootstrap block lengths: We subjectively applied a 50 year window. Generally it can be noted that longer windows reduce the width of the quantile ranges while shorter blocks widen it. The 50 year block length primarily was a compromise between the preferable lengths (according to the auto-correlation functions) for the ensemble mean target (longer window) and the simulation ensemble members (where shorter blocks are possible). Thus, for the case of the ensemble mean target we indeed use possibly too short blocks.

The degrees of freedom of the time-series are estimated following $N_{eff} = N * (1 - r_{1max} * r_{2max})/(1 + r_{1max} * r_{2max}) - 2$ which enters the calculation of the χ^2 statistic directly and automatically.

As discussed in the referenced literature and also mentioned in the draft, this has an essential influence on the results, which is from our point of view clearly detectable in the presented Figures and discussions. This is also the reason why we do not show an absolute result but allow the reader to make his own inferences from the results including and excluding uncertainties accounting and not accounting for serial correlations. We are happy to extend the discussion of the approach, of our choices and of related sensitivities even further.

The rest of this reply refers to the further issues noted by T.L. Edwards.

Page 2411 line 18 (2411/18): Agreeing with the other two referees, this has already been changed in the so far revised version.

2411/20, 2413/8: We will remove the first occurrence and weaken the second occurrence while still noting that the approach is less subjective than common "by eye" evaluations.

2414/18: Although we are happy to give more details on the block-bootstrap approach,

we do not think, that the manuscript benefits from such a discussion. Efron and Tibshirani (1994) is a standard reference for the bootstrap which outlines the essentials of the approach and should be referred to by the interested parties.

2414/24: The newest version includes more discussions on both aspects but after this review we may extend this even more.

2415/11, 2415/11: We do not think that the manuscript benefits from such a table as the ensemble has been discussed in depth by Jungclaus et al. (2010). Neither do we think that the manuscript benefits from one to three more Figures showing the forcing time-series. Rather the manuscript becomes unwieldy by including and discussing these series. However, if, in the end, the editor and the other referees agree on the necessity of table and Figure, this is easily included.

2425?/11: We refer to the descriptions by Jungclaus et al. (2010) on the ensemble.

2515/17: Yes all reconstructions are for annual temperature. We are going to try to clarify the sentence.

2415/18: While we would prefer to refer to Frank et al. (2010) for the information, we will include a description of the 9 original reconstructions and the re-calibration process applied by Frank et al. (2010).

2415/22: The sub-ensemble members are the 9 reconstructions in the re-calibration window 1920-1960. The end year is chosen as it is the last date available for all 9 reconstructions. For the start year, a discussion of all possible choices would be beyond the scope of this comment and even a reasonably structured manuscript. In short: any choice of a start year is arbitrary. An argument for the year 1920 is that one uses probably the most reliable data source. An argument for the year 1850 is that one uses the most complete data set. One could choose between the sub-ensembles based on the intra-ensemble and inter-annual variability, which would bias the inferences (e.g. choosing the ensemble re-calibrated to 1870-1910 likely gives different results than us-

ing the ensemble calibrated to 1850-1960). In the end, we decided that the reasonable choice is the sub-ensemble re-calibrated with the likely most reliable data.

2416/12: “Similar caveats” refers to sampling variability issues which are pronounced in the presented paleo-context as obvious in the discussion of the field reconstruction data. We think the discussion is generally sufficient to clarify these issues.

2416/23: We are going to clarify how added uncertainty estimates may compensate for the reduced variability.

2417/15: We are going to try to add a discussion of Fig. 1 at the start of section 3.

2417/21: Although we think that the reader benefits from both perspectives, we are going to follow the referee’s comment.

2418/14: The current status of the manuscript follows the comment of referee 1 and discusses this more explicitly. Besides these already implemented changes, extending the analysis to one or all (sub-)ensemble members does not improve the manuscript. We already discussed this to some extent in section 4 of the original manuscript. We will extend on this discussion.

2418/17: We will note the critical values.

2418/17: The calculation of the degrees of freedom is, in our understanding, appropriate. The result at this point depends slightly on the added uncertainty inflation and thus also on the sampling variability.

2418/20/21: We cannot fully follow these two comments but will clarify the references to bootstrapped envelopes and the 16th century especially in relation to the construction of the residual quantiles.

2418/24: The slope is most times within the bootstrapped intervals. However, the bootstrapped intervals include the zero line, therefore the slope is indeed not significant, but should be noted nevertheless.

2418/28: Correct, sorry for that, will be changed.

2419/11: Yes, it accounts for autocorrelations.

2419/14-17: The residual quantiles for the simulations can be inferred from Figure 3d, but the quantiles for the reversed analysis cannot be inferred from Figure 3f, thus they are not shown. We will add the panel to the reference to Figure 3.

2419/19: Will be done.

2419/19: The climatological assessment shows that despite the probabilistic consistency the climatological quantiles lack consistency. We will try to clarify.

2419/23: Generally refers to if the uncertainties are considered. Otherwise to if the uncertainties are not considered. We will try to clarify.

2419/26: While this formulation does not contradict line 19 since it mainly refers to the positive tails of the full evaluation, it is not as clearly formulated as it should be. We will improve this.

2420/5: We will clarify both aspects of the comment.

2420/8: We removed reference to the smoothed data since it doesn’t fit into the current revised status of the manuscript considering the comments by the anonymous referees.

Figure 1: We are going to try to optimize the figure.

Figure 2: Although we think that the reader benefits from both perspectives, we are going to follow the referee’s comment.

Figure 3: A clarification about which panel is considered would help. We assume that the comment refers to panels a to d. Indeed the simulation sub-ensembles are equally colored instead of one color for each member. We may possibly change this.

General comments: We will rewrite where appropriate. We have replaced “truth” due to the comments by the other two referees. The reference to Marzban et al. (2011) has

been updated already as has the reference to Wilks (2011).

References

Efron, B. and Tibshirani, R. J.: An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), Chapman and Hall/CRC, 1 edn., http://www.worldcat.org/isbn/0412042312, 1994.

Frank, D. C., Esper, J., Raible, C. C., Büntgen, U., Trouet, V., Stocker, B., and Joos, F.: Ensemble reconstruction constraints on the global carbon cycle sensitivity to climate, Nature, 463, 527–530, doi:10.1038/nature08769, 2010.

Hargreaves, J. C., Paul, A., Ohgaito, R., Abe-Ouchi, A., and Annan, J. D.: Are paleoclimate model ensembles consistent with the MARGO data synthesis?, Climate of the Past, 7, 917–933, doi:10.5194/cp-7-917-2011, 2011.

Johnson, C. and Bowler, N.: On the Reliability and Calibration of Ensemble Forecasts, Mon. Wea. Rev., 137, 1717–1720, doi:10.1175/2009MWR2715.1, 2009.

Jungclaus, J. H., Lorenz, S. J., Timmreck, C., Reick, C. H., Brovkin, V., Six, K., Segschneider, J., Giorgetta, M. A., Crowley, T. J., Pongratz, J., Krivova, N. A., Vieira, L. E., Solanki, S. K., Klocke, D., Botzet, M., Esch, M., Gayler, V., Haak, H., Raddatz, T. J., Roeckner, E., Schnur, R., Widmann, H., Claussen, M., Stevens, B., and Marotzke, J.: Climate and carbon-cycle variability over the last millennium, Climate of the Past, 6, 723–737, doi:10.5194/cp-6-723-2010, 2010.

Marzban, C., Wang, R., Kong, F., and Leyton, S.: On the Effect of Correlations on Rank Histograms: Reliability of Temperature and Wind Speed Forecasts from Finescale Ensemble Reforecasts, Mon. Wea. Rev., 139, 295–310, doi:10.1175/2010MWR3129.1, 2011.

Persson, A.: User Guide to ECMWF forecast products, Tech. rep., ECMWF, 2011.

Interactive comment on Clim. Past Discuss., 8, 2409, 2012.