Clim. Past, 7, 917–933, 2011
www.clim-past.net/7/917/2011/
doi:10.5194/cp-7-917-2011

Climate of the Past

Are paleoclimate model ensembles consistent with the MARGO data synthesis?

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

[1]RIGC/JAMSTEC, Yokohama Institute for Earth Sciences, Yokohama, Japan
[2]MARUM – Center for Marine Environmental Science and Department of Geosciences, University of Bremen, Germany
[3]AORI, University of Tokyo, Japan

Received: 10 February 2011 – Published in Clim. Past Discuss.: 1 March 2011
Revised: 5 July 2011 – Accepted: 22 July 2011 – Published: 22 August 2011

Abstract. We investigate the consistency of various ensembles of climate model simulations with the Multiproxy Approach for the Reconstruction of the Glacial Ocean Surface (MARGO) sea surface temperature data synthesis. We discover that while two multi-model ensembles, created through the Paleoclimate Model Intercomparison Projects (PMIP and PMIP2), pass our simple tests of reliability, an ensemble based on parameter variation in a single model does not perform so well. We show that accounting for observational uncertainty in the MARGO database is of prime importance for correctly evaluating the ensembles. Perhaps surprisingly, the inclusion of a coupled dynamical ocean (compared to the use of a slab ocean) does not appear to cause a wider spread in the sea surface temperature anomalies, but rather causes systematic changes with more heat transported north in the Atlantic. There is weak evidence that the sea surface temperature data may be more consistent with meridional overturning in the North Atlantic being similar for the LGM and the present day. However, the small size of the PMIP2 ensemble prevents any statistically significant results from being obtained.

1 Introduction

Recent work investigating the performance of the CMIP3 ensemble (Meehl et al., 2007) of climate models, has found that it may be considered to be reasonably "reliable", at least on the global scale, when tested against modern climatology (Annan and Hargreaves, 2010). By this we mean that we do not reject the hypothesis that the truth is statistically indistinguishable from the ensemble members, at least when subjected to simple (but standard) tests based on rank histograms, explained below. However this sort of testing against modern data does not address the question of the extent to which that reliability may hold for forecasts or projections of future change. It is possible that the models may share biases through their similar parameterisations, and there could also be processes which will affect future climate changes but which are not included in the models, either because the scientific understanding about them is as yet insufficient for them to be well represented in the models, or because they are (erroneously) not considered to be of sufficient importance (Hargreaves, 2010).

We will never be able to directly evaluate the performance of long-term climate model predictions, other than by the impractical method of waiting to see what happens. Therefore, we can only update our level of confidence in the existing models by more indirect methods, such as by evaluating their behaviour under a wide range of external forcings, preferably considering time periods and data that were not used during the model development and which can therefore provide independent validation. One of the most obvious such times is the Last Glacial Maximum (LGM, 21 ka before present). While the existence of the large ice sheets during that cold period may complicate the signal, this is the most recent time in the past when carbon dioxide level were significantly different to today (around 185 ppm), and a considerable amount of data has been collected which may, in principle, be used to evaluate the models.

In this paper we primarily investigate two ensembles of models which contributed to the Paleoclimate Modelling Inter-comparison Projects, PMIP (Joussaume and Taylor

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

(2000), hereafter PMIP1) and PMIP2 (Braconnot et al., 2007). The main difference between the two ensembles is that fully coupled ocean dynamics are included in PMIP2, whereas PMIP1 models used a slab ocean with ocean heat transport calibrated to pre-industrial values. An additional set of PMIP1 runs with prescribed sea surface temperature (SST) are not included in our analysis. In recent years, there has been an emphasis on developing ensembles from a single model by varying the parameters in that model. In order to consider the extent to which it may be possible to use single model ensembles (SME) as a replacement for the multi-model ensembles we also consider results from an SME which was generated by changing parameter values in the MIROC3.2 slab ocean model (Hasumi and Emori, 2004; Annan et al., 2005b).

It is not always straightforward to compare model outputs to paleoclimate data, as the former have low spatial resolution and substantial smoothness, whereas the latter are generally derived from point sources such as cores that sample small spatial scales. Additionally the paleoclimate data may have heterogeneous uncertainties arising from the use of different proxies and the representativeness of the individual estimate for the considered time period, which in turn depends on factors such as the number of samples per core and the accuracy of the dating of each sample. The data we consider here are the Multiproxy Approach for the Reconstruction of the Glacial Ocean Surface (MARGO) sea surface temperatures (MARGO Project Members, 2009). This dataset is in a very modeller-friendly form. It is a synthesis of six different proxies and includes estimates of the uncertainty in the temperatures obtained, so may be considered to represent the combined expertise of at least a sizeable fraction of the LGM paleo-data community. As such we consider it to be a powerful dataset against which to evaluate the multi-model ensemble, which likewise may be considered to represent the combined expertise of the modelling community (Hargreaves, 2010; Annan and Hargreaves, 2010). There have been previous attempts to compare PMIP and MARGO (e.g. Kageyama et al., 2006; Otto-Bliesner et al., 2009), but here we analyse each ensemble as a whole and are thus able to make an assessment of overall performance.

2 Reliability and the rank histogram

Reliability is a key concept in probabilistic prediction. Probabilistic predictions are described as reliable if the predicted probability of an event equals the frequency of its occurrence, over a large set of instances.

A standard paradigm for the interpretation of model ensembles is to consider reality as being a random sample from the same distribution as the models (Annan and Hargreaves, 2010). In this situation, a probabilistic prediction made from the ensemble by counting the relative frequencies (i.e. the proportion of members for which an event does/does not occur, or, if a gaussian approximation is suitable, through its mean and standard deviation (e.g. Figs. SPM.5 and SMP.7 of the IPCC AR4 Summary for Policymakers, Solomon et al., 2007)) will be reliable. If, instead, the ensemble spread is too large, such that observations are relatively closer to the mean than the ensemble members, then this indicates that a tighter prediction should be possible. On the other hand, a very narrow ensemble suggests that we may have a bias such that the ensemble rarely includes the truth. In both of these cases, a direct probabilistic interpretation of the ensemble would be misleading, but the second example is probably the more worrisome of the two, as it provides no bounds on the future outcome.

One standard test of ensemble reliability is to evaluate the rank histogram, also known as Talagrand diagram (Talagrand et al., 1997). If we take a single scalar observation and combine it with the ensemble of n equivalent observations, and rank these $n+1$ values in order from smallest to largest then, for a reliable ensemble, the observation should be equally likely to take each position in the rank ordering. The rank histogram is simply the histogram of ranks so obtained for a set of observations, and so will be flat (to within sampling error) for a reliable ensemble. It is worth noting that, even for a reliable ensemble, the truth would be expected to fall outside the ensemble range for a fraction of $2/(n+1)$ of the observations, where n is the number of models. In order to quantitatively evaluate the rank histograms, we use the method presented by Jolliffe and Primo (2008). This is based on chi-square tests on the contents of the bins, and allows us to efficiently check whether the ensemble is biased, or over- or under-dispersive. Computing the rank histogram, and checking for uniformity provides a necessary condition for an ensemble prediction system to be reliable, but it should be noted that it is not by itself a sufficient one (Hamill, 2001). For example, it is possible for a uniform histogram to arise from a set of predictions each of which has specific biases which cancel out overall, or alternatively the spatial covariances for the models may be inconsistent with the observations. The analysis here only considers the aggregated analysis of pointwise values. Moreover, there is no guarantee that the performance against a historical data set will be matched in the future. Nevertheless, it is reasonable to prefer an ensemble which does have a track record of good performance over one which does not.

One point that must not be overlooked, which may be expected to be more important for paleoclimate studies than those looking at modern climate, is the issue of uncertainty in the observational data. If the truth is sampled from the same distribution as the ensemble members, then the inevitable presence of this observational error will result in the observations themselves tending to have a somewhat broader distribution than the ensemble members. A standard method to account for this is to simply add equivalent (randomly generated) perturbations onto the model outputs (Anderson, 1996). Of course this requires some estimate of the magnitude of the

observational uncertainties. Fortunately, some estimates of uncertainty were provided for the MARGO synthesis, which we discuss further in Sect. 3.3.

3 Models and data

3.1 The PMIP ensembles

For the PMIP1 experiments (Joussaume and Taylor (2000), and other papers in the same volume), the focus was on atmospheric general circulation models (AGCMs), run with prescribed forcing to simulate the conditions of the mid-Holocene (6 ka BP) and the LGM (21 ka BP), and the pre-industrial control climate. For some of the models, the LGM and control runs were performed with the atmospheric climate model coupled to a slab ocean. In addition, one model (CLIMBER) was an EMIC (Claussen et al., 2002; Weber, 2010) with reduced complexity but including a fully coupled atmosphere-ocean system. It is this subset of models, which permit the SST to evolve, that we analyse here. The result is a 10 member ensemble including models of varying resolution and complexity (see Table 1). See also the PMIP1 website, http://pmip.lsce.ipsl.fr/, for further information about the PMIP1 database.

For a slab ocean AGCM, the model is first run with a prescribed modern SST field for the pre-industrial climate, and the heat fluxes (the Q-flux) required to maintain this SST, in addition to the heat flux due to the processes in the model, are calculated. The model is then run again, imposing the Q-flux but allowing the slab ocean to adjust the temperature. For the modern climate there should, therefore, be very little drift in SST away from the data that were used to calculate the fluxes. When these models are integrated for past or future climates this modern Q-flux field is applied, with the SST allowed to change. Running models of this type is far less computationally expensive than running a model with a fully coupled ocean, due primarily to the shorter spin-up time. The physical interpretation of this simplification is that the horizontal heat flux in the ocean is assumed to remain fixed, but the vertical flux between the atmosphere and ocean can vary. This has been described as allowing thermodynamic but not dynamic ocean processes to act (Ohgaito and Abe-Ouchi, 2007).

Unfortunately the SST outputs are not in the PMIP1 database, so here we used the air temperature variable (called "TAS", 2 m surface air temperature). As will be discussed in more detail in Sect. 4.1.1, this presents some problems for our analysis. While, for most of the ocean, the change in temperature between pre-industrial and LGM is similar for both the SST and air temperature, the air temperature over sea ice is generally very much colder than the SST beneath the ice. Thus, for high latitudes where the sea ice is present at the LGM, the PMIP1 results cannot be directly compared with the MARGO SST data, and so we analyse PMIP1 for the low latitude region only (35° S–35° N). Since annual average output was not available for one model (LMD 4) we used the monthly mean output. Details on the number of days in the months of the PMIP1 models is not available, so we used a simple average of the monthly means to make an annual mean. The potential error incurred in doing this is small in the context of the ensemble results presented here.

By the time of the second PMIP experiment, PMIP2 (Braconnot et al., 2007), new versions of the GCMs had been developed, with generally higher resolution. Another major difference was the coupling of fully dynamic ocean models to the AGCMs (to make AOGCMs). A small number of models also included coupled vegetation components (AOVGCMs). In addition, for the LGM experiment, the forcing protocol was slightly refined, but we do not expect this to have a major effect on the results. The PMIP2 database (see http://pmip2.lsce.ipsl.fr/) includes SST model output (called "tos"), thus enabling direct comparison with the MARGO data for 9 ensemble members. For these data (and the PMIP2 air temperature) the annual means were created from the monthly means. For PMIP2, the month length information was available in the netcdf files, so we could make annual means based on the actual number of days in the month. There is some inconsistency in variables in the database, particularly the ocean variables. For meridional overturning and northward heat transport, around half the models have annual averages available, one has only daily output, one has only some of the variables, and the rest have monthly means (see Table 1 for details). Two AOVGCMs, which are AOGCMs with a coupled vegetation model, are included in the ensemble. For one of these we also have the equivalent AOGCM. For two such closely related models we would expect some similarities between the two, but since the coupling of a whole new sub-model is a larger change than just a change in resolution, we do expect them to differ significantly, and so include both in our ensemble. For ECHAM there is also an AOGCM and an AOVGCM in the database. We use only the AOVGCM model, since SST, the principle variable for comparison with MARGO, was not available for the AOGCM.

3.2 JUMP ensemble

The single-model ensemble (SME) analysed here is the ensemble of MIROC3.2.2 that has been included in several previous analyses (Hargreaves et al., 2007; Hargreaves and Annan, 2009; Yokohata et al., 2010; Yoshimori et al., 2011). Created by the Japan Uncertainty Modelling Project, it is hereafter called the JUMP ensemble. This 40 member ensemble was derived by varying 13 parameters in a slab-ocean version of the MIROC3.2.2 (also called MIROC4) GCM, using the Ensemble Kalman Filter to tune the parameters to modern seasonal mean climatological data (20–30 yr climatological means from a variety of sources representing late 20th century climate) using the same methods described in

Table 1. Variables available for PMIP1 and PMIP2. 2-D model output: "tas", 2 m surface air temperature, "tos", SST. Zonally averaged regional output: "stfmmc", overturning steam function; "hfogo", northward heat transport. There are 4 regions (global, pacific, indian and atlantic) for the PMIP2 models, apart from CCSM, which has only 2 (global-marginal seas and atlantic + mediterranean + labrador + GIN + Arctic). Temporal resolution: "cm", monthly mean output from PMIP1;"'mo", daily data for 100 yr; "'se", monthly mean output for 12 months averaged over 100 yr; "an", 100 (or 99 for HADCM3 AOVGCM) years of annual average output.

Model	Model type	tas	tos	stfmmc	hfogo
PMIP1 models (Joussaume and Taylor, 2000; Petoukhov et al., 2000)					
CCC 2	AGCM	cm			
CCM 1	AGCM	cm			
CLIMBER 2	EMIC	cm			
GENESIS 1	AGCM	cm			
GENESIS 2	AGCM	cm			
GFDL	AGCM	cm			
LMD 4	AGCM	cm			
MRI 2	AGCM	cm			
UGAMP	AGCM	cm			
UKMO	AGCM	cm			
PMIP2 models (Braconnot et al., 2007; Randall et al., 2007)					
MIROC3.2.2(medres)[4]	AOGCM	se/mo	se	se	se
CCSM3	AOGCM	se	se	se	se[1]
CNRM-CM3.3	AOGCM	se	se	se	se
ECHAM5.3/MPIOM127/LPJ	AOVGCM	se	se	se	–
ECBILTCLIO	EMIC	se	mo	an[3]	an[3]
FGOALS_g1.0	AOGCM	se	se	se[2]	se
HADCM3M2	AOGCM	se	se	an	an
HADCM3M2	AOVGCM	se	se	an	an
IPSL-CM4_v1	AOGCM	se	se	an[5]	an[5]

Notes: [1] For CCSM, "hfogo" files were in error and new files (100 yr of monthly output) were obtained directly from B. L. Otto-Bliesner. [2] For FGOALS, the stfmmc files appear to be the negative of what was expected, but otherwise reasonable. Contact with the developers could not be achieved, and so this was not confirmed, but is assumed to be the case. [3] For ECBILT stfmmc and hfogo, the region labelled global appears to correspond to the atlantic, and vice versa. See the PMIP websites for more details of the models. [4] MIROC3.2.2, the official CMIP3 version of MIROC was used, rather than MIROC3.2. Not all variables for MIROC3.2.2 are available on the PMIP2 database, but they are available to the authors. [5] These variables for IPSL were made available after the on-line review of the original manuscript.

Annan et al. (2005a) and Annan et al. (2005b). As described in Annan et al. (2005a), a simple approximation to account for structural model error is made during the tuning. This error is approximated by the difference between the control model run with the default parameter set and the climatological data, which is then treated in the same way as data error. The result is intended to be an ensemble that is broadly consistent with climatological data. The model has the same atmosphere as the MIROC3.2.2 AOGCM submitted to the PMIP2 database, except, for reasons of computational cost, we ran the ensemble at the lower resolution of T21 (rather than T42). Although a couple of ensemble members are clearly too cold at the LGM (see Fig. 2c), in order to retain the maximum ensemble spread, in this analysis we retain the entire 40 member ensemble, run for both the LGM and pre-industrial simulations described in Hargreaves et al. (2007), using the PMIP2 forcing protocol for the LGM simulations.

3.3 MARGO synthesis

The LGM cold period has long been recognised as a target for evaluating the response of the climate system to large perturbations (cf. Randall et al., 2007, p.447). Consequently, a relatively large amount of paleo-SST data is available. The most recent synthesis of LGM SST data is presented by the MARGO Project Members (2009) as the result of a large international community effort (see also Kucera et al., 2005).

The MARGO synthesis is based on 696 individual SST reconstructions and combines the results of six proxies. Four of these are microfossil proxies based on the species composition of planktonic foraminifera, diatoms, dinoflagellates and radiolaria. The other two are geochemical proxies based on alkenones with 37 C atoms produced by unicellular algae (coccolithophores) in slightly different composition in relation to changes in temperature, as well as the ratio of magnesium to calcium found in planktic foraminiferal shells.

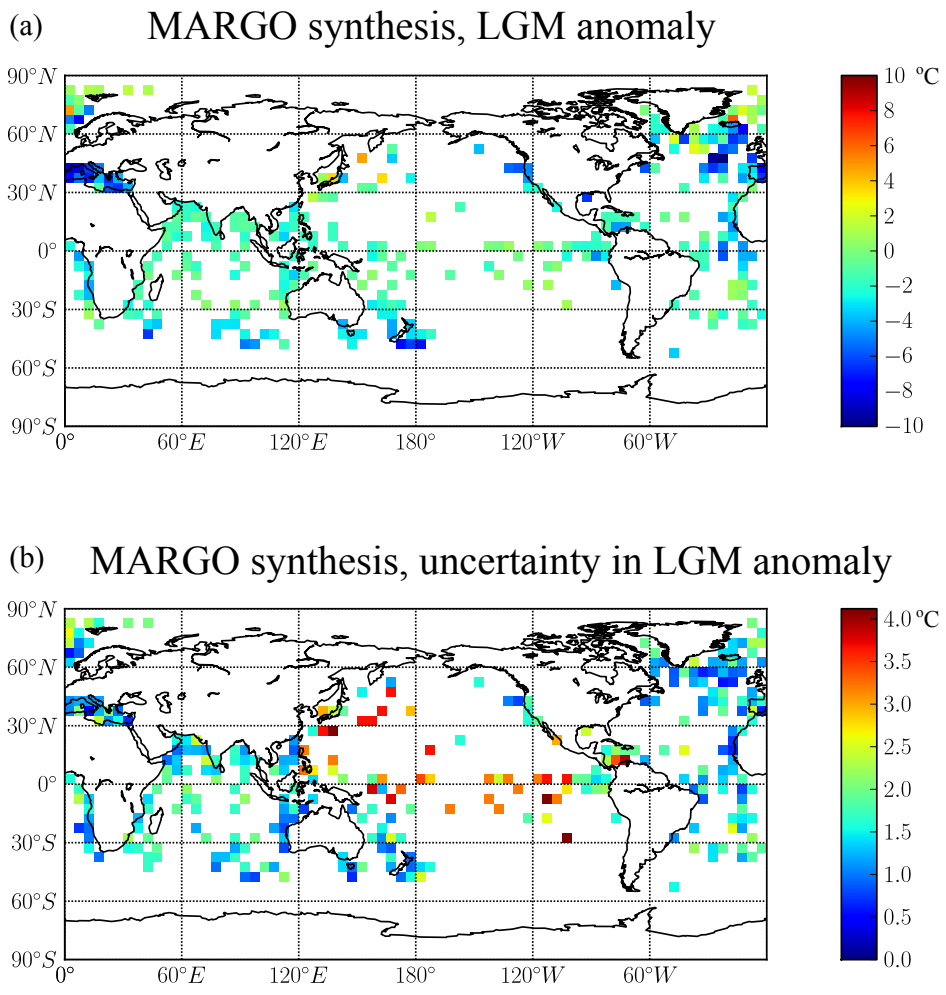

Fig. 1. (a) MARGO LGM SST anomaly with respect to WOA data. For ease of comprehension, zero is place at the centre of the colour bar, giving a range of −10 °C to +10 °C. The minimum value is actually −11.8 °C and the maximum 6.32 °C. **(b)** The value of the uncertainty on the annual mean included in the MARGO synthesis dataset.

Figure 1a shows the reconstructed LGM SST anomaly with respect to the present-day 10 m ocean temperature taken from the World Ocean Atlas 1998 (Conkright et al., 1998). Hereafter we use the term "LGM anomaly" to refer to the value of an annually-averaged variable at the LGM climate minus that at the control/present-day climate. It should be noted that the MARGO definition of SST was 10 m depth (Kucera et al., 2005; Kageyama et al., 2006) whereas the modelled "tos", although not explicitly defined, is probably calibrated to a shallower value of around 2m. Based on the MIROC3.2.2 model for which both depths were available, the LGM-CTL anomalies at the two depths generally agree to within 0.1 C, although some very localised differences in coastal areas can exceed ± 1 C. With no firm basis for correction, we ignore this detail in the analysis, and do not expect that this can have significantly affected the results.

The MARGO project members also present the first attempt at a quantitative treatment of uncertainty and the propagation of errors in a multi-proxy reconstruction of climate. It is based on a combination of expert judgment and some basic statistics, including the different sources of uncertainty and their propagation. Thus it takes into account (1) the error of calibration for each proxy, (2) its uncertainty due to the assumption of stationarity through time and in space, (3) the number of samples upon which each individual LGM SST reconstruction is based and (4) the quality of the age model for each ocean sediment core (MARGO Project Members, 2009). These uncertainties are propagated during the calculation of "block averages" and combined with the degree of convergence among the SST estimates within each block. The resulting block-averaged uncertainties clearly demonstrate more confidence in the reconstructed SST anomalies in some places than others (Fig. 1b).

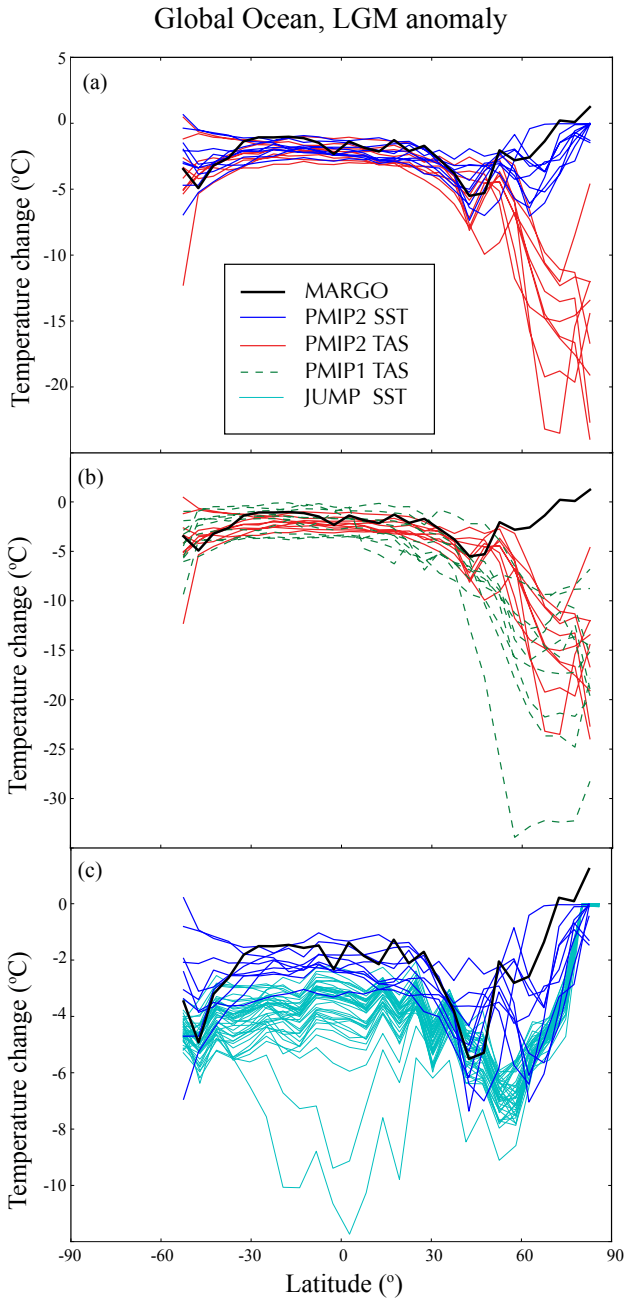

Fig. 2. The LGM anomaly for PMIP2 SST, PMIP2 TAS and PMIP1 TAS, and the MARGO data. As explained in the text, the PMIP model output is interpolated onto the MARGO grid. The zonal means are made by averaging only over the grid boxes for which there are MARGO data.

We note that the MARGO error estimate is only defined within a constant factor, because it is proportional to the so-called "mean reliability index", which is deliberately scaled such that its minimum value is one (MARGO Project Members, 2009). In order to incorporate this qualitative statistic in our analysis, we assume that the errors are Gaussian with a standard deviation at datapoint i, σ_i, given by $\sigma_i = A \times \mathtt{Err}_i$, where $\mathtt{Err}_i$ is the MARGO error estimate, and here we select the value $A = 1$. This assumption is considered reasonable in the expert opinion of the MARGO project members who worked on the derivation of the uncertainty estimate. The assumption of Gaussian independent errors is a very simple first-order approximation which could be further refined, although the selection of the overall scaling A is a dominant factor in the analysis.

It turns out that large discrepancies with respect to reconstructed LGM SST anomalies recorded by different proxies remain. Paradoxically, LGM conditions in the most densely sampled northern North Atlantic Ocean remain associated with large scatter and uncertainties (MARGO Project Members, 2009, p. 127). Possibly, the uncertainty assigned to each MARGO SST value does not fully capture the ambiguity of its attribution to a certain season and depth.

For both the PMIP1 and PMIP2 ensembles the 2-D model output was all interpolated onto the MARGO 5°×5° grid. The whole MARGO dataset has data in 307 grid boxes. For some of the interpolated grids of the PMIP2 models, MARGO data points coincided with land. These points were therefore excluded from the analysis of PMIP2, leaving 293 grid boxes in all. For PMIP1, the analysed region (35° S to 35° N) includes 190 data points. Since the T21 MIROC grid has a similar grid size to that of MARGO, but is displaced, the MARGO synthesis was re-derived from the original proxy data points onto the MIROC grid, for better comparison with that ensemble.

4 Results

4.1 Reliability of the paleoclimate ensembles

4.1.1 PMIP1

As briefly discussed in Sect. 3.1, sea surface temperature output is unavailable for PMIP1, so we must use the surface air temperature in our analysis instead. For the present day, the difference between surface air and sea temperatures are of the order of a degree, and over open ocean we would expect the air and surface ocean temperature changes to be about the same due to their tight coupling (Jones et al., 1999). The LGM was, however, a much colder climate than the pre-industrial climate, resulting in a considerable southward extension of sea-ice. The sea-ice acts as an insulating layer so that, while the water beneath the ice is at or near the freezing point of water, the air above the ice can get much

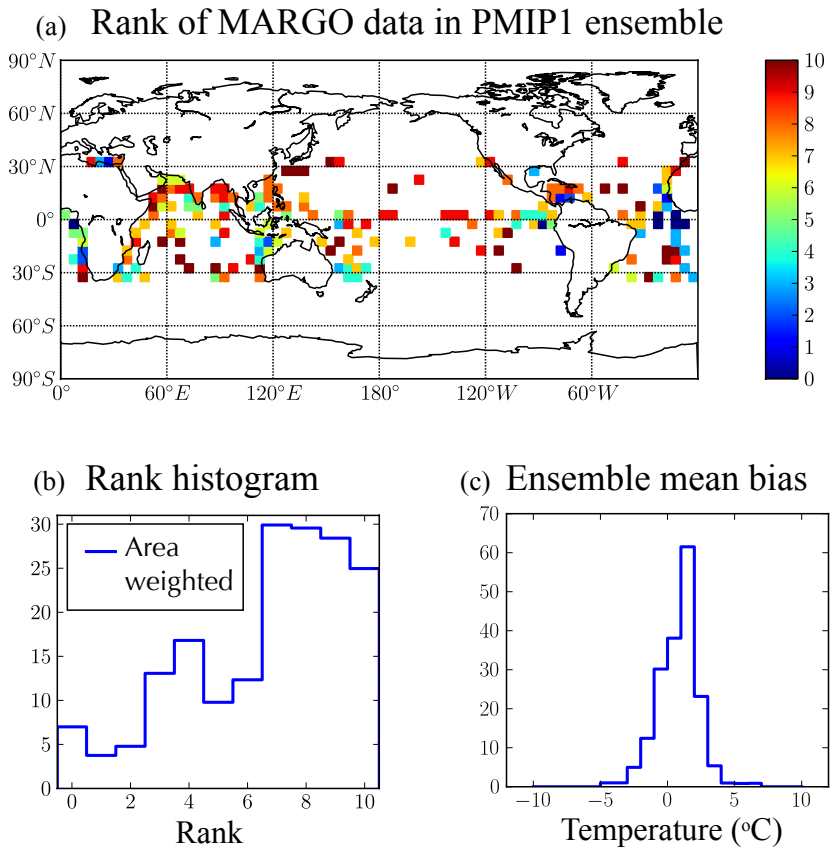

Fig. 3. **(a)** The rank of the MARGO data in the PMIP1 TAS for the range 35° S to 35° N. **(b)** Area-weighted rank histogram of the ranks in plot **(a)**. High rank (red on the colour scale) indicates that MARGO is warmer at the LGM than the ensemble. **(c)** The histogram of the difference between the PMIP1 TAS ensemble mean and the MARGO data for each data point in plot **(a)**. The uncertainties in the MARGO data are not taken into account

Table 2. Statistics for the χ-square tests of uniformity, with and without including the MARGO data uncertainty estimates. p-values for statistical significance of non-uniformity for the rank histograms, following Jolliffe and Primo (2008). We specifically test for bias, V-shape (including the inverted V) and whether one or both end bins are significantly different from that expected for a uniform distribution. Total refers to the chi-square test on the full histogram.The bold font indicates those statistics which indicate a distribution significantly different from uniform at the 5 % level.

	Shape being tested with χ-square test of non-uniformity					
Ensemble	Bias	Vshape	Both ends	Left end	Right end	total
PMIP1	0.1	0.9	1.	0.6	0.6	1.
PMIP1 including MARGO errors	0.5	0.3	0.5	0.5	0.7	1.
PMIP2	0.2	0.1	**0.05**	0.7	**0.03**	0.8
PMIP2 including MARGO errors	0.6	0.8	0.8	0.8	1.	1.
JUMP	<**0.0001**	<**0.0001**	<**0.0001**	0.56	<**0.0001**	<**0.0001**
JUMP including MARGO errors	**0.0001**	**0.02**	0.007	0.9	<**0.0001**	1

colder. Therefore, in the high latitudes, we expect the surface ocean and air temperatures to diverge significantly. Figure 2 illustrates the effect, by showing the zonally averaged LGM anomaly for the PMIP2 ensemble of surface air and ocean temperatures. We conclude that, in order to compare the PMIP1 TAS with the MARGO SSTs, we should restrict the region of analysis to the lower latitudes, 35° S to 35° N.

The results of the basic reliability analysis are shown in Fig. 3. Figure 3a shows the rank of the observations in the 10 member PMIP1 ensemble for each MARGO grid point

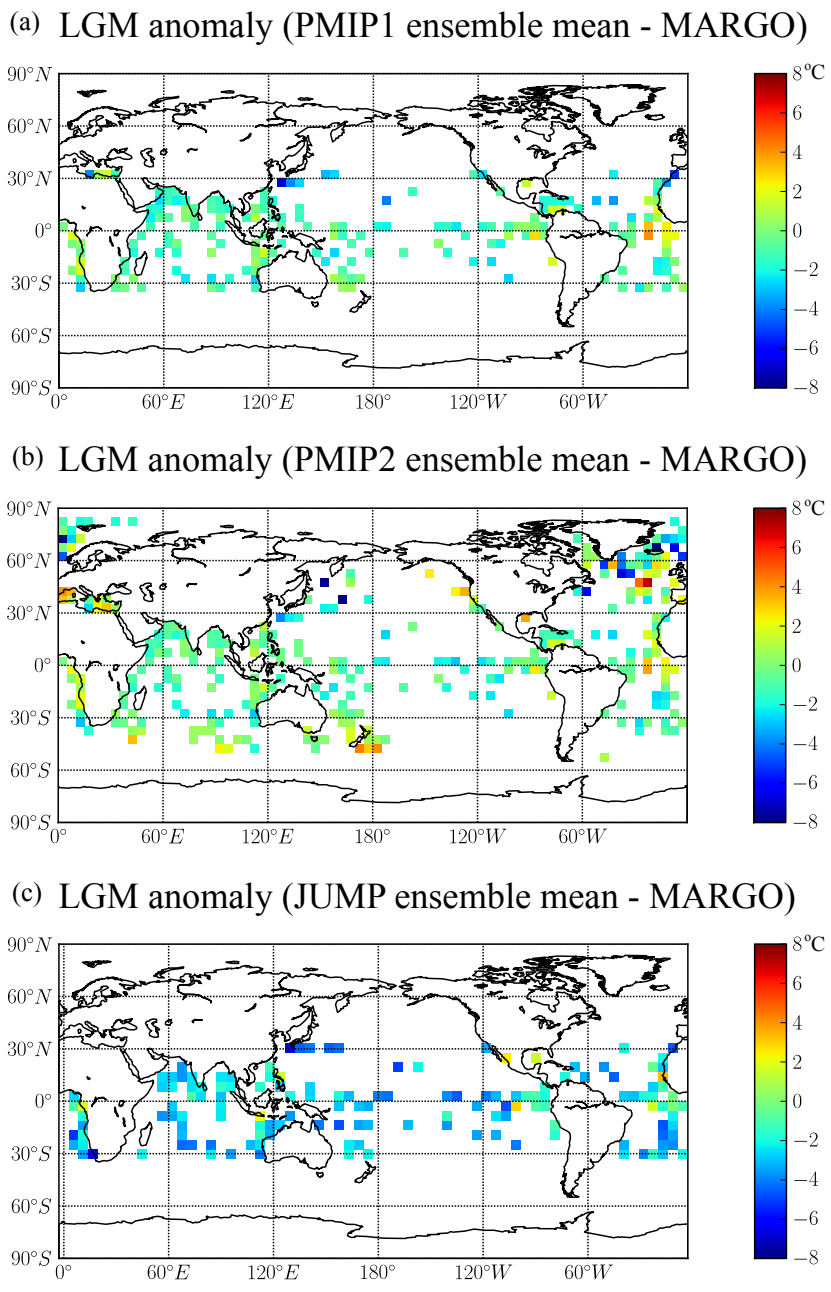

Fig. 4. The differences between the LGM anomalies for the model ensemble means and MARGO. **(a)** PMIP1, **(b)** PMIP2, **(c)** JUMP. The influence of the MARGO uncertainties is not included in these plots.

in the restricted latitude range. Throughout this work, a high rank indicates that the MARGO LGM state is relatively warm compared to the ensemble. While the plot appears red rather than blue overall, the eastern side of the Atlantic is blue, indicating that the models have generally more cooling in this region than the data. This area off the west coast of Africa, which is an upwelling region for the modern climate, is a region where models are known to perform poorly, being generally too warm (Randall et al., 2007, Figure 8.2), due partly to insufficient resolution (for CCSM3 model, Large and Danabasoglu, 2006; for HiGem model, P. L. Vidale, personal communication, 2011). Thus it is perhaps no surprise that the PMIP1 models are doing poorly in this region, and the result also indicates that accounting for biases in the base state may be important when using anomalies to estimate climate changes. Figure 3b shows the overall rank histogram. Assuming an effective dimension of 5, the histogram is statistically consistent with a flat distribution (see Table 2). Some

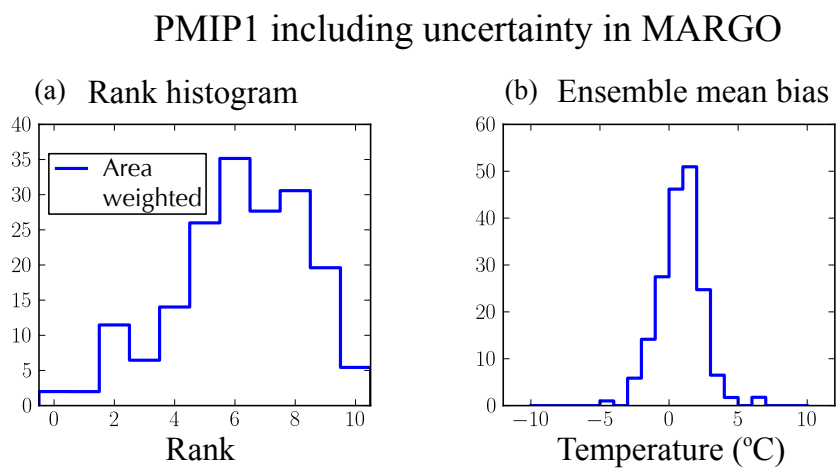

Fig. 5. As in Fig. 3b and c, but the uncertainty in the MARGO data are taken into account in the analysis of the PMIP1 ensemble. **(a)** Area-weighted rank histogram. High rank indicates that MARGO is warmer at the LGM than the ensemble. **(b)** Histogram of the difference between the PMIP1 ensemble mean and the MARGO data.

work has been done attempting to calculate the effective dimension of the CMIP3 ensemble (Annan and Hargreaves, 2011), which suggests that a lower value than 5 may be appropriate for a limited region such as the tropical ocean. We choose, however, to err on the side of caution, as assuming a higher value increases, rather than decreases the stringency of the statistical test. Figure 3c shows the histogram of the ensemble mean difference between the LGM anomaly for the MARGO synthesis and PMIP1 for each MARGO grid point, indicating that the ensemble mean error is only about 1 °C for much of the ensemble. Figure 4a shows the same result as a spatial plot.

The analysis was repeated, inflating the PMIP1 ensemble to account for the estimated data error, as described in Sect. 3.3. The results can be seen in Fig. 5a and b. The rank histogram appears dome shaped, compared to the histogram in Fig. 3, which is the indication of an ensemble that is too wide. The "Vshape" statistic (i.e. how similar the rank histogram is to a V shape or its inverse, Jolliffe and Primo, 2008) measures the significance of this shape and, as shown in Table 2 the ensemble remains consistent with a uniform distribution. As shown by Fig. 5b, the model-data differences are also slightly inflated. Overall the PMIP1 ensemble results are encouraging, although the fact that we have analysed only a subset of the latitudes for a single variable means we cannot make a very strong statement about the ensemble reliability.

4.1.2 PMIP2

Figure 6 illustrates the results for the PMIP2 ensemble compared to the global MARGO synthesis. Previously, (Figs. 3b, 5a, Annan and Hargreaves, 2010) we have shown only the area-weighted histograms, but because the distribution of the MARGO data over the globe is so far from uniform and many of the data points really do represent observed points rather than area averages, we also show the unweighted histogram. By far the most complete data coverage for the MARGO synthesis is in the North Atlantic region. In this region the pattern of the MARGO data in Fig. 1a is a band of small LGM anomalies or even warming closer to the Greenland coast, and a band of large cooling further away. While some of the PMIP2 models do show similar patterns, the amplitude of the pattern is smaller, with close to zero cooling near to Greenland and moderate cooling further away. Due to this difference in the amplitude of the pattern in the North Atlantic, this region contributes to both ends of the rank histogram. Since this region is a relatively high latitude area and the grid is regular in degrees, area weighting the ensemble tends to reduce the influence of these points and makes the rank histogram more uniform. Even so, the area-weighted rank histogram fails two of the statistical tests in Table 2. Of course, even after area weighting, the high density of points in the Atlantic means that the evaluation of the ensemble is weighted towards the performance in that region. It could be argued that giving prominence to the North Atlantic may not be unreasonable as it is seen as a key indicator of the general state of the ocean circulation (Randall et al., 2007). If we analyse just the limited region of tropical latitudes analysed for PMIP1, 35° S to 35° N, then the PMIP2 results also appear reliable, so, from this analysis alone we do not have evidence that either PMIP1 or PMIP2 produced superior results. As shown in Fig. 4b, the bias in the LGM anomaly in the region off the west coast of Africa may be very slightly improved in PMIP2, although, as the plot of the rank shows, this area is still poorly represented in PMIP2.

Of the 307 data points in the MARGO dataset, there are 48 points scattered around the globe for which the LGM

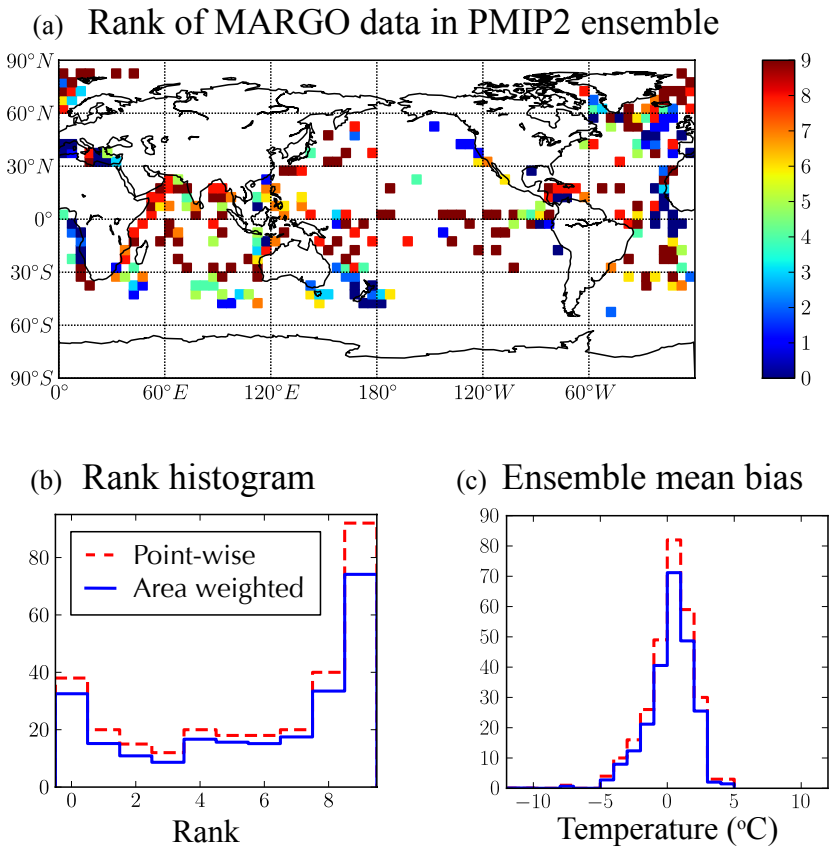

Fig. 6. (a) The rank of the MARGO data in the PMIP2 SST for the whole globe. **(b)** Area-weighted rank histogram of the ranks in plot **(a)**. High rank (red on the colour scale) indicates that MARGO is warmer at the LGM than the ensemble. **(c)** The histogram of the difference between the PMIP2 SST ensemble mean and the MARGO data for each data point in plot **(a)**. Both area-weighted and point-wise rank histograms are shown. The uncertainties in the MARGO data are not taken into account

anomaly is positive, indicating warming. Typically the magnitude of the warming at those points is less than a degree in the lower latitudes, but there are a few points in the northern high latitudes in which the warming is greater, up to a maximum of 6.3 °C. North of the UK, most points in the North Atlantic indicate warming. In contrast, points that warm are very rare in the PMIP2 ensemble; considering only the grid boxes populated with MARGO data, 3 models have no warming points, and there are only 11 warming points among the other 6 models, with only 4 of those points in the North Atlantic. This result is of some concern. To many modellers, it is considered counterintuitive to have warming in regions of the globe at the LGM, which causes them to question the quality of the data. It should be noted, however, that, for the MARGO data, there are only 8 points, all in the high latitudes, out of the 48 warming points, for which the amount of warming exceeds the estimated uncertainty of the data. Thus these data provide relatively low confidence that warming did in fact occur.

We repeated the reliability analysis, including the MARGO error estimates as described previously. On the spatial plot of the rank (Fig. 7a), the area of low rank off the west coast of Africa that was apparent in the PMIP1 results remains, and two further blue patches in the high latitudes are apparent, around New Zealand and the North Atlantic, as discussed above. On the whole, however, the rank seems quite variable, suggestive of reliability on scales less than global. Applying the statistics, we find that the reliability of the ensemble is increased to the point at which it passes all the statistical tests at the 5 % level (Table 2). This result indicates the great importance of consideration of the uncertainty in the data when comparing models and data, particularly for paleoclimates. Quantitative uncertainty estimates are far from ubiquitous in paleoclimate data, and the methods for deriving the estimates are not always well established, and open to development. For example, it may be desirable to indicate likely correlations between closely located points or points based on the same proxy types. Further work in this area is undoubtedly warranted.

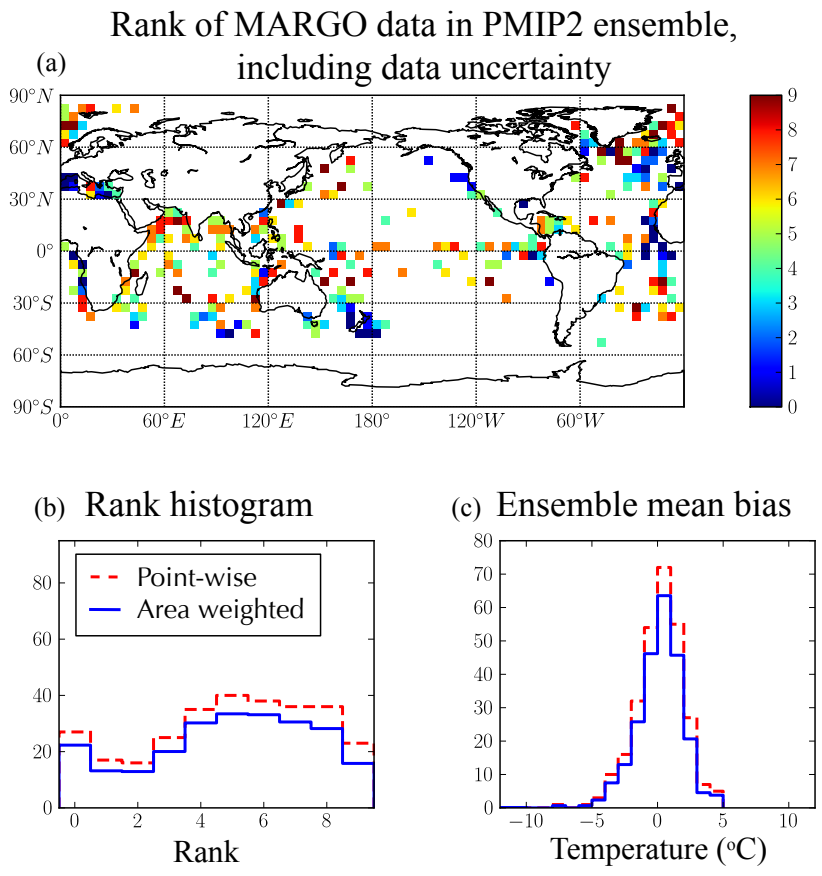

Fig. 7. (a) The rank of the MARGO data in the PMIP2 SST for the whole globe. **(b)** Area-weighted rank histogram of the ranks in plot **(a)**. High rank (red on the colour scale) indicates that MARGO is warmer at the LGM than the ensemble. **(c)** The histogram of the difference between the PMIP2 SST ensemble mean and the MARGO data for each data point in plot **(a)**. Both area-weighted and point-wise rank histograms are shown. The uncertainties in the MARGO data is taken into account

4.1.3 JUMP ensemble

Figure 8a–d and Table 2 show the overall results for the JUMP ensemble. Figure 8a and b show the results without including the MARGO error. The results are in stark contrast to those for both PMIP1 and PMIP2; the JUMP ensemble is clearly very biased and therefore too narrow and unreliable. In this case, only a slight improvement in the statistics occurs when the uncertainty in the MARGO data is included (Fig. 8c and d). As a sensitivity analysis, we tested a lower value for the assumed effective dimension, but the ensemble remains unreliable unless the assumed effective dimension is reduced to a value as low as 2, which seems implausibly low. The ensemble was created by varying 13 parameters found in previous experiments to affect the climate sensitivity and global LGM temperature anomaly. While not designed with the specific purpose of producing regional variability in LGM ocean temperatures, it is still of some concern that the range of the ensemble compares so poorly with the data. When considering the zonal means, the LGM anomaly of the JUMP ensemble is rather cold, although it does overlap with the PMIP2 ensemble. This is consistent with previous work with the JUMP ensemble, which found that the whole ensemble has climate sensitivity greater than 4 °C (Hargreaves and Annan, 2009), and so is at the high end of the range thought likely, whereas most GCMs have climate sensitivity spread throughout the canonical range. Figure 4c shows that the ensemble mean is generally biased cold compared to PMIP1 or PMIP2. It appears, therefore, that such high values for climate sensitivity may be harder to reconcile with the MARGO data than the more moderate climate sensitivities of the PMIP2 ensemble, although this could also just be an artefact of this specific model.

As can be seen in Fig. 2c, there is considerable variation between ensemble members in the latitudinal variation of the zonal mean LGM anomaly in the PMIP2 ensemble. For the JUMP ensemble (not shown), while the width of the ensemble of zonal means is comparable to that of PMIP2, this latitudinal variation pattern looks rather similar for all the ensemble members. The effect of varying the parameters has

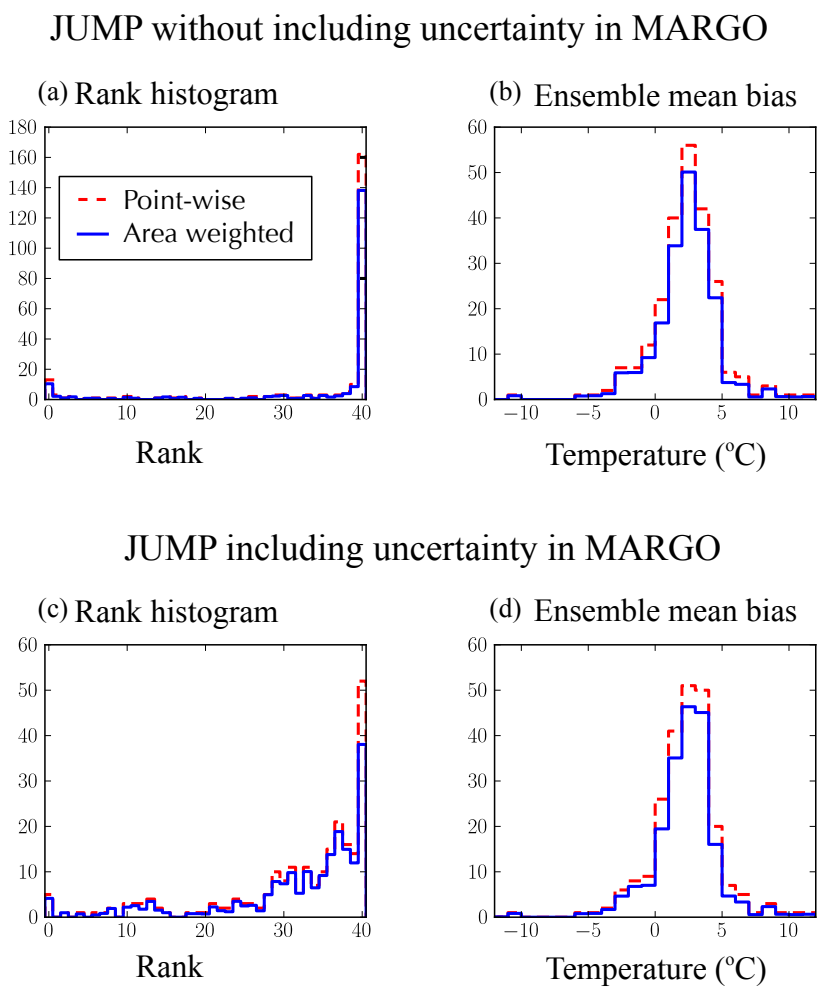

Fig. 8. Rank histogram analysis for the JUMP ensemble JUMP analysis with **(c–d)** and without **(a–b)** inclusion of MARGO data uncertainty. (**a** and **c**) Area-weighted, and point-wise rank histograms. High rank indicates that MARGO is warmer at the LGM than the ensemble. (**b** and **d**) Histograms of the difference between the JUMP ensemble mean and the MARGO data.

thus been to change the amplitude of the variations rather than produce different patterns of variation.These results are consistent with those found in other work analysing several different SMEs from different GCMs (Yokohata et al., 2011). While SMEs may be of great value for understanding the sensitivity and behaviour of a model, these results suggest that caution is required in their interpretation, since they appear to exhibit a more limited range of variation when compared to the PMIP multi-model ensembles.

4.2 Interpretation of the systematic differences between PMIP1 and PMIP2

While model improvements may be expected to have played a role, the principal difference between the PMIP1 and PMIP2 ensembles is probably the incorporation of a 3-dimensional coupled ocean module in all of the PMIP2 models. In this section we consider the effect of including this dynamical element of the climate system by looking at some of the systematic differences between the two ensembles. Since we do not have SST for PMIP1, we start by comparing air temperature (TAS) over the ocean in the two ensembles. The zonal means over the populated MARGO grid boxes of the LGM anomaly for TAS are shown in Fig. 2. The first clear difference is that there is a wider spread in the PMIP1 results. On the one hand we may expect general model improvements over the years between PMIP1 and PMIP2 to have caused the models to converge closer to the data. On the other hand, we might expect that increasing model complexity should increase the uncertainty in model outputs, which would therefore be expected to inflate the ensemble. It seems that at least in the case of TAS, the addition of a dynamic ocean has not increased the inter-model variability. Indeed, particularly around 20°–60° N the TASs seem constrained to follow very much closer to the MARGO SSTs than PMIP1.

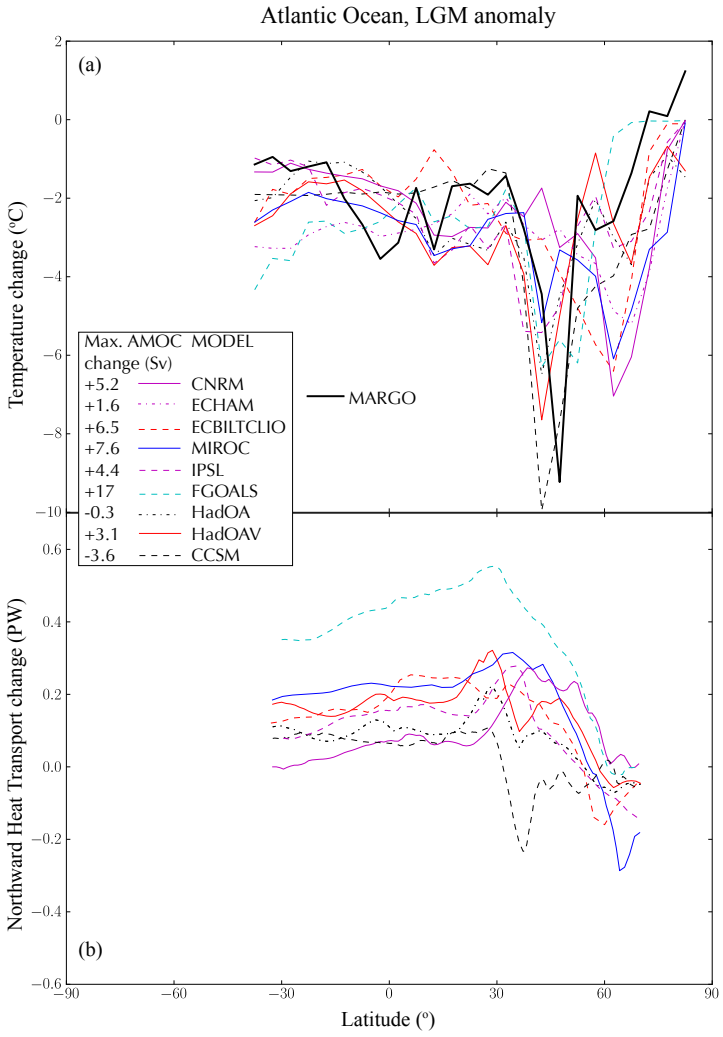

Fig. 9. PMIP2 and MARGO for the Atlantic ocean, zonal averages over the MARGO grid points: **(a)** SST; **(b)** Northward Heat Transport. The NHT is cutoff at 30° S, as further south some of the model outputs appear to include data from other ocean basins. The values of the LGM anomaly of the AMOC maximum are listed in the order of the magnitude of the minimum in the LGM SST anomaly around 40–50° N.

To investigate this further we looked at the sea-ice thickness (variable names in the database, "sit") in the PMIP1 and PMIP2 databases. We found that at the LGM in PMIP2 there is very little sea-ice south of 50° N, whereas there is sea-ice present at this latitude in PMIP1. This explains why the TAS more closely follows the MARGO SST in this region; the ocean is not covered by the insulating layer of ice. Thus it appears that the inclusion of the dynamical ocean components has prevented the over-extension of sea ice that was seen in the PMIP1 simulations, resulting in better agreement with the MARGO data.

We now move our attention to the Atlantic ocean. As the PMIP2 models incorporated 3-dimensional coupled ocean modules, an obvious target of the project was to estimate the state of the LGM circulation. While arguments based on paleoclimate data led to the hypothesis that the main meridional overturning circulation cell in the North Atlantic (AMOC) was both weaker and shallower than the present day (Labeyrie et al., 1992; Sarnthein et al., 1994; Lynch-Stieglitz et al., 2007), there is no direct evidence. The result from PMIP2 did little to either confirm or deny this hypothesis, since the PMIP2 models produced a wide spread of results, ranging from strongly strengthened to mildly weakened circulations (Otto-Bliesner et al., 2007; Weber et al., 2007). The small size of the PMIP2 ensemble means that relationships would have to be very strong in order to pass tests of statistical robustness. Thus our discussion is rather tentative in nature. Figure 9a shows the SST LGM anomaly for PMIP2 and MARGO averaged over the MARGO Atlantic grid points. The deep spike around 50° N indicates where there is sea-ice

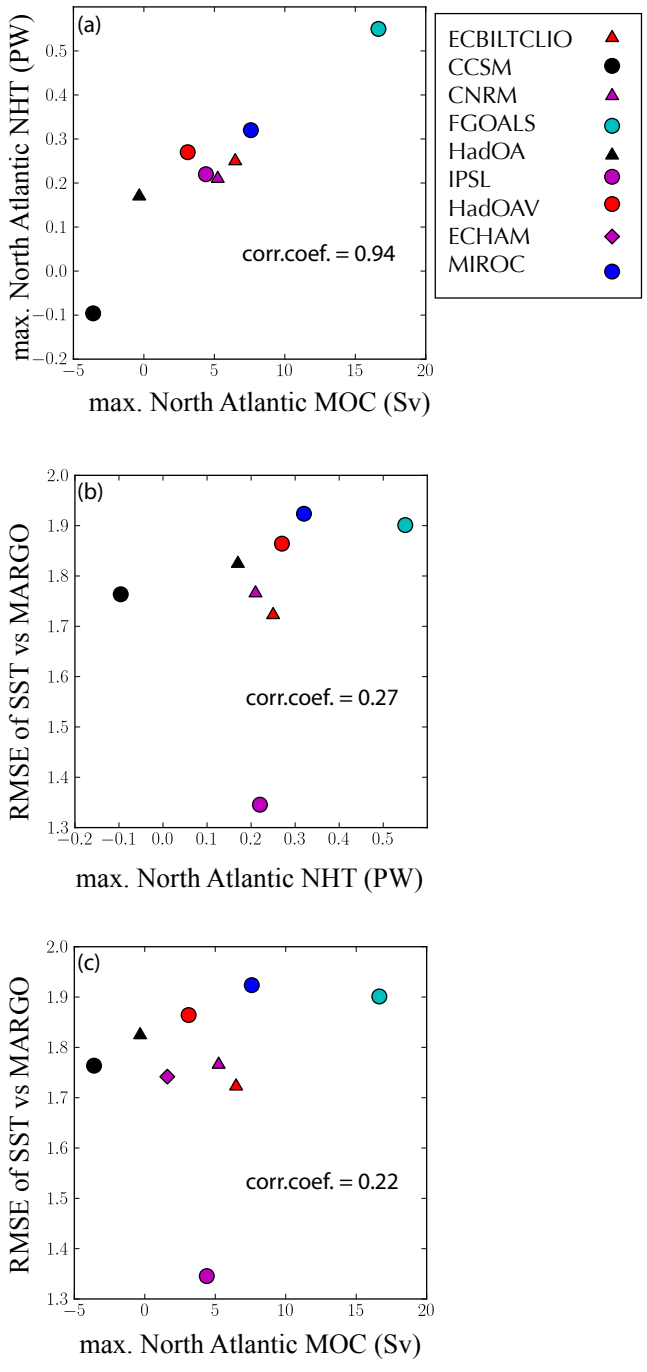

Fig. 10. Correlation between: the root mean square of the difference in Atlantic SST between PMIP2 and MARGO scaled by the MARGO uncertainty; the maximum northward heat transport in the North Atlantic; and the maximum meridional overturning in the North Atlantic.

for the LGM but not the present day, since the area further north where there is sea-ice for both periods does not cool so much at the LGM due to the insulating properties of the sea-ice. On the whole, the models reproduce this spike in qualitative terms, but the magnitude varies. The location of the spike for those models with a single clear sharp spike at close to the correct latitude tends to be a little to the south. In the box on the same Figure are shown the maximum AMOC anomalies for the PMIP2 models. Three of the four models with the smallest maximum AMOC anomalies have the deepest spikes, closest to the observations. Figure 9b shows that the Northward Heat Transport (NHT) in the Atlantic is increased at the LGM at least as far as 30° N for all 8 of the PMIP2 models for which this variable is available. This is another systematic difference caused by adding the dynamical ocean; the PMIP1 slab-ocean models impose the same oceanic heat transport, so the LGM anomaly for Northward Heat Transport is fixed at zero. This systematic difference (consistent with the results of Murakami et al., 2008) indicates that, irrelevant of the AMOC, the dynamical ocean is compensating for the cooling in the northern high latitudes by transporting more heat northward at the LGM. This seems a natural consequence of the greater latitudinal temperature gradient, since it implies that a given volume transport will carry more heat from the tropics to high latitudes than it does in the control climate.

In order to quantitatively compare the models and data, we also calculate the normalised area-weighted root mean square error (RMSE) in the Atlantic for MARGO and PMIP2 LGM anomalies for SST. We weight each squared model data difference by the appropriate grid box area and normalise by the relevant MARGO uncertainty (squared) in order to generate a nondimensional value. We then compare the values obtained to the maximum AMOC and maximum NHT anomalies for the North Atlantic, by correlation. The results are shown in Fig. 10. Most of the models have an RMSE around 1.7–1.9, but the IPSL model has much lower error of 1.35. As argued by Annan and Hargreaves (2011) it is generally expected that, depending on the dimension of the ensemble, the ensemble mean may outperform many ensemble members. In this case only the IPSL model is better; the RMSE of the ensemble mean is 1.46. The AMOC-RMSE and NHT-RMSE correlations are both positive, although not statistically significant for such a small ensemble. The correlation between AMOC and NHT is, however, strong and statistically significant. To sum up: in comparison to PMIP1, all the PMIP2 models have increased NHT in the Atlantic at least as far as 30° N; those with a larger increase in NHT also have a larger increase in AMOC at the LGM, but there is weak evidence that a smaller change in NHT and AMOC is preferred for a good fit to the MARGO data. A significantly larger ensemble size (20–40 members) would be required for robust results to be obtained for the AMOC-RMSE and NHT-RMSE relationships.

5 Conclusions

We have analysed the reliability of two PMIP ensembles and one single-model ensemble using the MARGO sea surface temperature data synthesis for the Last Glacial Maximum. Within the constraint that for PMIP1 only air temperature data can be analysed and this only for the lower latitudes, due to the unavailability of sea surface temperature data, we find that neither ensemble is shown to give unreliable predictions of the MARGO data. The PMIP2 ensemble is somewhat narrower than that of PMIP1, but once the uncertainty in the MARGO data is taken into account, there is no indication that it is too narrow. Rather it seems that model development including the addition of the coupled dynamical oceans in PMIP2 have caused the ensemble to be improved. This work indicates the vital importance of including uncertainty estimates along-side paleoclimate data. If we had not had the error estimate, we might have falsely concluded that the PMIP2 ensemble is too narrow. Further work to better model the data uncertainty estimate would be valuable for analyses such as these.

The JUMP ensemble is found to be extremely unreliable, even when the MARGO data uncertainty is included in the analysis. This result is consistent with ongoing work analysing several single model ensembles for the present day (Yokohata et al., 2011). It would appear that, in order to sample as wide a range as the multi-model ensembles, a much broader range of parametric changes would need to be varied, which would be computationally problematical. Comparing the JUMP and PMIP2 ensembles supports our previous conclusion (Annan et al., 2005b), that the JUMP climate sensitivities of greater than 4 °C are less consistent with the data than the generally lower climate sensitivities of the PMIP2 models.

Comparison of the PMIP2 and PMIP1 ensembles reveals that the addition of the coupled ocean models in PMIP2 have caused systematic differences in the modelled LGM state. It can be inferred from the surface air temperatures that the PMIP2 ensemble is probably more consistent with the data in the high latitudes than PMIP1. There is a less southward extension of sea-ice with the PMIP2 models; the dynamical oceans are causing northward heat transport to be increased at least as far as 30° N. Despite these systematic differences, and a general narrowing of the ensemble, there is a wide range of AMOC strength in the PMIP2 models. Stronger AMOC LGM anomaly correlates with stronger Atlantic northward heat transport LGM anomaly. There is also weak evidence for the PMIP2 models with lower AMOC (and NHT) LGM anomaly being more consistent with the data, which would also be consistent with the observationally derived estimates of the AMOC LGM anomaly. The ensemble size is not, however, sufficiently large to draw confident conclusions, and so it is important that this is increased to enable robust characterisation of the climate system behaviour. The ensemble size for the LGM is expected to increase considerably over the next few years, as the PMIP3/CMIP5 runs become available. Increasing the robustness of these result would also be helped by having available data representative of a range of variables including in the ocean at depth rather than only the surface, and initiatives are underway to increase the scope of ocean data syntheses (Paul and Mulitza, 2009).

Acknowledgements. We are grateful to the three reviewers, C. J. Van Meerbeeck, M. Kucera and T. L. Edwards, whose careful consideration and helpful comments have enabled us to greatly improve the manuscript. This work was supported by the S-5-1 project of the MoE, Japan, the Kakushin Program of MEXT, Japan, and by the DFG-Research Center/ Cluster of Excellence "The Ocean in the Earth System", Germany. We acknowledge the modelling groups involved in PMIP and PMIP2 in making available the multi-model datasets, and the MIROC development team in particular.

Edited by: V. Rath

References

Anderson, J.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Climate, 9, 1518–1530, 1996.

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

Annan, J. D. and Hargreaves, J. C.: Understanding the CMIP3 multi-model ensemble, J. Climate, in press, 2011.

Annan, J. D., Hargreaves, J. C., Edwards, N. R., and Marsh, R.: Parameter estimation in an intermediate complexity Earth System Model using an ensemble Kalman filter, Ocean Model., 8, 135–154, 2005a.

Annan, J. D., Hargreaves, J. C., Ohgaito, R., Abe-Ouchi, A., and Emori, S.: Efficiently constraining climate sensitivity with paleoclimate simulations, SOLA, 1, 181–184, 2005b.

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, 2007.

Claussen, M., Mysak, L., Weaver, A., Crucifix, M., Fichefet, T., Loutre, M., Weber, S., Alcamo, J., Alexeev, V., Berger, A., Calov, R., Ganopolski, A., Goosse, H., Lohmann, G., Lunkeit, F., Mokhov, I. I., Petoukhov, V., Stone, P., and Wang, Z.: Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models, Clim. Dynam., 18, 579–586, 2002.

Conkright, M., Levitus, S., OBrien, T., Boyer, T., Antonov, J., and Stephens, C.: World ocean atlas 1998 CD-ROM data set documentation, National Oceanographic Data Center (NODC) Internal Report, Silver Spring, Maryland, 1998.

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

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

Hargreaves, J. C. and Annan, J. D.: On the importance of paleoclimate modelling for improving predictions of future climate change, Clim. Past, 5, 803–814, doi:10.5194/cp-5-803-2009, 2009.

Hargreaves, J. C., Abe-Ouchi, A., and Annan, J. D.: Linking glacial and future climates through an ensemble of GCM simulations, Clim. Past, 3, 77–87, doi:10.5194/cp-3-77-2007, 2007.

Hasumi, H. and Emori, S.: K-1 coupled model (MIROC) description, K-1 technical report 1, Tech. rep., Center for Climate System Research, University of Tokyo, 2004.

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

Jones, P., New, M., Parker, D., Martin, S., and Rigor, I.: Surface air temperature and its changes over the past 150 years, Rev. Geophys., 37, 173–199, 1999.

Joussaume, S. and Taylor, K.: The Paleoclimate Modeling Intercomparison Project, in: Paleoclimate Modelling Intercomparison Project (PMIP): proceedings of the third PMIP workshop, edited by: Braconnot, P., Canada, 1999, 43–50, 2000.

Kageyama, M., Laine, A., Abe-Ouchi, A., Braconnot, P., Cortijo, E., Crucifix, M., Vernal, A. D., Guiot, J., Hewitt, C. D., Kitoh, A., Kucera, M., Marti, O., Ohgaito, R., Otto-Bliesner, B., Peltier, W. R., Rosell-Mele, A., Vettoretti, G., Weber, S. L., and Yu, Y.: Last Glacial Maximum temperatures over the North Atlantic, Europe and western Siberia: a comparison between PMIP models, MARGO sea-surface temperatures and pollen-based reconstructions, Quaternary Sci. Rev., 25, 2082–2102, doi:10.1016/j.quascirev.2006.02.010, 2006.

Kucera, M., Rosell-Melé, A., Schneider, R., Waelbroeck, C., and Weinelt, M.: Multiproxy approach for the reconstruction of the glacial ocean surface (MARGO), Quaternary Sci. Rev., 24, 813–819, 2005.

Labeyrie, L., Duplessy, J., Duprat, J., Juilletleclerc, A., Moyes, J., Michel, E., Kallel, N., and Shackleton, N.: Changes in the vertical structure of the north-Atlantic ocean between glacial and modern times, Quaternary Sci. Rev., 11, 401–413, 1992.

Large, W. and Danabasoglu, G.: Attribution and impacts of upper-ocean biases in CCSM3, J. Climate, 19, 2325–2346, 2006.

Lynch-Stieglitz, J., Adkins, J. F., Curry, W. B., Dokken, T., Hall, I. R., Herguera, J. C., Hirschi, J. J.-M., Ivanova, E. V., Kissel, C., Marchal, O., Marchitto, T. M., Mccave, I. N., Mcmanus, J. F., Mulitza, S., Ninnemann, U., Peeters, F., Yu, E.-F., and Zahn, R.: Atlantic Meridional Overturning Circulation During the Last Glacial Maximum, Science, 316, 66–69, 2007.

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

Meehl, G., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J., Stouffer, R., and Taylor, K.: The WCRP CMIP3 multimodel dataset, B. Am. Meteorol. Soc, 88, 1383–1394, 2007.

Murakami, S., Ohgaito, R., Abe-Ouchi, A., Crucifix, M., and Otto-Bliesner, B. L.: Global-Scale Energy and Freshwater Balance in Glacial Climate: A Comparison of Three PMIP2 LGM Simulations, J. Climate, 21, 5008–5033, 2008.

Ohgaito, R. and Abe-Ouchi, A.: The role of ocean thermodynamics and dynamics in Asian summer monsoon changes during the mid-Holocene, Clim. Dynam., 29, 39–50, 2007.

Otto-Bliesner, B. L., Hewitt, C. D., Marchitto, T. M., Brady, E., Abe-Ouchi, A., Crucifix, M., Murakami, S., and Weber, S. L.: Last Glacial Maximum ocean thermohaline circulation: PMIP2 model intercomparisons and data constraints, Geophys. Res. Lett., 34, L12706, doi:10.1029/2007GL029475, 2007.

Otto-Bliesner, B. L., Schneider, R., Brady, E., Kucera, M., Abe-Ouchi, A., Bard, E., Braconnot, P., Crucifix, M., Hewitt, C., and Kageyama, M.: A comparison of PMIP2 model simulations and the MARGO proxy reconstruction for tropical sea surface temperatures at last glacial maximum, Clim. Dynam., 32, 799–815, 2009.

Paul, A. and Mulitza, S.: Challenges to Understanding Ocean Circulation During the Last Glacial Maximum, Eos, 90(19), p. 169, 2009.

Petoukhov, V., Ganopolski, A., Brovkin, V., Claussen, M., Eliseev, A., Kubatzki, C., and Rahmstorf, S.: CLIMBER-2: a climate system model of intermediate complexity. Part I: model description and performance for present climate, Clim. Dynam., 16, 1–17, 2000.

Randall, D. A., Wood, R., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R., Sumi, A., and Taylor, K.: Climate Models and Their Evaluation, in: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, chap. 8, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Sarnthein, M., Winn, K., Jung, S., Duplessy, J., Labeyrie, L., Erlenkeuser, H., and Ganssen, G.: Changes in east Atlantic deepwater circulation over the last 30,000 years – 8 time slice reconstructions, Paleoceanography, 9, 209–267, 1994.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., Tignor, M., and Miller, H. (Eds.): IPCC, 2007: Summary for Policymakers, in: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Proc. ECMWF Workshop on Predictability, 1–25, 1997.

Weber, S.: The utility of Earth system Models of Intermediate Complexity (EMICs), WIRES Climate Change, 1, 234–252, 2010.

Weber, S. L., Drijfhout, S. S., Abe-Ouchi, A., Crucifix, M., Eby, M., Ganopolski, A., Murakami, S., Otto-Bliesner, B., and Peltier, W. R.: The modern and glacial overturning circulation in the Atlantic ocean in PMIP coupled model simulations, Clim. Past, 3, 51–64, doi:10.5194/cp-3-51-2007, 2007.

Yokohata, T., Webb, M., Collins, M., Williams, K. D., Yoshimori, M., Hargreaves, J. C., and Annan, J. D.: Structural similarities and differences in climate responses to CO_2 increase between two perturbed physics ensembles by general circulation models, J. Climate, 23, 1392–1410, 2010.

Yokohata, T., Annan, J., Collins, M., Jackson, C., Tobis, M., Webb,

M. J., and Hargreaves, J. C.: Reliability of multi-model and structurally different single-model ensembles, Clim. Dynam., submitted, 2011.

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