Clim. Past Discuss., 7, C557–C561, 2011
www.clim-past-discuss.net/7/C557/2011/

Climate of the Past Discussions

Interactive comment on "Are paleoclimate model ensembles consistent with the MARGO data synthesis?" *by* J. C. Hargreaves et al.

M. Kucera (Referee)

michal.kucera@uni-tuebingen.de

Received and published: 21 May 2011

The paper presents a simple but innovative approach for testing the adequacy of model ensembles by comparison of ensemble output with proxy-based sea surface temperature data for the last glacial maximum. It overcomes the difficulty of quantifying the differences between model outputs and proxy data by using rank histogram statistics combined with an explicit treatment of proxy data uncertainty. This is an elegant approach, which is shown not only to effectively characterize the properties of a model ensemble but also to identify specific features of model-data discrepancy. The methodology is not devoid of assumptions, but it in my view represents a significant step towards a meaningful framework for model-data comparison. The particular strength (and elegance) of the approach lies in the substitution of absolute differences in the

compared variables by their ranks, thus allowing a robust statistical analysis of the position of the observational (reconstructed) field within the model ensemble. The paper is clearly written, focused and presents a comprehensive analysis of the results. The interpretation of the analysis and the conclusions draw from it represent a significant advancement in the field. They appear well supported by data, are logically argued and their consequences for the understanding of how climate models perform in conditions outside of the present climate are clearly highlighted. There are several points that require clarification, but in general, I believe this paper is acceptable for publication with minor revisions.

General comments

1) The rank histogram method is simple, but it may appear difficult to conceptualize for a reader not familiar with it. I have struggled with understanding what exactly the method does until the results section and figure 3. The reason is partly found in a somewhat unclear formulation in the introductory section (see comments on the terminology below) and a lack of an explicit description of how the method works. I believe the paper would benefit from making the method more accessible for a wider audience in the introductory chapters.

2) I agree that the method of accounting for the uncertainty in the reconstructions by adding a random noise of the same statistical properties to each of the model outputs represents an easy and efficient solution to the probabilistic nature of the target variable, but I would like to authors to specifically mention that this approximation is based on the assumption that it is possible to correctly recover the size and spatial structure of the uncertainty and that this is random and symmetrically distributed around the central value for each discrete reconstruction. I concede that I cannot offer any better way of dealing with this issue but it appears fair to note that the way in which the uncertainties were assigned to MARGO SST values does not permit the recovery of the "true" shape and spatial distribution of the uncertainty interval, because many sources of uncertainty have been arbitrarily simplified by parametrisation to a normal

distribution, as stated on page 783 of the manuscript. A rudiment of such discussion is present at the end of the section 4.1.2. I believe this is an important issue, especially since the inclusion of data uncertainty has been shown to have a significant impact on the interpretation of the reliability of the PMIP2 ensemble.

3)I wonder whether the "tos" SST (Page 781, Line10) is really fully compatible with the MARGO SST, which is calibrated to represent the 10-m depth. I guess this will not make a large difference, but perhaps a small systematic offset, which could affect the rank of the observation with respect to the ensemble members?

4)Differences among the PMIP model generations in the extension of sea ice have been identified as a critical factor that could explain much of the high-latitude data-model discrepancy. In view of this, it would have been interesting to discuss how does the sea-ice extent implied by the ensembles actually compare with the sea-ice extent reconstructed by the MARGO project?

5)Considering the success of the rank histogram approach in comparing model output with the MARGO LGM reconstructions, I believe the authors should include an explicit discussion on how this work goes beyond the earlier attempts to compare MARGO and PMIP runs (Otto-Bliesner et al., 2009, Kageyama et al., 2006). This is justifiable, because although the main aim of the present paper is initially stated as a test of the reliability of the ensemble, the rank analysis allowed the authors to discuss spatial patterns and physical processes associated with the observed data-model differences and in this aspect the analysis overlaps with the aims of those (and other) papers.

Minor comments

Like Referee 1, I have problems with the legend to Table 2. In particular, the column headings are confusing and should be linked more to the actual meaning of the variables (Left and Right=?).

Page 776, line 20: "the ensemble is statistically indistinguishable from the truth". Here,

it is not clear to me what is meant by the "truth". If "truth" represents the actual state of the climate then this is a singular state and it cannot be compared to an ensemble? Perhaps the authors meant to state that the "truth" is contained in the ensemble? This seems to be the case based on the statement on Page 778, line 13, although in this case, the comparison is not with a known climatology, but with a probabilistic reconstruction. Perhaps the usage of the word truth is confusing in this context?

Page 777, line 25: "which additionally have heterogeneous uncertainties arising from the calibration of the proxies". This statement about paleoreconstructions is not entirely correct. The heterogeneous uncertainties do not arise from the calibration of proxies (such uncertainties are more likely to be homogenous) but by the use of different proxies and due to uncertainties in the representativeness of the obtained reconstruction for the considered LGM interval.

Page 778, line 1: the MARGO synthesis is based on six proxies, not three.

Page 778, line 10: "Probabilistic predictions are described as reliable if the frequency of occurrence equals the predicted probability, over a large set of instances." It is not clear from this sentence the frequency of what should equal the predicted probability?

Page 778, line 25: this sentence is absolutely critical to understand the method and yet it is presented in a confusing manner: "A consequence of reliability is that when the ensemble of outputs for a particular (scalar) variable, together with the observation representing the variable, is sorted from highest to lowest, the data lies equiprobably at each position in the rank order". What is meant by data? The next sentence exacerbates the confusion: "The rank histogram is simply a histogram of these ranks for all the data points under consideration, and so will be flat (to within sampling error) for a reliable ensemble." What exactly are the data points under consideration?

Page 791 Line 5: "We weight the squared model data differences by the grid box area and the MARGO uncertainty." This needs explanation – how exactly was the weighting carried out?

Interactive comment on Clim. Past Discuss., 7, 775, 2011.