Clim. Past Discuss., 8, 3481–3511, 2012
www.clim-past-discuss.net/8/3481/2012/
doi:10.5194/cpd-8-3481-2012

Climate of the Past Discussions

Skill and reliability of climate model ensembles at the Last Glacial Maximum and mid Holocene

J. C. Hargreaves[1], **J. D. Annan**[1], **R. Ohgaito**[1], **A. Paul**[2], **and A. Abe-Ouchi**[1,3]

[1]RIGC/JAMSTEC, Yokohama Institute for Earth Sciences, Yokohama, Japan
[2]University of Bremen, Bremen, Germany
[3]AORI, University of Tokyo, Kashiwa, Japan

Received: 20 July 2012 – Accepted: 24 July 2012 – Published: 13 August 2012

Correspondence to: J. C. Hargreaves (jules@jamstec.go.jp)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Paleoclimate simulations provide us with an opportunity to critically confront and evaluate the performance of climate models in simulating the response of the climate system to changes in radiative forcing and other boundary conditions. Hargreaves et al. (2011) analysed the reliability of the PMIP2 model ensemble with respect to the MARGO sea surface temperature data synthesis (MARGO Project Members, 2009) for the Last Glacial Maximum (LGM). Here we extend that work to include a new comprehensive collection of land surface data (Bartlein et al., 2011), and introduce a novel analysis of the predictive skill of the models. We include output from the PMIP3 experiments, from the two models for which suitable data are currently available. We also perform the same analyses for the PMIP2 mid-Holocene ensembles and available proxy data sets.

Our results are predominantly positive for the LGM, suggesting that as well as the global mean change, the models can reproduce the observed pattern of change on the broadest scales, such as the overall land-sea contrast and polar amplification, although the more detailed regional scale patterns of change remains elusive. In contrast, our results for the mid-Holocene are substantially negative, with the models failing to reproduce the observed changes with any degree of skill. One likely cause of this problem may be that the globally- and annually-averaged forcing anomaly is very weak at the mid-Holocene, and so the results are dominated by the more localised regional patterns. The root cause of the model-data mismatch at regional scales is unclear. If the proxy calibration is itself reliable, then representation error in the data-model comparison, and missing climate feedbacks in the models are other possible sources of error.

1 Introduction

Much of the current concern over climate change is based on long-term forecasts from climate models forced with increased GHG concentrations due to anthropogenic

emissions. However, a direct assessment of the predictive performance of the models is not generally possible because the time scale of interest for climate change predictions is typically for decades or centuries into the future, so we cannot build up confidence and experience via repeated forecasts on a daily basis as is typical in the field of weather prediction. Therefore, in order to have confidence in the ability of the ensemble to provide a believable projection of future climates, we must try to develop other methods for assessing the performance of models in simulating climates which may be very different to today.

Paleoclimate simulations provide us with an opportunity to critically confront and evaluate the performance of climate models in simulating the response of the climate system to changes in radiative forcing and other boundary conditions. A particularly attractive feature of using paleoclimate simulations is that (in contrast to the situation regarding more recent climate changes) it is uncontentious that the performance of models over these intervals has not been used in their development. Therefore, these simulations provide a truly independent test of model performance and predictive skill under substantial changes in external forcing. The extent to which such assessments may then be used to imply skill for future forecasts is still, however, open to some debate, since not all the past climate changes are necessarily relevant for the future. Therefore, models which provide the most realistic simulations of past changes may not necessarily provide the most accurate predictions of future change. Nevertheless, the potential of such assessments to help in evaluating model performance provides strong motivation for research in this area. The sparse and semi-qualitative nature of paleoclimatic data (for example, interpretation as vegetation type) has motivated the development of advanced but semi-quantitative methods, such as using fuzzy logic to measure model-data mismatch (Guiot et al., 1999) and cluster analysis to classify types of model behaviour (Brewer et al., 2007). With the advent of new more comprehensive syntheses of gridded paleodata (MARGO Project Members, 2009; Bartlein et al., 2011), it becomes possible (if not compelling) to attempt more directly quantitative analyses of model performance, which we undertake here.

The second phase of the Paleoclimate Modelling Inter-comparison Project, PMIP2, established a common protocol of boundary conditions for two different paleoclimate intervals, the Last Glacial Maximum (LGM, 21 ka BP) and the mid Holocene (MH, 6 ka BP). Of the two, the LGM represents by far the greatest change in climate with significantly decreased concentration of atmospheric carbon dioxide (and other greenhouse gases), and large ice sheets over the northern hemisphere high latitudes. The forcing of the mid-Holocene is more subtle, with the only changes considered by the models being that of orbital forcing, and a moderate decrease in atmospheric methane. While this results in substantial changes in the seasonal and spatial pattern of the insolation, the net change in the annual and global mean is rather small. Rather than large global changes, perhaps the largest climatic changes during the MH related to shifts in monsoon patterns, with associated vegetation changes (Braconnot et al., 2007b).

In this paper we extend our previous work presented in Hargreaves et al. (2011), hereafter H11, which assessed sea surface temperature at the LGM. We use several state of the art proxy data syntheses for surface temperatures for both the LGM and MH, and compare them to outputs from the coupled atmosphere and ocean (AOGCM) and coupled atmosphere, ocean and vegetation (AOVGCM) general circulation models in the PMIP2 database. For the LGM we additionally include two models from the 3rd phase of PMIP, PMIP3 (those for which sufficient output are available at the present time). We perform analyses based on quantitative model evaluation methods which are widely used in numerical weather prediction. We present a rank histogram analysis to indicate reliability, Secondly we introduce a skill analysis using two different reference baselines. We also we present Taylor diagrams (Taylor, 2001). These diagrams are a way of summarising three conventional statistics, and have been widely used to analyse climate model ensemble output in the context of the modern climate. In addition, we introduce some simple modifications to these conventional statisics to account for observational uncertainty.

In this paper, we introduce the models and data used in the analysis in Sect. 2. Then we overview the methods for analysis of reliability and skill in Sect. 3. In Sect. 4 we

present the results from the LGM and MH, and this section is followed by the conclusions.

2 Models and data

2.1 Last Glacial Maximum

All the PMIP2 models analysed here are either physical coupled climate models comprising atmosphere, ocean and sea ice components (AOGCMs), or additionally including vegetation modules (AOVGCMs). One of the models, ECBILT, is an intermediate complexity model (see Table 1). For the LGM, the forcing protocol (Braconnot et al., 2007a) comprises a set of boundary conditions including large northern hemisphere ice sheets, altered greenhouse gases including a reduction to 185 ppm for atmospheric carbon dioxide, a small change in orbital forcing, and altered topography. There are some minor changes in the multi-model ensemble compared to our previous analysis of this ensemble (H11). Firstly, the output of the run from IPSL has been updated. In addition, previously the number of days in each month (which differs between models) was not taken into account when calculating the annual mean from the monthly data. This has been corrected, making a small difference to the values of the annual mean obtained. We use both surface air temperature (SAT) and sea surface temperature (SST) and therefore use the 9 models in the database for which both these variables are available. From the new PMIP3 experiments, which were downloaded from the Coupled Model Intercomparison Project (CMIP5) database, we include the two models in the database for which both SAT and SST are available. Thus we have a total of 11 models. For PMIP3 the boundary conditions are slightly revised, primarily in relation to the ice sheet reconstruction (see: http://pmip3.lsce.ipsl.fr/) but still remain sufficiently similar to those for PMIP2 that a priori it seems reasonable to investigate the features of the larger ensemble by combining both PMIP2 and PMIP3. We also consider down-weighting models from centres which have contributed more than one model to our

ensemble, to account for their likely similarity (Masson and Knutti, 2011). Table 1 indicates which model versions are used in each ensemble.

For comparison with the models we use an updated synthesis of temperature data representing sea surface temperature, plus land temperature from pollen data and a small number of estimates from ice cores. This dataset has been previously used by Schmittner et al. (2011). The ocean data comprise the MARGO synthesis (MARGO Project Members, 2009) with a small number of points having been updated. These updated points may not be fully homogenous with the original MARGO dataset as the data error has not necessarily been estimated in an identical way. The land pollen data points come from Bartlein et al. (2011) (hereafter B11), which is a somewhat more ad-hoc data set than the MARGO synthesis, in that the error estimates have been directly drawn from the original literature in which the underlying data were presented, rather than being recalculated homogeneously across the data set as in MARGO Project Members (2009). In addition, the temperature anomalies in B11 are taken relative to the core tops in contrast to the modern World Ocean Atlas data that were used to anchor the MARGO anomalies. Ice core error estimate were derived through a variety of methods. The SST data are analysed on the MARGO 5 degree grid, while all the land points are on a 2 degree grid. After removing grid points for which SST information is unavailable in one or more models (due to their differing land sea masks), there are 309 SST points left for comparison with PMIP2, and 300 points for comparison with PMIP2 + PMIP3. Our goal is to assess the model response to imposed forcing, rather than the forcing itself, so for the land data we remove those points for which 50 % or more of the grid point lies under the model's ice sheet. This affects 11 points, leaving 95 land points for both PMIP2 and PMIP3. Thus we have a total of 404 points for comparison with PMIP2 and 395 for comparison with the combined PMIP2 and PMIP3 ensemble.

Estimates of the data error uncertainty are included for all the data points, although we note that the MARGO errors are only defined in terms of their relative reliability, so as in H11, we assume Gaussian uncertainties scaled by 1 °C. The resulting errors

range from 0.24 °C to 6.4 °C across the data set. The model SST output was interpolated on to the 5 degree MARGO grid and the SAT onto the 2 degree B11 grid. We use equal weighting for each grid box. The data, multi-model mean and data uncertainty are shown in Fig. 1.

2.2 Mid Holocene

For PMIP2 the mid Holocene protocol includes only two changes compared to the pre-industrial climate. The orbital forcing is changed, and the atmospheric methane is moderately decreased (from 760 ppb to 650 ppb). The orbital forcing changes the seasonal and large-scale spatial pattern of insolation. Globally and annually averaged, the insolation is the same for pre-industrial and 6ka, and so we expect to see only a small signal in the annual temperature. Therefore, in addition to annually averaged temperature, we consider representations of changes in the seasonal temperature signal. There are 11 AOGCM and 6 AOVGCMs in the PMIP MH ensemble that have both SAT and SST data available. There is only one AOVGCM, ECHAM, that does not have a counterpart AOGCM in our ensemble (see Table 1).

For the land temperatures at the MH we use the pollen-based dataset of B11. This synthesis includes estimates of annual average temperature as well as the temperatures of the hottest and coldest months, which indicate changes in the seasonal cycle. An uncertainty estimate is also included for all points, which ranges from 0.04 to 4.8 °C over all the variables. The number of data points varies slightly between the different variables (between 615 and 638), and the data are very clustered with high density in Europe and North America. The data, and the data error for the hottest month are shown in Fig. 2.

The "GHOST" SST dataset (Leduc et al., 2010) contains annual average estimates of annually averaged SST at both 6 ka and core-top for only 81 sites, and, while recognising that the core top is not an ideal or wholly consistent reference point (as the dates, and therefore climates, represented by the core tops may vary across the data set), we nevertheless take the difference between these two values to represent the annually

averaged MH temperature anomaly with respect to the pre-industrial climate. Seasonal SST data are, as yet, unavailable. Many of these points are quite close to the coast, and due to varying coastlines in the models, SST output from all models is available for only 42 points. Data uncertainties are not readily available so for this analysis we assumed a 1 standard deviation error of 2 °C for all the points, which is representative of the data uncertainty of the MARGO SSTs.

It is clear that for the MH the data coverage is substantially more sparse and less uniform than for the LGM. As with the LGM analysis, we give equal weight to each data point. It is possible that the hottest month in the tropics may not be the same for the models and data due to the ways the calendars are configured (Joussaume and Braconnot, 1997). We have few data in the tropics, and the actual error in the value of the anomaly is expected to be small, so we do not expect this to have a significant effect on our results.

3 Ensemble analysis methods

3.1 Reliability

To assess the reliability of the ensembles we adopt the same approach used in several recent papers (Annan and Hargreaves, 2010; Hargreaves et al., 2011; Yokohata et al., 2011), in which we interpret the ensemble as representing a probabilistic prediction of the climate changes and assess its performance by means of the rank histogram (Annan and Hargreaves, 2010) formed by ranking each observation in the ensemble of predictions for each data point. In the case of a perfectly reliable ensemble (meaning that the truth can be considered as a draw from the distribution defined by the ensemble), the rank histogram would be flat to within sampling uncertainty. For an ensemble that is too wide such that the truth is close to the mean, the histogram is dome shaped. Conversely an ensemble that is too narrow (often not including the truth) has a U-shaped rank histogram, with large values in one or both end bins. The analysis for

the PMIP2 LGM ensemble using only the MARGO dataset was already performed in H11, and overall produced encouraging results. In this analysis we extend this test by including land data. We also analyse the performance of the ensemble at the MH, for which a moderately large ensemble of runs is available.

In order to make the models and data comparable, we follow the same procedure as H11 to account for data uncertainty, by adding random deviates of appropriate magnitude to each model output before calculating the rank histogram. We use the same statistical tests as H11 to quantify the significance of any divergence from flatness of the rank histogram (Jolliffe and Primo, 2008). The rank histogram test is only a necessary but not sufficient test of reliability, in that an ensemble which does not have a flat rank histogram may be considered unreliable, but a flat rank histogram does not necessitate reliability (Hamill, 2001). Indeed, with a large enough ensemble and fine data coverage we generally expect the ensemble to be unreliable at some level, and thus this test is not whether the ensemble is perfect, but rather whether the limitations are so substantial as to be immediately apparent with this standard test. In previous applications such as H11 and Yokohata et al. (2011), it has been found that the ranks of nearby observations are often highly correlated, and therefore the effective number of degrees of freedom are substantially lower than the number of data points. In order to estimate the number of degrees of freedom, we adopt the EOF analysis approach of Bretherton et al. (1999), as used by Annan and Hargreaves (2011). Analysing the LGM fields of SAT and SST for the model ensembles, we estimate a value of 8 degrees of freedom, which we use throughout our analyses. Our results are insensitive to reasonable changes in this value.

3.2 Skill

The concept of model skill in climate science has been occasionally touched upon, but it has rarely been clearly and quantitatively defined (Hargreaves, 2010). Here we use the term skill in the sense in which it is in common use in numerical weather prediction, which is as a relative measure of performance: skill compares the performance

of the forecast under consideration, to that of a reference technique (Glickman, 2000). Perhaps the most straightforward of these is the skill score (SS) defined by

$$\mathrm{SS} = 1 - \left(\frac{E_\mathrm{f}}{E_\mathrm{ref}}\right) \tag{1}$$

where E_f is the error of the forecast under evaluation, and E_ref is the error of the reference technique. A perfect forecast will have a skill score of 1, one which has errors equal to that of the reference will have a skill score of 0, and a negative skill score implies that the errors of the forecast are greater than that of the reference. It is conventional to describe a forecast with positive skill score as "having skill", but note that this is always defined relative to a specific reference. So for example, using a root mean square difference to evaluate the forecast, one has,

$$S = 1 - \sqrt{\frac{\sum_i (m_i - o_i)^2}{\sum_i (n_i - o_i)^2}} \tag{2}$$

where m_i are the forecast results, o_i the data and n_i the reference. According to this equation, however, in the presence of error on the data, even a perfect model (where the model prediction precisely matches reality) would not achieve a skill of 1. Especially in the case of paleoclimates, the uncertainty on the data is often substantial (Hargreaves et al., 2011), and must be taken into account for a fair evaluation. We therefore modify Eq. (1) as follows:

$$S = 1 - \sqrt{\frac{\sum (m_i - o_i)^2 - \sum (e_i^2)}{\sum (n_i - o_i)^2 - \sum (e_i^2)}} \tag{3}$$

where e_i are the one standard deviation uncertainties on the observations. The numerator and denominator in the fraction have here been adjusted to account for the observational errors, under the assumption that these are independent of the forecast errors.

Note that the skill score here becomes undefined if either the model or the reference agrees more closely with the data than the data errors indicate should be possible. Such an event would be evidence either that the data errors are overestimated, or else that the model had already been over-tuned to the observations. In principle, no model should agree with the data with smaller residuals than the observational errors, since even reality only just achieves this close a match, and then only if the observational errors have not been under-estimated.

For the obvious reason that forecasts are have not been generally realised in climate models, skill analyses for climate model predictions are rare. One simple analysis was performed by Hargreaves (2010), which indicated that, at least on the global scale, the 30 yr forecast made by Hansen to the USA congress in 1988 had some skill (regional data were not available for testing). We are not aware of previous analyses of model skill for paleoclimates and it has not been established what might be an appropriate reference forecast for such calculations. In numerical weather prediction, persistence (that tomorrow's weather is the same as today's) is a common baseline for short-term forecast evaluation, and seasonal prediction (where persistence is clearly inappropriate) may use the climatology as a reference. An analogous reference for climate change predictions might be that the climate persists, that is, a reference of no change. It should be clear that this is a rather minimal baseline to beat, only requiring that the model predicts any forced response at each location to within a factor of anywhere between 0 and 2 times the correct amplitude (on average). We might reasonably hope for our models to perform rather better than this, and provide a useful prediction not only of the overall magnitude, but also the spatial pattern of change. Thus we also employ a second reference to tests the pattern of the change more directly. For this reference forecast, the climate change is assumed to be a uniform change equal to the mean change of the available data. This represents the case of a perfectly-tuned zero dimensional energy balance, in which the global mean temperature change is predicted which optimally matches the data, but without any information on the spatial pattern.

In order to have skill with respect to this reference, the model must also represent the spatial pattern of change.

3.3 Conventional Taylor diagram analysis

We also present an analysis of the model outputs in terms of the conventional statistics of (centred) RMS difference, correlation and field standard deviation which will be familiar to many readers. Such values are conveniently presented in a Taylor diagram (Taylor, 2001) which summarises these three values with a single point. The usual calculation and presentation of these statistics does not account for observational uncertainty, and Taylor (2001) only suggests investigating the effect that this might have on the results through the use of multiple data sets, which we do not have here. However, since we do have estimates of observational uncertainty, we can instead adjust the statistics to account for this. We present our results based on two approaches. First, we present the conventional results, without accounting for observational uncertainty, but also indicate where a hypothetical "perfect model" (which exactly matches the real climate system) should be located. This is straightforward to calculate, as its RMS difference from the actual observations should equal the RMS of the observational errors, and the observational standard deviation is the sum (in quadrature) of that of the underlying (but unknown) true field, and the errors.

An alternative approach, is to correct the statistics for each model, to indicate where they should be relative to perfect observations. This is also a simple calculation when error estimates are known.

4 Results

4.1 Last Glacial Maximum analyses

4.1.1 LGM reliability

Figure 3 shows the reliability analysis for the combined LGM ensemble of 11 models for all the points at which we have data. Overall, the ensemble has a rank histogram which cannot be statistically distinguished from uniform (Fig. 3b), and the differences between the data and the ensemble mean (Fig. 3c) are mostly of similar magnitude to the uncertainty in the data. Looking at the map in Fig. 3a, there are some patches that are predominantly red or blue, indicating the spatial limit to the reliability. Analysing the ocean and land data separately (Fig. 4) we find that, assuming 8 degrees of freedom, the ensemble is statistically reliable with respect to both. However we note that the histogram for the land (Fig. 4c) has a fairly large peak at the left hand side, indicating that the ensemble tends to have a greater anomaly than the data. It is also apparent that the difference between the ensemble mean and the data is larger for the land than for the ocean (Fig. 4d). This model-data difference exceeds the quoted data uncertainty much more frequently for the land than for the ocean. Paleoclimate data are derived from measurements made from cores drilled into the surface of the earth at discrete locations. The open ocean may be considered quite well mixed, whereas land has many more local features due primarily to high resolution topography. Thus it may be more difficult to derive a representative grid box average temperature from the data for direct comparison with the models over the land, than over the ocean. On the grid scale of the models, one sees more variation over land than ocean, but even so it is likely that the resolution at the finest scales there is under-represented (due to smoothing in the forcing, boundary conditions, and dispersion in the model numerics). Thus it is understandable that the model data mismatch is greater over the land. It is important that more work is done to identify, quantify and, if possible, reduce these

kinds of representation errors between the models and data so that future model-data comparisons can be as informative as possible.

To create the ensemble, we simply aggregated all the AOGCM and AOVGCMs available in the PMIP2 and CMIP5 databases. While there has been discussion in the literature about the similarity of models based on their origins and design (Masson and Knutti, 2011) there is little consensus about how to treat this, so we start from the premise of assigning equal weight to each model. Including two model versions that are identical to each other would be clearly pointless and indeed harmful as it would be equivalent to double-counting one model and would actually reduce the effective sample size of our ensemble, thereby degrading the results. We do not have identical models in this ensemble, but it is possible that some of the models are very similar. For example, we already know that the only difference between MIROC3.2 and MIROC3.2.2 (which are both in the PMIP2 database) is that one minor bug has been fixed in the latter, so we included only the latter model in the ensemble. For substantially updated versions of the same model (such as might exist in consecutive iterations of the PMIP experiments) we would expect the differences to be rather greater. But, ideally we should wish for each new model included in the ensemble, to be not particularly more similar to one existing model than it is to all the others.

With this in mind, we performed some sensitivity analyses into our treatment of the ensembles. Excluding the two PMIP3 models which provide both SAT and SST does not make a difference to the level of reliability, a result consistent with the suggestion that CMIP3 and CMIP5 models do not appear to have substantially different behaviour in distribution (though with only two PMIP3 models, any difference would be hard to detect in any case). Analysing SAT alone for the four PMIP3 models presently available for which earlier model versions were also presented in PMIP2, and comparing them to the combined PMIP2 and PMIP3 database (14 models in total) in terms of bias, correlation and mean square difference, we find that the PMIP2 version of the same model in the PMIP3 database is not usually the most similar model but is usually in the top three, although it may be as low as sixth. We also analysed the PMIP2 AO and

AOV models in the same way, and found that they were very similar in terms of RMS difference and spatial correlation, although they typically have a significant mean bias. From the similar pairs, it is not clear which model should be excluded. Thus, in order to make a conservative ensemble on which to test the robustness of our results, we reanalysed the ensemble giving half weight to each member of each pair of related models in the ensemble. In the case of our 11 member ensemble, this means that the PMIP2 and PMIP3 IPSL models had half-weight each as did the PMIP2 HadCM2 AO and AOV models. The results of the rank histogram analysis are not significantly different for those using the whole ensemble equally weighted.

4.1.2 LGM skill

Figure 5 shows the results for the skill calculations for the LGM anomaly. For the PMIP2 and PMIP3 models we analysed ocean and land both separately and together, and we also calculated the skill for the multi-model mean along with each model individually. For the first reference forecast, of a zero LGM anomaly, there is skill for both the land and the ocean individually, and both combined. As expected (Annan and Hargreaves, 2011) the multi-model mean performs relatively well. Thus we can see that in general the models are producing a cooling that, overall, is of the same scale as the data. As mentioned above, this is, however, a rather limited test. A skill score of 0.5 indicates that the modelled anomalies are typically 50 % greater or smaller than observed.

The second reference forecast is of a uniform field equal to the data mean. This provides a much greater challenge to the models, as they have to not merely reproduce a broad-scale cooling of the correct magnitude (which the reference forecast already achieves), but must also represent the spatial pattern and magnitude of changes. While on the face of it, this still does not seem like a highly challenging requirement (given the well-known phenomena due to land-ocean contrast and polar amplification), none of the models have high skill against this reference, and in fact more than half the models have negative skill when assessing the land and ocean separately (which eliminates the strong influence of the land-sea contrast). The skill of the multi-model mean is

generally greater, especially for the ocean where it outperforms all of the ensemble members, and is also positive for the land. This indicates that the broad scale features which are what remains after the models are averaged, do bear some relation to the spatial pattern in the data.

The combination of land and ocean together shows much improved skill for all the models compared to land and ocean separately, indicating that the models are at least to some extent capturing the land-sea contrast in the changes at the LGM. This is encouraging, especially in light of recent work (Schmittner et al., 2011) in which an intermediate complexity model underestimated this land-sea contrast significantly. In the data, the LGM anomaly is larger over the land than the ocean, with the simple averages of the data points over these regions being −6.53 °C and −1.98 °C respectively, giving a land-ocean ratio of 3.30. For the models, the averages over the datapoints are from −3.59 °C to −9.00 °C over the land and from −1.88 °C to −2.95 °C over the ocean, and the land-ocean ratios range from 1.90 to 3.09. So the data ratio is outside the model range, but not necessarily to an alarming degree. It seems plausible that missing forcings (such as dust forcing) may have more effect over land than ocean (Schmittner et al., 2011), which could imply the error is more in boundary conditions rather than models themselves. Overall it must be noted that the levels of skill are not particularly high based on either of the two reference forecasts, suggesting that much of the regional to fine scale spatial pattern of the change is not being reproduced by the models and that there is plenty of scope for improvement.

4.1.3 LGM Taylor diagram

Conventional statistics are presented in the form of a Taylor diagram in Fig. 6. The modelled LGM anomalies shown in the top plot have correlations with the data which range from around 0.4 to almost 0.7, with a centred RMS difference which is somewhat lower than the standard deviation of the data itself (which is 3.5 °C in this case). However, observational errors are quite large, as is indicated by the location of the theoretical “perfect model”. The lower plot shows that if instead we had “perfect data” with

no observational uncertainty, we could expect the correlations to mostly lie in the range 0.6–0.8 and the RMS difference from the data would also be substantially smaller.

4.2 Mid Holocene analyses – reliability and skill

For the MH we have no seasonal information for the ocean so most of our analyses were for the land only. These comprised the anomaly for the annual average, and the hottest and coldest months. Given the nature of the forcing, the change in the magnitude of the seasonal cycle would seem the most obvious target for the MH, but there is some concern that the calculating the anomaly of the hottest month minus coldest months may not provide a completely fair comparison with the data as the same proxies are not used to compile the two datasets. For the land and ocean together we analysed the annual average anomaly only. For the MH we have several models with both AOGCM and AOVGCMs versions (see Table 1), so, as well as the full ensemble we also analyse a conservative ensemble where versions of the same model are downweighted so that each underlying model has a total weight of 1.

In comparison with the LGM, for the MH we have far more land data points and far fewer ocean points. As discussed in Sect. 4.1.1, the models tend to match the data less well over land than ocean, so a larger overall model-data mismatch for these analyses does not necessarily mean that the models are intrinsically worse at simulating this interval. For the LGM the model skill was poorer when tested on smaller spatial scales (through the second reference forecast), but for the MH the climate forcing is not expected to cause climate change at the larger scales. For these reasons, we expect this interval to provide a greater challenge for the ensemble.

The results obtained are indeed mostly negative for both the reliability and skill analyses. Of all the analyses, only three ensembles are not shown to be unreliable (through significantly non-uniform rank histograms), and even these are clearly tending towards being U-shaped (see Fig. 7). These are the mean temperature anomaly for land and ocean for the conservative ensemble, and the coldest month anomaly for both the full and conservative ensembles. The anomaly for the hottest month is unreliable. This

anomaly is also considerably larger in the models than that of the coldest month, and thus although we do not directly test it, we anticipate that the anomaly in the seasonal cycle is also unlikely to be reliably predicted by the models.

For the skill analysis, the picture is even worse, with most models having negative skill and no models or model means having skill greater than 0.01 (not shown). There is no indication in the skill results, or in an analysis of the RMS model-data differences, that the AOVGCMs models perform any better than the AOGCMs. This is somewhat disappointing as it is widely thought that vegetation feedbacks had a strong influence on the climate of the MH interval. The nearest thing to a positive result in these analysis is the relatively good result for the land and ocean together in the conservative ensemble, which suggest that model-data comparison may be more successful if better data coverage of the oceans could be obtained. The Taylor diagram (Fig. 6) also indicates similarly poor results for the MH hottest month. Correlations are typically negative (albeit small), and the model fields exhibit substantially too small variability. In contrast to the situation for the LGM in Sect. 4.1.3, accounting for observational uncertainty (which is relatively small) does not improve these results significantly.

The model forcing for the MH is principally a smooth temporal and latitudinal variation in the insolation, and the model responses are similarly smooth. The data, however, are highly spatially variable (Fig. 2). As discussed above in reference to the LGM, there are a number of possible causes of this high frequency mismatch, including both representation error in the data and a lack of high frequency information in the model at the smallest scales. In order to more clearly show the regional patterns we have rebinned the data and the multi-model mean for the hottest month to 10 degree boxes. The result is shown in Fig. 8. In Europe and Africa, the anomaly appears predominantly positive and is greater at lower latitudes. In North America, the anomalies are smaller and more mixed. The data are, however, generally sparse so it is far from certain whether or not there is a significant spatial pattern in the data. What is clear is that to the extent that there is a pattern in the data across these regions, it is substantially different to that of the multi-model mean response to the forcing. Thus it appears that the model-data

mismatch is not just due to high frequency noise. The reasons behind this model-data mismatch on the regional scale require further investigation from both the model and data communities. Better global coverage of data is clearly required, but it also seems plausible that the model forcings or feedbacks are inadequate.

5 Conclusions

In this paper we extended our previous analysis of the LGM to include more data, more models, and the MH interval. We also performed the first conventional analysis of predictive skill for paleoclimate GCMs, and present Taylor diagram summaries for both intervals. In these model-data comparison exercises, we have obtained generally positive and encouraging results for the LGM, showing that the models produce generally reasonable and informative predictions of the large-scale response to strong forcing. However, limitations are apparent at finer scales. The model-data mismatch is quite large but it is possible that representation error in the data is obscuring the signal, particularly on land.

The MH, with its much smaller net climate forcing, clearly highlights the difficulties of reproducing regional-scale patterns of climate change. For this experiment the global climate change signal is very small, and the changes are regional and seasonal in nature, possibly involving significant vegetation feedbacks. We find that for the MH, the ensemble is largely unreliable, with zero skill. Furthermore, it appears possible from examination of the data that there are coherent spatial patterns in reality that are not quantitatively reproduced by the models. While more qualitative approaches have found some positive results (Brewer et al., 2007), our direct comparison of gridded data to model output highlights the substantial discrepancies. On the other hand, it should be noted that the models are responding to the applied forcing much in the way that would be expected from simple physical intuition, with changes in seasonality directly relating to the changes in radiative forcing. A likely cause of the mismatch is missing or erroneous feedbacks in the model, perhaps due to poor representations of vegetation.

However, it should also be noted that the data are not well distributed around the globe, with high density in Europe and North America, but very poor coverage elsewhere on land and in the oceans. Data coverage must be improved for us to be confident that the models are really missing some major feedbacks.

For the point of view of directly using existing models to constrain future climate, the LGM with its large forcing seems the most promising of the two experiments. However, climate science is now facing the challenge of predicting future changes on regional scales, which includes the requirement to correctly model vegetation and many other feedbacks. Our results provide some sobering evidence of the limits to the ability of current models to accurately reproduce the local patterns of change that are seen in paleoclimate data. Therefore, unlocking the reasons for the local to regional model-data mismatch for paleoclimates should be a powerful contribution to furthering progress in this area.

Acknowledgements. We are particularly grateful to Pat Bartlein for his considerable patience while helping us to use the pollen data sets for the LGM and MH, and the GHOST dataset for the MH. This work was supported by the S-10-3 project of the MoE, Japan and by the Sousei project of MEXT, Japan. We acknowledge all those involved in producing the PMIP and CMIP multi-model ensembles and data syntheses, without which this work would not exist. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

Annan, J. D. and Hargreaves, J. C.: Reliability of the CMIP3 ensemble, Geophys. Res. Lett., 37, L02703, doi:10.1029/2009GL041994, 2010. 3488

Annan, J. D. and Hargreaves, J. C.: Understanding the CMIP3 multimodel ensemble, J. Climate, 24, 4529–4538, 2011. 3489, 3495

Bartlein, P. J., Harrison, S. P., Brewer, S., Connor, S., Davis, B. A. S., Gajewski, K., Guiot, J., Harrison-Prentice, T. I., Henderson, A., Peyron, O., Prentice, I. C., Scholze, M., Seppä, H., Shuman, B., Sugita, S., Thompson, R. S., Viau, A. E., Williams, J., and Wu, H.: Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis, Clim. Dynam., 37, 775–802, 2011. 3482, 3483, 3486

Braconnot, P., Otto-Bliesner, B., Harrison, S., Joussaume, S., Peterchmitt, J.-Y., Abe-Ouchi, A., Crucifix, M., Driesschaert, E., Fichefet, Th., Hewitt, C. D., Kageyama, M., Kitoh, A., Laîné, A., Loutre, M.-F., Marti, O., Merkel, U., Ramstein, G., Valdes, P., Weber, S. L., Yu, Y., and Zhao, Y.: Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: experiments and large-scale features, Clim. Past, 3, 261–277, doi:10.5194/cp-3-261-2007, 2007a. 3485

Braconnot, P., Otto-Bliesner, B., Harrison, S., Joussaume, S., Peterchmitt, J.-Y., Abe-Ouchi, A., Crucifix, M., Driesschaert, E., Fichefet, Th., Hewitt, C. D., Kageyama, M., Kitoh, A., Loutre, M.-F., Marti, O., Merkel, U., Ramstein, G., Valdes, P., Weber, L., Yu, Y., and Zhao, Y.: Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 2: feedbacks with emphasis on the location of the ITCZ and mid- and high latitudes heat budget, Clim. Past, 3, 279–296, doi:10.5194/cp-3-279-2007, 2007b. 3484

Bretherton, C., Widmann, M., Dymnikov, V., Wallace, J., and Bladé, I.: The effective number of spatial degrees of freedom of a time-varying field, J. Climate, 12, 1990–2009, 1999. 3489

Brewer, S., Guiot, J., and Torre, F.: Mid-Holocene climate change in Europe: a data-model comparison, Clim. Past, 3, 499–512, doi:10.5194/cp-3-499-2007, 2007. 3483, 3499

Glickman, T.: 2000: Glossary of Meteorology, Amer. Meteor. Soc, 2000. 3490

Guiot, J., Boreux, J., Braconnot, P., and Torre, F.: Data-model comparison using fuzzy logic in paleoclimatology, Clim. Dynam., 15, 569–581, 1999. 3483

Hamill, T.: Interpretation of rank histograms for verifying ensemble forecasts, Mon. Weather Rev., 129, 550–560, 2001. 3489

Hargreaves, J. C.: Skill and uncertainty in climate models, Wiley Interdisciplinary Reviews: Climate Change, 1, 556–564, doi:10.1002/wcc.58, 2010. 3489, 3491

Hargreaves, J. C., Paul, A., Ohgaito, R., Abe-Ouchi, A., and Annan, J. D.: Are paleoclimate model ensembles consistent with the MARGO data synthesis?, Clim. Past, 7, 917–933, doi:10.5194/cp-7-917-2011, 2011. 3482, 3484, 3488, 3490

Jolliffe, I. and Primo, C.: Evaluating Rank Histograms Using Decompositions of the Chi-Square Test Statistic, Mon. Weather Rev., 136, 2133–2139, 2008. 3489

Joussaume, S. and Braconnot, P.: Sensitivity of paleoclimate simulation results to season definitions, J. Geophys. Res, 102, 1943–1956, 1997. 3488

Leduc, G., Schneider, R., Kim, J.-H., and Lohmann, G.: Holocene and Eemian sea surface temperature trends as revealed by alkenone and Mg/Ca paleothermometry, Quaternary Sci. Rev., 29, 989–1004, 2010. 3487

MARGO Project Members: Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum, Na. Geosci., 2, 127–132, doi:10.1038/NGEO411, 2009. 3482, 3483, 3486

Masson, D. and Knutti, R.: Climate model genealogy, Geophys. Res. Lett, 38, L08703, doi:10.1029/2011GL046864, 2011. 3486, 3494

Schmittner, A., Urban, N., Shakun, J., Mahowald, N., Clark, P., Bartlein, P., Mix, A., and Rosell-Melé, A.: Climate Sensitivity Estimated from Temperature Reconstructions of the Last Glacial Maximum, Science, 334, 1385–1388, 2011. 3486, 3496

Taylor, K.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res, 106, 7183–7192, 2001. 3484, 3492

Yokohata, T., Annan, J. D., Collins, M., Jackson, C. S., Tobis, M., Webb, M. J., and Hargreaves, J. C.: Reliability of multi-model and structurally different single-model ensembles, Clim. Dynam., 39, 599–616, doi:10.1007/s00382-011-1203-1, 2011. 3488, 3489

Table 1. Overview of the model versions used in the different ensembles analysed. The names correspond to the filenames in the PMIP2 and CMIP5 databases.

Model	PMIP2 LGM AOGCM	PMIP2 LGM AOVGCM	PMIP3 LGM	PMIP2 MH AOGCM	PMIP2 MH AOVGCM
CCSM	CCSM3			CCSM3	
CNRM	CNRM-CM33				
CSIRO				CSIRO-Mk3L-1.1*	
ECBILT*	ECBILTCLIO			ECBILTLOVECODE	ECBILTLOVECODE
ECHAM		ECHAM53-MPIOM127-LPJ			ECHAM53-MPIOM127-LPJ
FGOALS	FGOALS-1.0g			FGOALS-1.0g	
FOAM				FOAM	FOAM
GISS				GISSmodelE	
HadCM3	HadCM3M2	HadCM3M2		UBRIS-HadCM3M2a	UBRIS-HadCM3M2
IPSL	IPSL-CM4-V1-MR		IPSL-CM5A-LR	IPSL-CM4-V1-MR	
MIROC	MIROC3.2.2				
MPI			MPI-ESM-P		
MRI-nfa				MRI-CGCM2.3.4nfa	MRI-CGCM2.3.4nfa
MRI-fa				MRI-CGCM2.3.4fa	MRI-CGCM2.3.4fa

* ECBILT is the only EMIC in the ensemble. All the other models are full general circulation models.

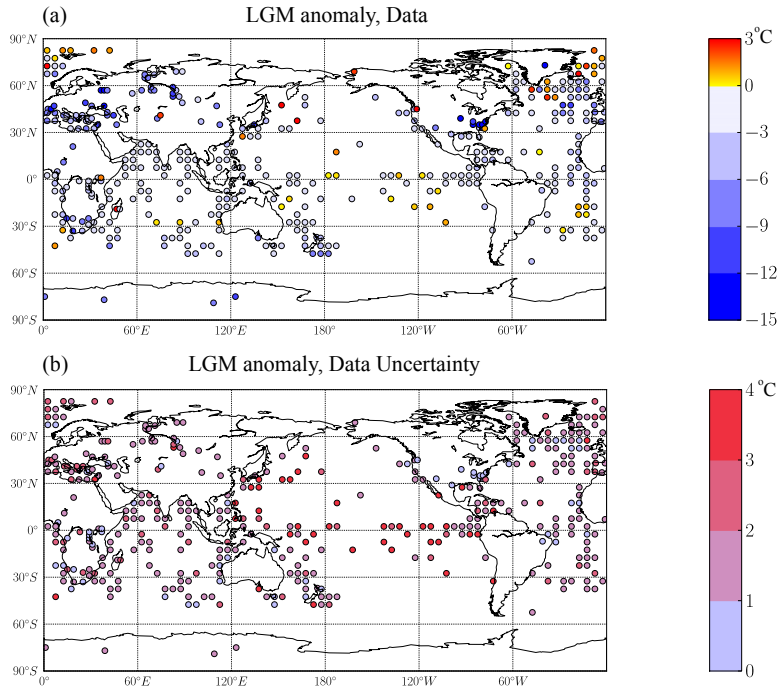

Fig. 1. (a) LGM temperature anomaly for the data.The colorbar axes are chosen to best display the data. The actual minimum and maximum are −16˚C and 6.32˚C. **(b)** The value of the uncertainty on the annual mean included in the data synthesis. Max. = 6.42, Min. = 0.24 Mean = 1.73.

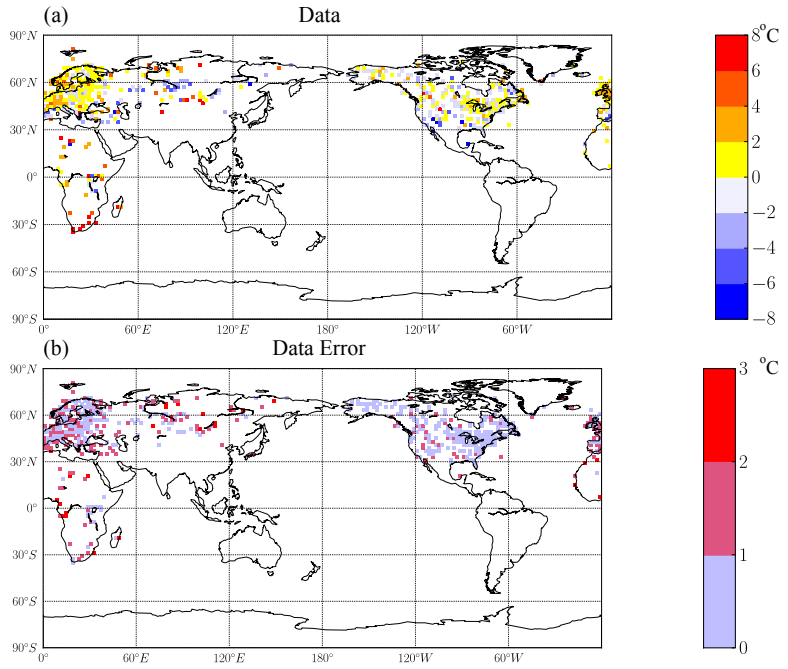

Fig. 2. Mid Holocene temperature anomaly for the hottest month: **(a)** Data, Max. = 10.0 °C, Min. = −20.1 °C, **(b)** Data error, Max. = 3.3, Min. = 0.05, Mean = 0.96.

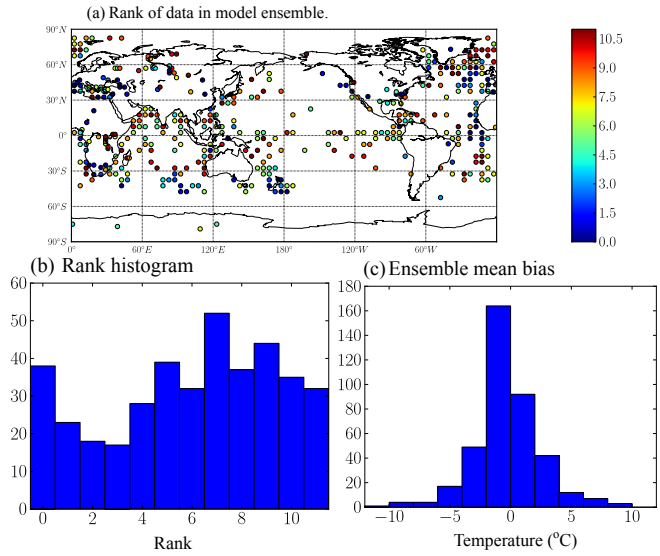

Fig. 3. (a) The rank of the data in the 11 member LGM ensemble. **(b)** Rank histogram of the ranks in plot **(a)**. **(c)** The histogram of the difference between the ensemble mean and the data for each data point in plot **(a)**. A low rank indicates that the climate change is greater in the models than the data.

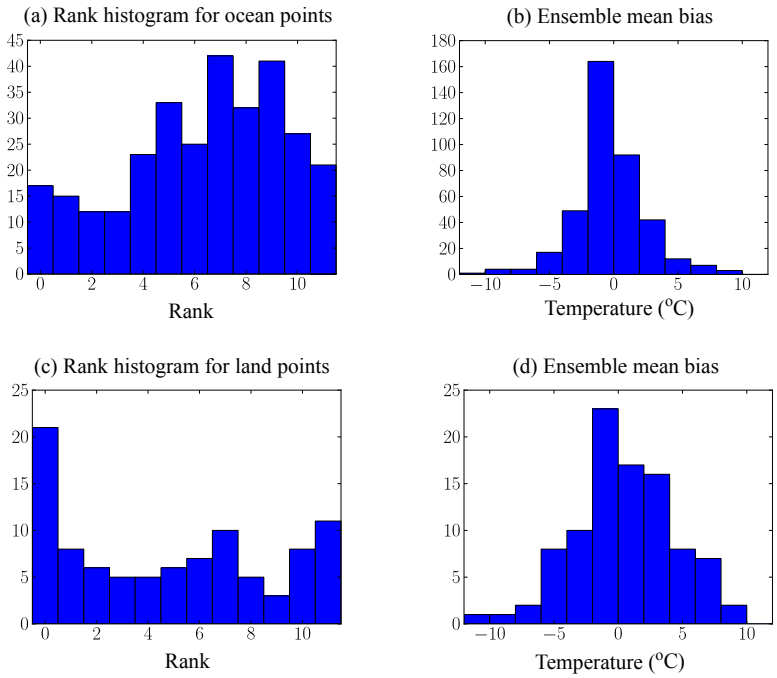

Fig. 4. Rank histograms and mean bias histograms for the LGM anomaly, considering the ocean and land separately.

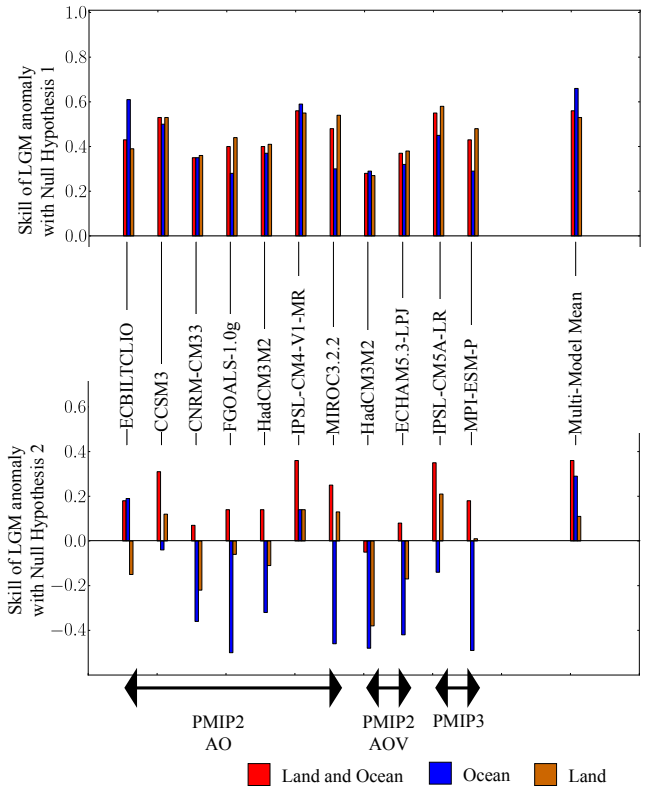

Fig. 5. Skill for the LGM anomaly. The top plot shows the result using the first reference, that the LGM anomaly is zero, and the lower plot the results using the second reference, that the LGM anomaly is equal to the data mean.

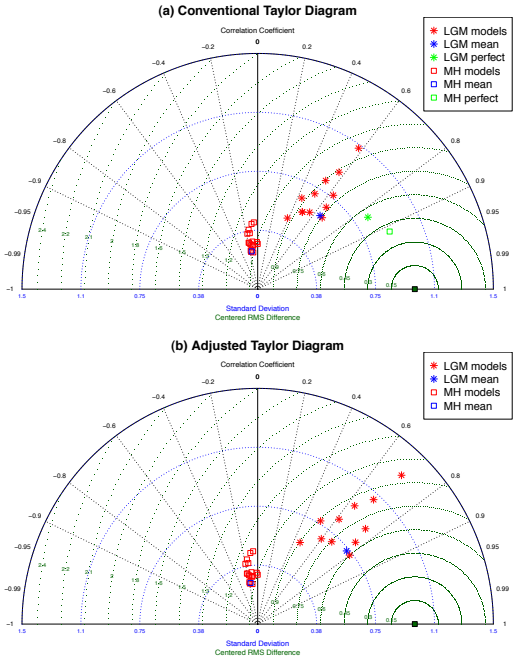

Fig. 6. Taylor diagrams for the LGM mean temperature anomaly and MH hottest month anomaly. The top plot shows conventional analysis, with the location of the "perfect model" indicated for comparison. The lower plot shows the analysis where model statistics are corrected to account for observational errors. All results are normalised by the standard deviation of the data fields.

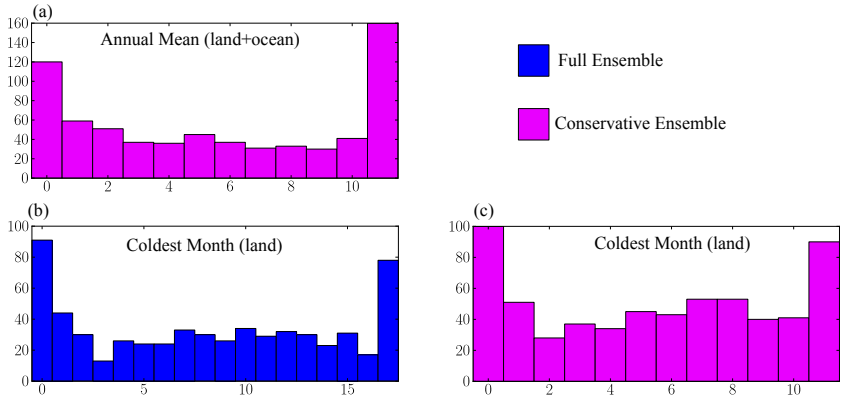

Fig. 7. Rank histograms for the three out of the eight ensembles analysed for the mid-Holocene which passed the statistical test for reliability. They nevertheless show a tendency towards being too narrow.

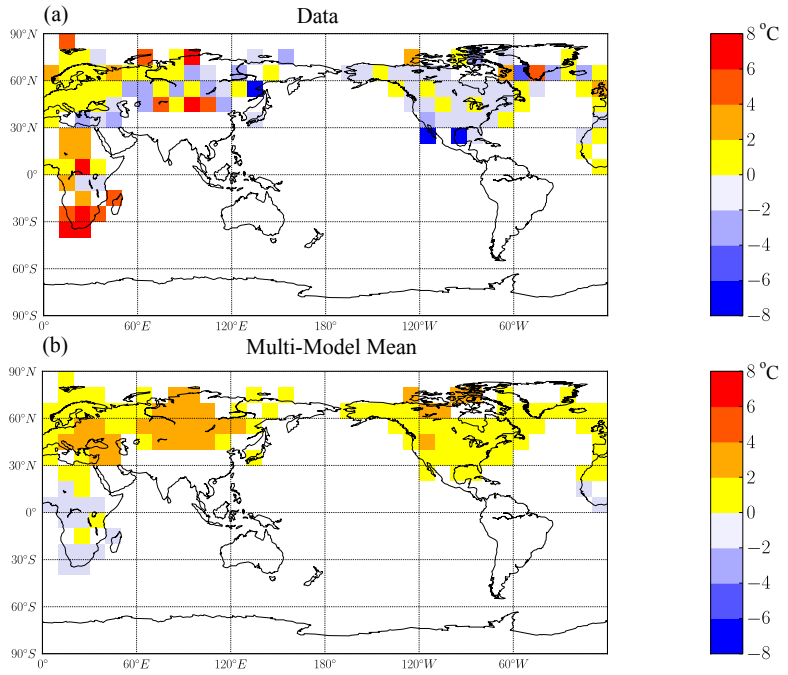

Fig. 8. The hottest month MH anomaly rebinned onto a 10 degree grid, to more clearly illustrate the model-data mismatch on the regional scale.