Review 1: C.J. Van Meerbeeck (Referee)

Specific comments:
1. References:
One concern for the overall quality of the manuscript is the frequent self-referencing to either papers that have been submitted to – but are not yet accepted for publication in – peer reviewed journals (i.e. Annan and Hargreaves, submitted, 2010b; Yoshimori et al., in revision, 2011) or papers that are published in journals that are not peer reviewed or of which the status is not clear to me (i.e. Annan et al., 2005b; Hargreaves, 2010). Combined with the self-referencing of papers in peer reviewed journals, this practice appears somewhat inappropriate. Since, however, the present manuscript does not draw much on the conclusions of the referenced papers, the authors should reconsider whether those references are at all needed and delete them where not.

This issue resolved through an on-line author-reviewer exchange on 6 May, and the references have been updated accordingly.

2. p777 l8 and elsewhere:
The authors should avoid the use of 'epochs' which in geochronological terms refers to intervals of a typical duration of millions of years (thus far with the exception of the Holocene). The Last Glacial Maximum, however, is not considered an epoch in geochronology. Rather, Mix et al. (2003) define the LGM chronozone as the interval ~23-19ka BP.

Thanks for the correction. We have replaced "epoch" with "time" or "time period" throughout the manuscript.

3. p777 l12: . . . carbon dioxide level[s] substantially. . . 'substantially' is a matter of choice. Perhaps better would be less than half of today's (or 2011) values.

We changed "substantially" changed to "significantly", and mentioned the value of 185ppm.

4. p781 l18: The authors should include a brief justification of the use of both AOGCMs and AOVGCMs. In my opinion, one of the strengths of PMIP2 in this is its diversity of coupled model types.

We think this is inappropriate. It is not the role of this paper to justify the choices made by the model development community. We are deliberately stepping back from model development and analysing the ensembles based purely on their results, rather than on whether, theoretically, they ought to be better.

5. p782 l8: control model run
In terms of control run, the authors should specify which conditions are meant with 'modern'. Do they mean pre-industrial or present-day.

This is an interesting point and the answer is that it is neither, hence the use of the term "modern"! The climatologies used are typically 20-30 year climatological means, mostly from reanalysis projects and other data reconstructions. I have added to the text, "(20-30 year climatological means from a variety of sources representing late 20th century climate)". Fields representing "true" pre-industrial climate would, of course, be preferable, but they do not exist. We believe that other model errors dominate the possible inconsistency or bias caused by this weak tuning of the model towards these datasets.

6. p784 l9-11: T21 MIROC grid
Since the T21 spectral resolution corresponds to ~5.625○ horizontal grid resolution, which is slightly lower than the MARGO 5○ horizontal resolution, should that not in- crease the error of the re-

interpolated MARGO SST estimates? If so, then the MARGO SST constraints are less tight for MIROC evaluation.

No, the opposite is true. It is the PMIP1 and PMIP2 analyses which have error introduced from interpolation, not the JUMP analysis. MARGO was not *interpolated* onto the MIROC grid. As was stated in the text, the “MARGO dataset was recalculated from scratch onto the MIROC grid”. The “raw” data are located at the sites of the sediment cores. In the MARGO synthesis procedure, for any particular grid, those cores which lie within a grid box are averaged. We have re-phrased the sentence to say, “...the MARGO synthesis was re-derived from the original proxy data points onto the MIROC grid, for better comparison with that ensemble.“

7. p785 l17: effective dimension of 5
The authors should explain what is meant by effective dimension and why 5 is a fair assumption (without referring to the as yet unpublished Annan and Hargreaves 2010b paper).

The paper is now in press at Journal of Climate. In general you may make a request to the editor to see unpublished papers referenced in a paper you are reviewing. I would very gladly have shared them with you.

8. p786 l24-27: add reference to underpin the important conjecture that ‘prominence to the North Atlantic may not be unreasonable’, since it affects the interpretation of the results.

We thought this was probably widely accepted, but have now cited the AR4 Chapter 8.section 8.7.2.1.

9. p788 l24-26: the authors should better argue the use and presentation of the JUMP ensemble, since up to this point in the text it remains unclear. Moreover, the rationale of using a slab ocean version of MIROC is unclear.

We use the JUMP ensemble because we wish to compare the performance of single and multi-model ensembles. We have adjusted the introduction to make this clear. As for the rationale of using a slab ocean, the answer is for computational efficiency, but this is really point 4 again; the ensemble has previously been published in a number of papers cited in the text and here we test its performance rather than justify its existence.

10. Table 2 caption: it is unclear to me which values are associated with a statistically different from uniform distribution. Also, the use of bold type face denoting ‘unreliability’ should be explained.

Thanks for the correction. We have adjusted the caption.

Editorial comments:
1. p776 l2 ensembles of [climate] models
done

2. p776 l17 . . . CMIP3 ensemble [of global climate models] (Meehl . . .)
done

3. p776 l19 ... the the ... remove the
done

4. p777 l23 define ASGCM
done

5. p781 l6 define AOGCM
done

6. p781 l24 AOVG[C]M
done

7. p782 l23-24 The MARGO acronym has already been defined. Delete definition.
done

8. p783 l25 . . . the expert opinion [of] the MARGO . . .
done

9. p785 l1-3 We conclude that . . .
this sentence repeats what has been mentioned on p780 l20-228
The first mention is a summary and the second explains the details. We agree that it is a little awkwardly written, but leave it as it is as we feel that both parts need to mention what was concluded.

10. p787 l3-4 Combine the first two sentences of this paragraph into one topical sentence
done

11. p787 l17 Split up the paragraph after ‘occur.’ since the remainder discusses a different topic than the 307 MARGO data points.
done

12. p788 l20 Split up the paragraph after ‘model.’ References: Mix A (2003) Chilled out in the ice-age Atlantic. Nature 425: 32-33
done

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 temper- ature 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 to- wards 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 termi- nology 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.

We have attempted to do this, and hope you find it improved.

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 tat 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.

We agree that the parameterisation of “observational error” is an important aspect of the analysis and have added a brief discussion of this.

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?

This seems a little unfortunate. In Waelbroeck et al, the term “sea surface temperature” or “SST” is used throughout, with no mention of the calibration depth. Although it is also not explicitly stated, we suspect that “tos” output in the PMIP experiments is generally rather shallower (and know this to be the case for MIROC3.2). Perhaps the lack of clarity is evidence that most people expect the discrepancy to be modest. We have checked with MIROC3.2 and for the comparison of LGM-CTL anomalies, the discrepancy is indeed usually rather small (<0.1C) except for some very localised coastal areas where it may exceed 1C (based on results from MIROC3.2). Therefore, we don’t expect this to affect the results, and have added some clarification in the discussion of the MARGO data, Section 3.3.

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?

This is interesting information. We find that the data are not presently in a form that is easily accessible for model comparison (the type of information on sea-ice coverage - concentration, occurrence and/or presence - varies between the proxies), but hopefully this may be the subject of future work.

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.

Added a couple of sentences to the introduction.

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 vari- ables (Left and Right=?).

We have adjusted the caption for this Table in response to all 3 reviewers.

Page 776, line 20: “the ensemble is statistically indistinguishable from the truth”. Here, C559 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 ensem- ble? 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?

We have modified the wording. You are correct that the concept is of the truth (state) being contained in the ensemble (drawn from the same distribution as the ensemble members).

We don’t think “truth” itself needs changing as we are distinguishing between the actual underlying climatic state, and the imperfect observations. (Of course, only the latter is accessible to us.)

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.

We agree that the source of heterogeneity is the use of different proxies and the representativeness of the individual estimate for the considered time period, rather than the calibration error, and we changed the wording accordingly.

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

Corrected

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?

Wording clarified.

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 exac- erbates 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?

We have clarified the demarkation between observations and model output (“data” referred to the former).

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?

Explained more clearly.

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 Harg- reaves, 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 ex- pands 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 palaeocli- mate 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 di- mension
2. An unwarranted inference of reliability and sufficient spread in the PMIP ensem- bles
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).

We deliberately selected a value at the high end of the range that we considered plausible, as this provides the most stringent test, which the PMIP ensembles still pass (note that the value of 11 in the J Clim paper refers to precipitation, which has far more intricate spatial variation than temperature). Therefore the results of calculations with a lower assumed effective dimension are a foregone conclusion. The JUMP ensemble, on the other hand, remains irredeemably poor unless the effective dimension is set as low as 2. We've added a bit to describe this in 4.1.3.

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.

The effective dimension we are interested in is that of the underlying sampling distribution, not the apparent dimension calculated from the finite sample (which we know to generally be an underestimate). We have already compensated (to the best of our ability) for the bias of a small sample size, so the different ensemble sizes isn't an issue here.

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).

This is a good point, which we thought we had mentioned in Annan and Hargreaves 2010 (including the same Hamill reference), but which it turns out was cut for brevity there! So we are happy to have an opportunity to mention it here (2nd to last paragraph of Section 2).

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.

We have changed the text in 4.1.2 and 5.

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.

We are not attempting to assess the usefulness of the ensemble, which will be very much application-specific.

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

As above, we have changed the text in 4.1.2 and 5.

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?

It is a warm bias not a cold bias. Additionally, Figures 3(a) and 5(a) indicate the rank of each ensemble, not the size of the bias. Thus it is not possible from this plot to tell which ensemble has the larger bias in this area. Looking at the ensemble means, PMIP2 has a slightly smaller bias in this area than PMIP1. It is not very different, which is why we did not comment in the original manuscript. PMIP2 is still too low resolution to capture this kind of feature. We had thought that the histogram included sufficient information, but we have now included an additional figure showing the spatial differences in the anomaly between each ensemble and MARGO. We have also added some text to the manuscript describing the PMIP2 result off the west of Africa.

Adding the map of PMIP1 ranks with MARGO uncertainties to Fig. 4 would enable robust comparison between the two ensembles.

We do not think such a figure adds sufficient information to warrant another plot. The histograms in Figure 4 already present the information.

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.

Perhaps the wording was a little confusing, but a “high” rank means a high number (eg 11 rather than 1) and also that the observation is warm relative to the models. Hopefully the new figure and improved wording will help. Red is high rank, which means that the MARGO data indicate a warmer LGM relative to the present day than the ensemble. We have also added additional information to the figure captions.

The text states the opposite. It would help to see the individual PMIP2 SST anomaly maps to compare with Fig. 1(a).

Hopefully the new plot showing the ensemble means will be sufficient basis for comparison.

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.

For models with the spike in more or less the right place, the location tends to be a little to the south. I have adjusted the text to make this clear. I do not have an explanation for why in some models the spike is in completely the wrong place, other than what is already stated in the text (ie it indicates that the sea-ice in the model does not come far enough south at the LGM) .

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.

We agree with all this and also were not happy! The problem is that the IPSL data were not on the PMIP2 database. Only once the paper had already been drafted did we become aware that a max. AMOC value for the IPSL model was in the literature. However we wanted all the models to have been analysed the same way, and for the max. AMOC there are small variations in the calculation which can influence the value obtained (latitude and depth range considered, grid resolution used), so we were not so confident about the IPSL value in the context of our analysis, which is why the result was down-played.

However, IPSL researchers saw this paper in CPD, and decided to create the variables. We received them by personal communication, but understand that they have just been added to the

PMIP2 database. The value I calculate for the max AMOC in IPSL is indeed a little different from the previously quoted value.

Now we have NHT and AMOC for IPSL, we have replaced the table of correlations with a Figure and re-written the text as appropriate. Plotting the figure was very useful as I found a mistake in the correlation analysis. There is actually a strong correlation between the NHT and AMOC.

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.

Fixed

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.

Histograms corrected. As explained in the text, a random value scaled by the MARGO uncertainty value is added to each model point. As the random number seed was not frozen before initially submitting the paper, for the calculations where MARGO uncertainty is included, the histograms look slightly different in the manuscript.

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.

We disagree with the suggestion to remove this section. Although the importance of accounting for observational uncertainty is often mentioned (eg Knutti et al 2010 "Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections"), this detail is not infrequently overlooked in practice. Moreover, it is still quite rare to find usable quantitative error statistics accompanying paleoclimate data analyses, and even the MARGO "relative error" value is a rather vague concept. It also may not always be obvious as to when the observational errors are large enough to matter. Therefore, we think it is very important to highlight the importance of future development in this area.

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.

We are not sure of the point here. The paradigm is of a distribution from which the models are sampled randomly. Obviously this is a somewhat hypothetical construct as we have no practical means of drawing an unlimited (or even large) number of samples from this distribution (which itself evolves through time).

p779

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

It is not so much a space for time substitution, as that we are only sampling in space. Generic NWP applications may sample in both space and time (eg Jolliffe and Primo). There is no assumption, for example, that a positive bias in one location will or will not change sign at other time periods, merely an assessment that (under this analysis) the existing data are not inconsistent with the paradigm of a reliable ensemble.

It would also seem natural to use a larger number of paleoclimate times in the future, depending on the availability of adequate data, hence filling out the time dimension.

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?

While this may be technically correct, we didn't want to seem too dogmatic about it, especially as the estimated magnitude of errors (ie the scaling on the MARGO relative error) was based on our own judgement. Our analysis does not seem sufficient basis to actually disagree with Waelbroeck et al who do refer to the local warming.

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).

We have not done the feedback analysis of all the PMIP models, so cannot comment on what is causing the differences between models. Actually this work is underway by a colleague at Tokyo University, as a follow-up to Yoshimori et al 2011.

p789

L5-7 Mention resolution too.

There have been a wide range of general model improvements and changes, including resolution. The sentence in question is a statement of belief rather than fact and we still think that the largest difference in the ensemble in general is the incorporation of dynamical oceans. We have, however, now mentioned higher resolution as a feature of model improvement in section 3.1.

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).

It seems like you might have misread the figure. Figure 2(b) clearly shows that for most latitudes there are several PMIP1 ensemble members that are outside the range of the PMIP2 ensemble. Remember, we are focussing on mid-latitudes, not the polar region where this is less clear.

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.

No, this conclusion was arrived at before the analysis of the Atlantic was done. The purpose of Figure 8 is to show the results for just the Atlantic ocean.

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.

The correlation analysis has been completely re-written. The observationally based estimate of the AMOC is referred to again in the Conclusion.

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

No longer relevant now that the section has been re-written.

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 ex- tent) 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.

We have re-written the paragraph. For the last question, see Paragraph 1 section 4.2. The Figure has been added. The data and models are global in longitude. If you mean the Atlantic, the longitudinal bounds vary with latitude.

3 Technical corrections

3.1 Suggestions for clarity

Use human-readable variable names (e.g. AMOC rather than stfmmc) throughout. C620

These variable names are only used in the Table, and their description is included in the table caption. As the PMIP databases are not particularly easy to navigate (and we found several errors) we consider it important to make it clear precisely which variables were used in the analysis. The variables names are the unique identifiers.

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.

We have changed the title of the section to broaden its scope.

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

There are many model differences that could be included, and we do not see a particular reason to single out resolution. Basic information on the model components, resolution etc may be found in the cited references.

p783 L11 add "assumption of" before stationarity

Done

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.

We think it is worth emphasising both that it is necessary to introduce a conversion factor to use the MARGO value as a numerical error estimate, and that in this case, we specifically chose the factor 1.

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

It is the other way round. We have clarified the figure captions.

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

Now refers to subplot (c) so all should be clear.

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."

Wording changed.

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?

More quantitatively than the discussion up to this point in the paper. We have removed “more” for clarity.

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

Section re-written in light of new data and corrected NHT calculation.

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.

Now that the third paragraph of the conclusions has been largely re-written, these sentences should remain.

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

done

Figure 2. Remove PMIP2 TAS from (a).

The point of this subplot is to compare TAS and SST. Thus SST is retained. As SST is the primary variable for comparison with MARGO we have, however, plotted the SST on top of the TAS to make it more visible.

3.2 Typing errors

p776 L20 the the

done

p782 L7 to accounting for

done

p783 L25 expert opinion *of* the

done

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

wrong

p788 L6 therefor

done

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

done

Table 1. Delete question mark

done

References:

Yoshimori, M., Hargreaves, J., Annan, J., Yokohata, T., and Abe- Ouchi, A.: Dependency of Feedbacks on Forcing and Climate State in Perturbed Parameter Ensembles, Journal of Climate, (in press), 2011.