Clim. Past Discuss., 7, C614–C622, 2011
www.clim-past-discuss.net/7/C614/2011/

Climate of the Past Discussions

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

T.L. Edwards (Referee)

tamsin.edwards@bristol.ac.uk

Received and published: 24 May 2011

1 General comments

I am relatively new to refereeing, but I hope these comments are useful.

The main part of this study is an application of a previous study (Annan and Hargreaves, 2010) to the field of palaeoclimate simulation. Rank histograms have been used for climate simulations before, but the recent suite of papers by the authors expands the field and opens the area up for debate. Some important questions must be asked about which assessments of weather forecast skill may be used in climate simulations, and I therefore very much welcome the authors' contribution. I also whole-

heartedly agree that palaeoclimate eras provide an important, independent test of model success and am happy to see the authors make comparisons with palaeoclimate reconstructions in parallel with those of the instrumental period. More traditional, PMIP-type, model-data comparisons are also made.

However, I find this paper is not as carefully put together as it could be. The most important problems for me are:

1. No sensitivity study for the assumption of spatial independence and effective dimension
2. An unwarranted inference of reliability and sufficient spread in the PMIP ensembles
3. Some parts of the text appearing to contradict the figures
4. Seemingly arbitrary downplaying of IPSL AMOC result (affects one conclusion)
5. Displaced MARGO data in Fig. 1 with respect to other figures
6. Asymmetric, wide, and inconsistent bins across histograms

If the additional sensitivity studies are carried out, and the text and figures are revised as set out below, I think it will be a useful contribution to the literature and appropriate for this journal. I have also made several requests below for the addition of figures and results to aid clarity.

2 Specific comments

2.1 Main scientific points

1. No sensitivity study for the assumption of spatial independence and effective dimension.

There is an implicit assumption of independence of model-data discrepancies when using rank histograms (Hamill, 2001). Applied to equilibrium climate simulations, rather than time-evolving weather forecasts, the assumption is of spatial rather than temporal independence. This is not valid for short length scales, which necessitates a reduction in the number of degrees of freedom for the chi-square tests, as described by Annan and Hargreaves (2010) and Jolliffe and Primo (2008). The use of fixed grid box sizes in degrees (rather than km) is likely to make this problem more severe for high latitude comparisons.

The effective dimension is crucial for the chi-square rejection tests, but no sensitivity to this independence assumption, or equivalently to the choice of effective dimension, is investigated in this study. The conclusions hinge on the choice made.

One method of testing the sensitivity would be to repeat the analysis with more sparsely sampled locations. Another would be to change the effective dimension, perhaps within the range 4-11 suggested by the cited paper (Annan and Hargreaves, 2011, J Clim).

A related point is this: given the strong link between ensemble size and effective dimension (Fig. 3 in Annan and Hargreaves, 2011, J Clim), can the authors explain why they appear to use the same effective dimension for the PMIP1, PMIP2 and JUMP ensembles? The figure would indicate values of 6, 5, and 9-10 respectively.

2. An unwarranted inference of reliability and sufficient spread in the PMIP ensembles.

Reliability implies a uniform rank histogram, but uniformity does not imply reliability. This is pointed out by Jolliffe and Primo (2008) and Hamill (2001). Therefore if the hypothesis of the ensemble being statistically indistinguishable from truth is rejected, we can say the ensemble is unreliable, but if it is not rejected, we cannot say that it is reliable. A separate, but related, point is made by Anderson (1996): consistency with observations does not guarantee usefulness. The text should be adjusted to reflect these.

e.g. p792 L1: PMIP1+2 ensembles are not shown here to be "reliable": rather they are not found to be unreliable.

3. Some parts of the text appearing to contradict the figures.

p785 L12 The text states that upwelling regions are too cool in the PMIP1 ensemble due to insufficient resolution, but the cool bias is worse in the higher resolution PMIP2 models (c.f. figures 3(a) and 5(a)). Can this difference be explained? Adding the map of PMIP1 ranks with MARGO uncertainties to Fig. 4 would enable robust comparison between the two ensembles.

p786 L14-18 This text appears to contradict Fig 5(a). In the figure, the models are warmer than MARGO near Greenland (red, high values on scale, "low rank") and cooler than MARGO further away. The text states the opposite. It would help to see the individual PMIP2 SST anomaly maps to compare with Fig. 1(a).

p790 L18 In four of the models the deepest SST spike is at 60-70N, not 40-50N, but the opposite is stated in the text ("location tends to be a little to the south"). The reason for this bimodal position is also not discussed.

4. Arbitrary downplaying of IPSL AMOC result.

p790 L20. I find it confusing and misleading to talk about 3 of 4 of the lowest AMOC without IPSL, then to add the IPSL value later. I would say 4 of 5 instead, or else 3 of 4 with the deepest spike at 40-50N have low AMOC.

p791 L20, p792 L28, Table 3, Figure 8. I think the IPSL max AMOC value (which weakens the correlation between small LGM AMOC anomaly and model-data SST agreement) should either be trusted or, if not, then left out entirely. At the moment it reads as if a conclusion was reached before obtaining the IPSL result and the authors are now reluctant to let it go.

In Table 3, I would ask that the AMOC-RMSE results are replaced with those including IPSL (both with and without ECBILT), i.e. n=9 and n=8. It will be interesting to see if there is still a correlation when including IPSL and excluding ECBILT.

5. Displaced MARGO data in Fig. 1 with respect to other figures.

In Figure 1, the map points are displaced by 5 degrees latitude with respect to all other figures (grid boxes have different locations with respect to lat/lon values and coastline). I don't think this is related to recalculation of MARGO on the displaced MIROC grid (p784). Even if it were, the visualisation should be the same throughout the paper for clarity.

6. Asymmetric, wide, and inconsistent bins across histograms.

In figures 3(c) and 4(b), the bins are asymmetric about zero (and rather wide). They should be symmetric, and ideally 1degC width. In figures 5-7 (c), the bins are symmetric and a different width to the above: the figures should be consistent. These choices make the results less clear and comparable.

2.2 Other scientific points

General

The "without uncertainties" analyses (Figs. 3 and 5) can be dispensed with, given that adding noise is considered essential by Hamill (2001) when observational errors are large. This does remove one of the conclusions of the paper.

Hamill (2001) point out that the interpretation of rank histograms under non-random sampling strategies is "not clear". Correlations across the ensemble are discussed by Annan and Hargreaves (2010) but should also be commented on here.

p779

L2 I would like to see some reference to the ergodic assumption (space-for-time substitution) inherent in the use of rank histograms for equilibrium climate modelling rather than time-evolving weather forecasting.

p787

L14-16 I would say 8/48 (16.7%) of MARGO points greater than 1 standard deviation from zero is about what is expected from a normal distribution (15.9%), i.e. MARGO is consistent with zero anomalies: not "provide low confidence of warming", as stated. Are any of these points outside 2 or 3 standard deviations?

p788

L20 It's probably worth commenting that the SAT variability at high latitudes is due to strong positive albedo feedbacks in this region (presuming this is the case).

p789

L5-7 Mention resolution too.

L13 "Wider spread in the PMIP1 results". They look the same to me, bar one outlier (and the PMIP1 ensemble is one larger, though of course the other 9 models are not the same).

L26 "over-extension of sea-ice" - this is conclusion is inferred jointly from Figures 2(b) and 8(a), not Fig. 2 alone, so it should be moved later in the paper and clarified.

p790 L5 The hypothesis that the AMOC was weaker and shallower isn't referred to again, which seems strange given the positive AMOC-RMSE correlation (though this may not survive the inclusion of IPSL, even excluding ECBILT).

p791

L9 Ideally I would like to see plots of the correlation analyses.

L11 Given the comment that ECBILT may be behaving rather differently and its exclusion from some results, can we see a map of model-data differences for ECBILT and another model with similar RMSE?

p792

L21 "agreement with the data" - The comparison between PMIP1 and PMIP2 is judged only by comparing model TAS with MARGO SST at low latitudes: this should be pointed out here.

L21-22, L25 There is confusion between comparing PMIP2 to PMIP1 (for sea ice extent) and PMIP2 LGM-present (for NHT) here.

L27 "PMIP2 is probably more consistent with the data in the high latitudes" - how can this be inferred when only the SATs are available for PMIP1?

Figures

2. Can we also have the JUMP LGM zonal means?

8. Please give the longitudinal bounds of the region.

3 Technical corrections

3.1 Suggestions for clarity

Use human-readable variable names (e.g. AMOC rather than stfmmc) throughout.

p789 L4 The majority of Section 4.2 is not a comparison of PMIP1 and PMIP2 (only p789 and one sentence about heat transport later). Suggest dividing this into two sections.

p780 L5 Can the model resolutions be given in Table 1?

p783 L11 add "assumption of" before stationarity

p783 L23 I'm not sure why the assumption that the MARGO error is the standard deviation of a normal distribution is this convoluted, e.g. introducing A and immediately setting it to 1.

p785 L7 Clarify that "high rank" corresponds to low values and blue on scale.

p788 L20 "considerable variation": clarify this refers to SSTs, 2(a), because there is more variation in TAS.

p790 L24 "PMIP1 slab-ocean models require..." -> "require" is strange. Replace with something like "have the same prescribed ocean heat transport in both simulations so the LGM anomaly is zero."

p791 L2 "more quantitatively" - than what: inspection of figures, or rank histograms? Or does this refer to studying this particular region rather than all MARGO data?

p791 Probably worth clarifying that small NHT corresponds to small AMOC error etc.

p792 Delete lines 4 and 7: "Rather it seems...improved" and "If we had not had...narrow" for clarity, as these points are made later in the page.

Table 2. Clarify with something like "Values less than or equal to 0.05 (bold) are significantly non-uniform (95

Figure 2. Remove PMIP2 TAS from (a).

3.2 Typing errors

p776 L20 the the

p782 L7 to accounting for

p783 L25 expert opinion *of* the

p787 L8 "area of low rank" -> high rank / low value (blue)

p788 L6 therefor

p792 L19 + L27 delete "the" or add "ensemble" (and check throughout paper)

Table 1. Delete question mark

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