

# Reply to Jessica Tierney

Authors

10 May 2022

Thank you for the detailed and interesting questions and comments. To aid readability of our reply, specific changes proposed for our manuscript will be *highlighted in red*. Reviewer text quoted below is *highlighted in blue*.

---

One major change that should be highlighted at the outset is that *we will make our code and results available*. We probably should have done this initially, but wanted to make it a final version rather than having the risk of multiple versions being used. Therefore interested researchers will be able to check our result and further test it themselves in any way they wish. As so much of the review considers the contrast between our results (Annan et al, AEA subsequently) and those of Tierney et al (TEA), we will initially focus on this issue in some detail, before returning to the review comments on a point-by-point basis. We show below that a major cause of the differences between our results is in the respective choices of prior. The priors are substantially different and it is inevitable that this influences the results, especially in this situation where data are sparse and imprecise. Our prior does include a simulation from CESM1-2, the model that TEA used, and our posterior is sufficiently broad as to include much of the TEA range as well as milder values. Indeed as we show below, the ensemble member in our posterior corresponding to the CESM1-2 model is very close to the TEA reconstruction mean.

## 1 Comparison with Tierney et al.

*I think this is a useful contribution to the problem of reconstructing past climate fields and it is very interesting to see the outcome when a slightly different DA approach and multiple different model priors are used. However, the analyses in the paper as it is now don't identify \*why\* these new results are different than Tierney et al.*

While (as we replied earlier to Dan Lunt) our initial intention was not to investigate the analysis of TEA in detail, we do acknowledge that some further analysis is warranted, and interesting to the readers, and therefore have undertaken some further investigations which we describe in this Section. *We will*

*include in the paper extended discussion and comparison with TEA including a direct difference between the reconstructions (similar to Figure 1 attached here).*

## **1.1 Description of differences between our AEA and TEA**

We agree that the Pacific plume in Fig 2a of TEA is not so anomalous, as you correctly stated in your review. Figure 1 displays the difference between the means of the AEA and TEA reconstructions. Two large scale discrepancies seem apparent. Firstly, there is a very strong anomaly in the Southern Ocean, where TEA is cooler by at least 4C (and up to 13C) over a substantial region. It seems possible that the strong cooling in TEA could be due to an increase in sea ice extent, as the SSTs in this area are much more similar between the two reconstructions, and there is a decoupling here between SST and SAT in the TEA reconstruction. However the important point is that the discrepancy exists, regardless of physical origin. The region shown in Figure 2 (defined as the region south of 46S where the TEA anomaly is at least -4C cooler than ours) represents about 5% of the global surface area, but contributes almost -0.4C to the difference in global mean temperature anomalies. PMIP models show a wide variety of results in this area as we shall illustrate shortly. There are very few data points in this region and SST proxy data cannot inform directly on SAT over sea ice as the ice insulates the air from the ocean surface.

The central latitudes of the Pacific basin as a whole (excluding the Southern Ocean areas described above, i.e. as defined in Figure 3) also has a substantial difference of around -2C in TEA relative to our reconstruction. Due to the large area this represents, this discrepancy contributes around -0.5C to the difference in our global means. While there are a few data points here, they are extremely sparse, and some show warming. There are a number of other regions showing differences in Figure 1, including the poles and Greenland, but the areas of these regions are too small for them individually to make really substantial contributions to the global mean temperature difference.

These two regions combined (covering a little more than a quarter of the globe) can explain a large majority of the difference between the global means of the reconstructions, which provides an account for the difference in our headline result, and also explains why both the TEA reconstruction, and our own, can agree reasonably with ice core data which are remote from these areas and hence uninformative regarding them.

## **1.2 Explanation of differences between AEA and TEA**

In this subsection we show that the differences between the results are directly traceable to the difference in our priors, as we had surmised, and Figures 4 and 5 underpin our reasoning.

We used as our prior a large number of PMIP models that explore a wide range of uncertainty in feedbacks that determine their LGM states. Figure 4 shows the PMIP simulations that we use in our calculation (thin black lines, with

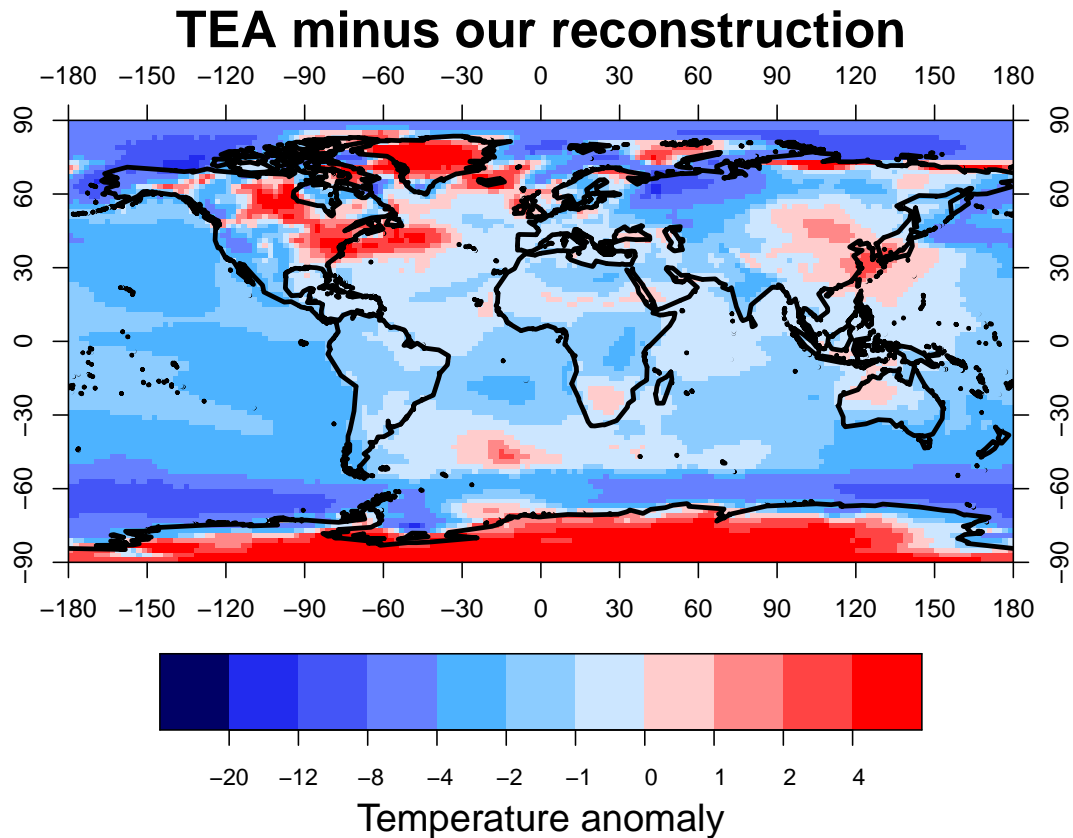


Figure 1: Difference between posterior means of TEA and this work.

the CESM1-2 simulation highlighted in red). We observe here that the CESM1-2 model exhibits the most extreme cooling in the Southern Ocean, and also shows the strongest cooling of the entire ensemble across the tropics especially south of the equator, which is dominated by the Pacific Ocean.

The posterior ensemble in the Ensemble Kalman Filter is generated by updating each member of the prior ensemble according to the Kalman Equation, and thus our posterior contains a member derived from applying this Kalman update to the CESM1-2 simulation. Figure 5 shows the latitudinal variation of our posterior ensemble members, together with the TEA reconstruction mean for comparison. The overall similarity between the latitudinal profile of our posterior sample based on CESM1-2 (red line), and that of the TEA reconstruction mean (purple line), is striking. This member of our ensemble appears more similar to the TEA reconstruction in structure and magnitude than does any other member of our ensemble.

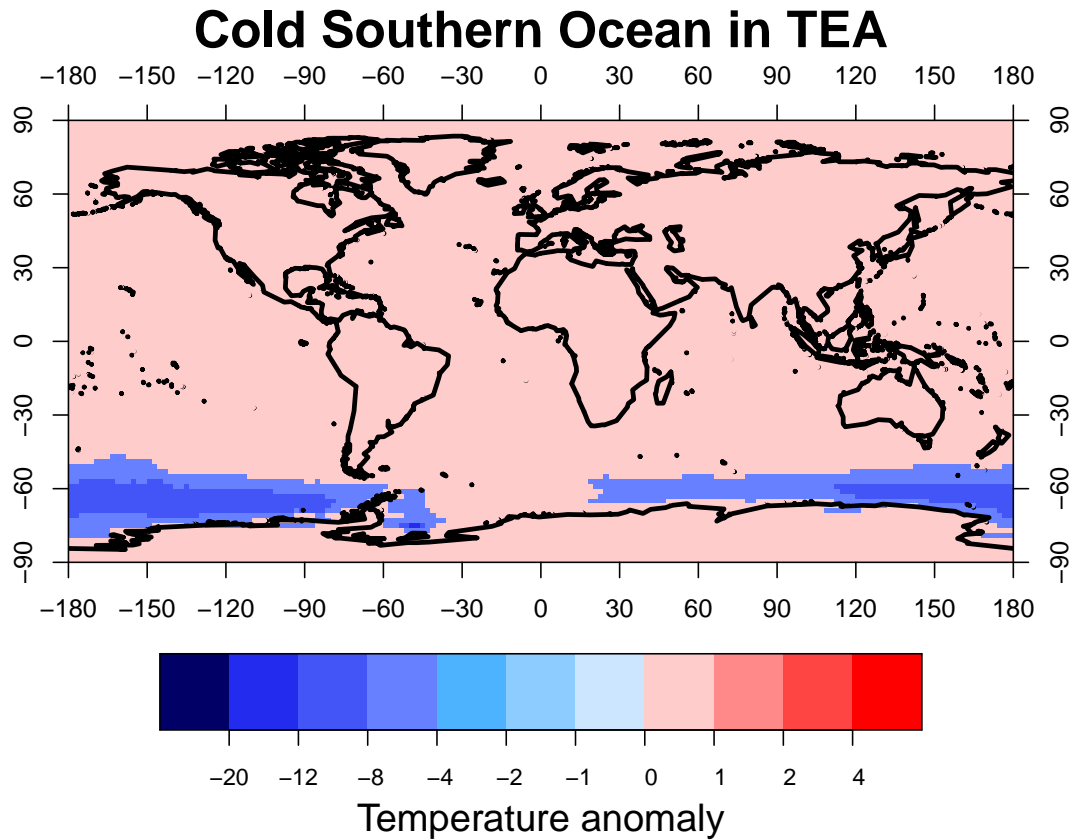


Figure 2: Masking to show region of extreme cold in Southern Ocean. Difference in anomalies exceeds  $-4\text{C}$  in blue area, and this contributes almost  $-0.4\text{C}$  to the difference in global means between TEA and this work.

A simple numerical analysis supports this claim quantitatively. Table 1 summarises some statistics of our posterior ensemble. For each posterior ensemble member we list firstly the model it is sourced from in the prior, then the mean SAT anomaly of this posterior member, and lastly the area-weighted grid point RMS difference between this sample and the TEA posterior mean. The 12th member in the list, based on CESM1-2, is highly similar to the TEA posterior mean, having a SAT anomaly of  $-6\text{C}$ , and a grid point RMS difference from the TEA posterior mean of well under  $2\text{C}$ , which is by some way the lowest value among all of our ensemble members. The ease with which this CESM1-2-derived sample can be identified by to its similarity to the TEA posterior mean is a strong indication of the influence of the prior, which persists despite substantial differences in prior covariance matrix, data set, and other methodological details. In both our calculation and that of TEA, the update of the

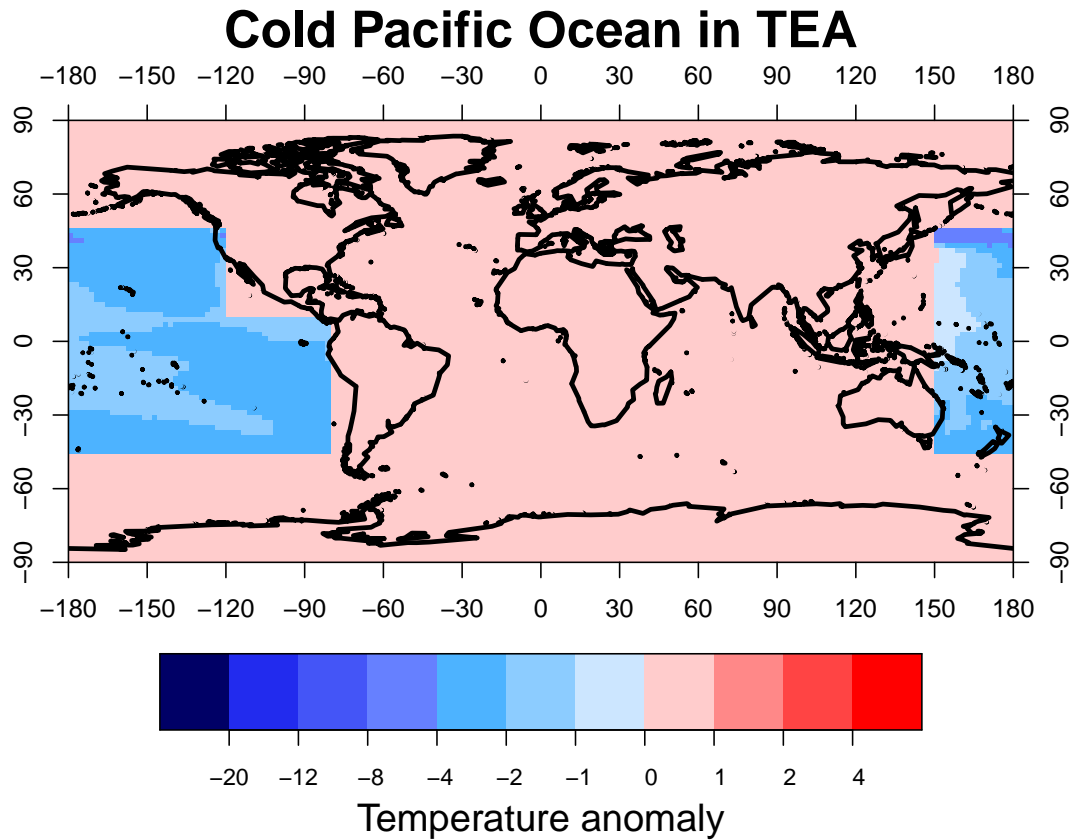


Figure 3: Masking to show region of strong cooling in Pacific Ocean. Difference between LGM anomalies averages in excess of  $-2\text{C}$  in this area, and this region contributes  $-0.5\text{C}$  to the total difference in global means between TEA and this work.

CESM1-2 model leaves us with something that is still recognisably CESM1-2.

We therefore see that our posterior ensemble contains a member which is close to the TEA ensemble mean, and as a result, our posterior result more-or-less includes that of TEA. Specifically, the TEA reconstruction's global mean temperature estimate is within our 2sd range, as is most (67%) of the mean spatial map on an area-weighted gridpoint basis. Conversely, however, our posterior global mean temperature estimate of  $-4.5\text{C}$  is far outside of the TEA posterior 95% range, and a large majority of our spatial map (over 60% of the globe) is outside the 4sd range of the TEA spatial estimate. This is despite our reconstruction showing a high level of agreement with not only the data used in TEA, but the other diverse and independent data compilations that we used.

Model name	Global mean SAT anomaly	RMSD vs TEA posterior mean
CCSM	-4.18	3.05
CNRM	-3.13	4.75
FGOALS	-5.16	2.87
HadCM3	-4.62	3.36
IPSL	-3.59	4.05
ECHAM	-4.96	2.9
FGOALS	-4.25	3.58
GISS-E2	-4.48	3.47
MPI	-4.03	3.5
MRI	-4.34	2.87
CCSM4	-4.81	2.52
<b>CESM1-2</b>	<b>-6.02</b>	<b>1.88</b>
HadCM3-GLAC1D	-5.73	2.24
HadCM3-PMIP3	-6.18	2.9
INM-CM4-8	-3.56	4.3
IPSLCM5A2	-4.54	3.48
MIROC-ES2L	-3.74	3.56
MPI-ESM1-2	-3.79	3.43
iLOVECLIM1-1-1-ICE-6G-C	-3.71	3.45

Table 1: Summary of posterior ensemble members, showing mean temperature anomaly at LGM and also RMS difference of spatial map with that of TEA posterior mean

## Latitudinal variation of initial LGM anomalies of PMIP ensemble

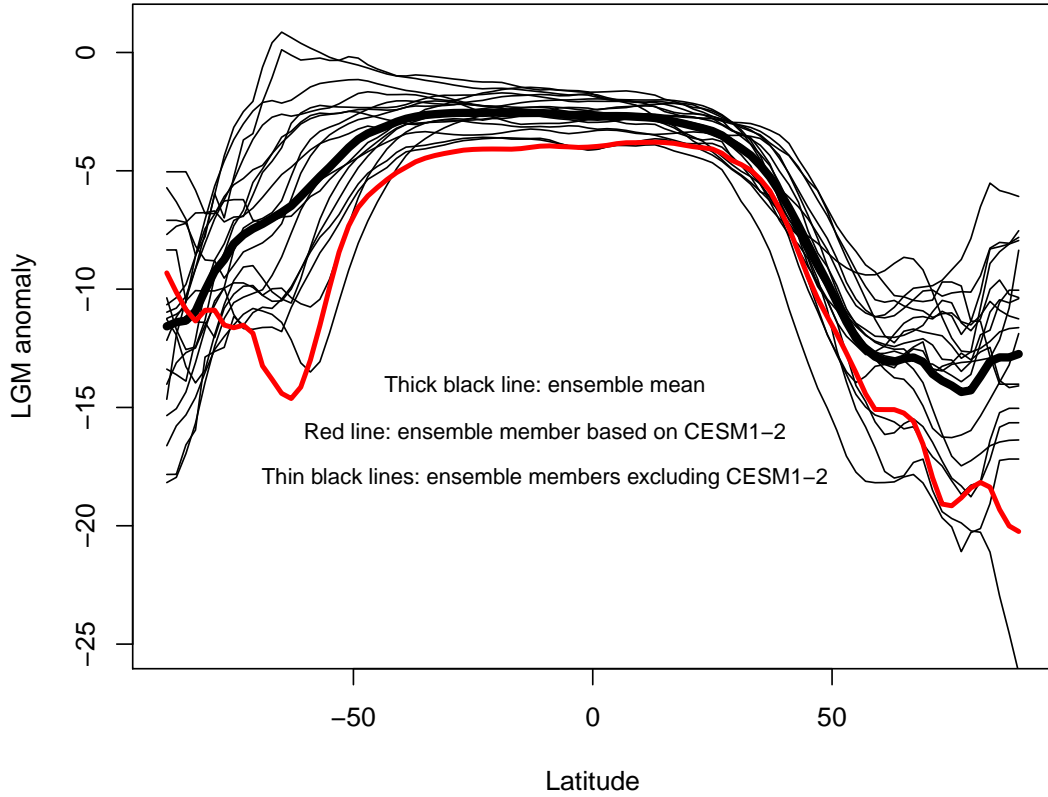


Figure 4: Latitudinal summary of our prior ensemble. CESM1-2 simulation is highlighted in red, with the other ensemble members shown as thin black lines. Thick black line is ensemble mean.

### 1.3 New sensitivity test - Recentering the prior

As a corollary of this analysis, if we recenter our entire ensemble on the CESM1-2 model output, rather than centering it on the data as described in the manuscript, then we can anticipate that our posterior mean would be very similar to that of TEA. In fact when we perform this experiment, our posterior global mean temperature anomaly becomes -6.4C and our mean field is again within 2C of TEA on a gridpoint RMS basis. Thus, the difference in the reconstruction means of TEA and AEA can be directly traced to the difference in prior means. *We will report this experiment in our manuscript.*

## 1.4 Comments on choice of prior

While the prior is a personal choice that researchers may reasonably differ over, it seems doubtful to us that the prior of TEA is suitable for this problem. The prior of TEA does not consider any of the uncertainties in model structure or feedback strengths that contribute to our uncertainty regarding the LGM state. As a consequence, all of the model instances in TEA will have the same equilibrium climate sensitivity. Sampling internal variability does little to mitigate these limitations. The overconfidence of the prior in TEA is apparent from the rank histograms in the supplementary data; the reconstruction of TEA could not allow for the possibility of a significantly milder climate than was obtained, because this was excluded from the prior. This can be seen directly from Fig 2d in TEA where the prior 95% range is restricted to values for global mean temperature of colder than  $-5.6\text{C}$ . This excludes not only the majority of the PMIP simulations — 15 from the 19 we used — but also the mean of the previous reconstruction of Annan and Hargreaves 2013 (AH2013). While a Gaussian has no absolute boundaries, many of these values are outside even the  $\pm 3\text{SD}$  (99.7% probability) range of the TEA prior. It was therefore inevitable that the TEA posterior would also be restricted to a much colder result than AH2013 and would exclude the values generated by many of the PMIP simulations, regardless of the data.

## 1.5 Summary of comparison with Tierney et al.

Based on this analysis, we do not think there is further reason to perform the many additional tests and calculations suggested in order to explore why the AEA and TEA differ. We agree that the data are not inconsistent with a cold ( $\simeq -6\text{C}$ ) LGM, at least at one extreme of our posterior range. Indeed we generate a posterior sample that is very close to the TEA reconstruction mean, directly through applying the Kalman update to the member of our prior ensemble generated by the model that was used in TEA. We have however shown that the data are also consistent with much less extreme cooling than TEA supported, with substantial uncertainty appearing to arise from the different ways that models extrapolate into poorly observed areas. TEA excluded such possibilities *a priori* and therefore cannot attribute this aspect of their result to the new data. Indeed a direct comparison of the gridded proxy data of TEA with that of MARGO at coincident locations suggests that it represents rather similar conditions, albeit marginally cooler, and this is also broadly consistent with the pollen data of Bartlein et al. The TEA prior is very narrow and centred on far colder values than the AEA prior. We do not believe there is any further discrepancy to be explained between our results, beyond what would be reasonably expected as a direct result of these two factors.



## 2 Specific comments

We now address the specific points in your review.

*Osman et al., (2021) which finds an even cooler LGM (-7C) with a more limited proxy network yet much wider priors (this follow-up paper isn't discussed here, but should be!)*

We didn't discuss Osman et al as it is not primarily a reconstruction of the LGM and itself makes the point that it uses fewer data points at each interval in time, therefore cannot be expected to be as accurate a result for the LGM specifically. It seems inevitable that it also suffers from the problem that TEA has of relying on the internal variability of a single model to generate an ensemble, which by construction does not include uncertainties in climate system feedbacks. If the topic of our manuscript were a multi-model data assimilation reconstruction of the deglaciation, it would be more relevant to include comparison with Osman et al.

*Conversely, it seems like the relatively warm priors here are influencing the result also, given that the adjusted prior mean is -4.9 p/m 1.1C.*

The prior is centred on the data via the initial translation step and thus is neither relatively warm nor cool.

*The authors also find that if they inflate the prior mean cooling to -9.7C, they get a substantially cooler posterior (Page 14), unless they do their "translation procedure", which re-centers the prior on the data a priori. But then, isn't the result basically dependent on pre-centering the prior on the data – which means you presuppose what level of cooling the data show? I think this could be problematic. The data might show a mean cooling of say -4.5C, but the data don't sample the coolest places on Earth (like the Laurentide) so this is almost certainly an underestimate. To what extent does the translation step bake-in the posterior result?*

Since we do the recentering on a spatial basis using gridded data, the location of data points will not in itself bias the result if the spatial patterns of the models are reasonable. Thus cold unobserved places like the Laurentide ice sheet do not affect the translation.

*In Tierney20, we calculated a proxy-only global mean cooling to compare with the DA and found it has a median of -5.5C. This was based only on SST, so reasonably could be too warm since it excludes land data. So first order, the -4.5C result here just does not seem cold enough. Granted the -5.5C we calculated has to assume a scaling b/t SST and SAT (we followed Snyder et al, 2016 and drew this scaling from PMIP). Nevertheless you could calculate a proxy-only value here for comparison, and also compute a proxy-only global SST (which doesn't have to be scaled). Can you calculate proxy-only global SST change and*

*compare to the DA results (as we did in Fig. 2 of Tierney20)? We got -2.9C with the TEA network. I'm just curious whether the DA proxy posterior comes out similar to this or warmer.*

We emphasise that our posterior range includes -5.5C (indeed this value is only just over 1sd from our mean) and such a result is therefore not particularly unlikely. We do not think there is any substantial discrepancy to explain here. The simple mean of the 405 data points we used is -3.2C, which is certainly an underestimate of the global mean temperature change as the data are dominated by SST proxies and are not missing at random. Performing the fit at the correct locations accounts for this, on the assumption that ensemble of models contains reasonable spatial patterns.

*In Tierney20, for sure our prior was a bit too tight than we liked which we discussed in the paper and is evident in the U-shape of the rank histograms. However, interestingly, there was no strong mean bias in the LGM rank histogram (but there was a slight mean bias for the Late Holocene) which suggested to us that the LGM prior was not too cold; otherwise we would expect to see a skewed shape in the rank histogram. This is different than the rank histograms shown here which are much more flat. I'm not sure how to reconcile this. Our rank histograms are based on withheld validation data only (not sure if that is the case here or not). They were also calculated in proxy space, taking advantage of the forward modeling in our DA technique. So how to compare those to these?*

The direct reconciliation of the U-shaped rank histogram of TEA versus the flatter one of AEA is that the TEA prior is strongly overconfident, which is as expected given the use of a single model. Our prior has a much broader range of model responses, due to including structural uncertainties (as represented by the PMIP models) in its construction. This result is consistent with several papers we have published over the past decade (eg Yokohata et al, 2012 and 2013), and also an independent analysis that was published more recently (Parsons et al, 2021).

We agree that your rank histograms do not show a strong bias, and the same is true of ours. While this might on the face of it seem inconsistent, the data are not missing at random, as we noted in AH2013, which makes the extrapolation into data voids a highly model dependent process. Therefore, it is necessary to consider the full range of plausible model results rather than focussing on just one model, especially given that the CESM1-2 model appears to be such an outlier for the data-poor Southern Ocean and Pacific mid-latitudes. It is also worth mentioning that the rank histogram is not a very comprehensive test, and while failing to obtain a reasonably flat rank histogram is certainly indicative of a problem, it is also quite possible for a poor ensemble to pass this test satisfactorily. A flat rank histogram is a necessary but not sufficient condition.

*ERSSTv5 is used here as a preindustrial baseline instead of the Late Holocene*

*values. What happens if you use the Late Holocene SSTs in TEA instead? Our reasoning for using Late Holocene SST was that in some locations there is a strong bias in the proxies (relative to climatological SST) so subtracting LH SST would correct for this. This said, LH SST estimates from the TEA network might suffer from non-modern coretops. In Tierney20, we did not screen the data for age control, so some of the coretops might have pre-modern (likely colder) proxy information. We speculated that this is one reason why Osman et al LGM cooling is larger. In that study, we only used proxy data with good age control so the Late Holocene data are for sure Late(est) Holocene. Perhaps it is worth running your DA using the screened proxy network of Osman et al. alone?*

As described in the text, we found a slightly better agreement between MARGO and TEA when we used their absolute values and a common pre-industrial baseline, versus the published anomalies for both (which use different baselines). However it is worth emphasising that the difference is minor. Since our method is based entirely on anomalies, using the core-top data of TEA would restrict us to the subset of gridpoints where both core-top and LGM are co-located. We accept there are arguments in favour of using core-top in that biases may cancel, but on the other hand, it requires another set of imperfect measurements and calibration. *We will publish our code so anyone interested in performing a wider set of calculations will be free to do so.* However, we think our choices are reasonable. If there were major internal inconsistencies between the respective anomaly and absolute data sets of either MARGO or TEA, this would be problematic, but we don't see any evidence for this.

*Working in temperature vs. proxy space. I suspect this could produce some of the differences between the result here and Tierney et al (and Osman et al). For one, in these latter studies the forward modeling allows us to consider the seasonality of the proxies and assimilate them during different production seasons. In contrast, it seems that here all of the proxies are treated as annual mean T (?). What happens if, instead, you consider the seasonality and assimilate the proxies in their respective seasons (as we did)?*

We performed seasonal reconstructions in Annan and Hargreaves, Quaternary Science Reviews 107 (2015) using MARGO and Bartlein et al data and found only modest differences from our annual mean calculation. *We will cite this paper and mention the seasonal aspect.* We do agree that working in proxy space should be intrinsically superior if the proxy models are good enough. Of course our results depend on the accuracy of all of the gridded data sets we use, a point that we make in the manuscript. We have however already shown that the choice of prior can explain the differences between our results. If the proxy assimilation produces significantly different values from the calibrated estimates at those grid points then that would be interesting, but we suggest that is a question for the respective scientists to address, not us.

*Page 5, Line 20: I agree that CESM2 is an outlier and probably needs to*

*be removed to satisfy the condition of normality, but it is noteworthy that the resulting ensemble mean is 4.5 p/m 1.1, which is rather tight.*

Please note that the uncertainty quoted here is 1sd and thus our prior uncertainty on the global mean is rather more than double that used by TEA. We therefore do not agree that it is rather tight, but we do report on tests where it was inflated.

*The authors can test the theory that the CESM priors are the main source of the difference by applying their approach to a CESM-only ensemble.*

We have shown above that centering our prior on the CESM1-2 simulation provides a posterior mean very close to that of TEA. Further tests of the priors don't seem necessary.

*Finally, validation. One of the strongest pieces of evidence that the solutions of Tierney20 and Osman21 are robust is the fact that we excellent independent validation with the ice core data. If LGM cooling were only -4.5C globally, I doubt we would get as good of validation. In contrast, there is no external validation in this study*

The reason that there is no additional data other than that we used, is that we used all the data that were readily available so as to generate as high a quality result as possible. While we could have presented a main result that only used some of the data, it would be less well constrained and thus we do not consider this criticism valid. The result we obtain is essentially unchanged for all of the sets of 404 data points anyway, and the left out data provides the validation in these cases.

*Finally, Seltzer et al. (2021) ...*

There are many interesting analyses of a small number of data points giving a range of results often focussed on particular regions. However, we are not experts in the interpretation of data from proxies. Our aim with this work is rather to generate a comprehensive global analysis using the available large and widely accepted compilations of data, and publicly available model outputs, in order to estimate what these information sources combined can tell us about LGM temperatures.

*Page 8, line 15: The TEA database includes all of the data from MARGO that are d18O, Mg/Ca, and UK37, so if you are adding MARGO estimates double check that it is only the assemblage data, which we did not use. Otherwise the same data might be used more than once when you combine the datasets.*

Where data are co-located, we use the TEA data set in preference to MARGO, rather than averaging the two together. Thus there is no double-counting.

*Page 10, line 10: Can you clarify here whether you are doing serial updates (one observation at a time) or a joint update (all observations together)? Tierney et al. and Osman et al. are joint updates. There can be some differences*

*b/t using serial vs. joint although they are usually small.*

The calculation is joint.

*Page 11, line 14: The authors take issue with the cold N. Pacific gyre signal in T'20, but they also show this feature in their reconstruction with a cooling of -4 to -8. Indeed, most PMIP models indicate a strong cooling here and the proxy data also suggest this. In T'20 it is indeed larger (-6 to -10), but I don't think we can say which magnitude is more accurate given the uncertainties in the proxies, and also the lack of proxy data from the center or eastern side of the gyre. Note that Gray et al. 2020 (PaleoPaleo) document gyre cooling of about -5C on average, but deeper cooling is seen in some of the raw Mg/Ca and UK37 data. It seems to me that it's hard to argue what the exact magnitude is. I would either drop this sentence or introduce more nuance here.*

Agree as mentioned previously. *We will refer instead to the Southern Ocean and Pacific more generally, and intend to display the direct comparison of our gridded mean results in the form of Figure 1 here.*

*Page 12: Rank histograms. Are these calculated on the withheld validation data (they should be)? Please clarify.*

We did not combine the test of biased priors, together with the leave one out analysis. Thus the question does not arise here.

*Page 17, Line 9: I would not go so far as to argue that using calculated SST/SAT from the proxies is better than forward modeling, i.e. working in proxy space. This \*might\* be true for proxies that are univariate, but it is definitely not true for proxies that are not, which includes d18O, Mg/Ca, foram transfer functions, and pollen. To translate these to temperature, one has to make assumptions about the other environmental influences on these proxies (i.e. pH, salinity, or pollen, moisture balance) that are going to be imperfect compared to the forward modeling scenario in which multiple environmental parameters can be accounted for. Unless you can prove otherwise, using derived temperatures will be inferior to working in proxy space where inversion is not needed and you can make fewer static assumptions.*

We agree and explicitly state that forward modelling of proxies, where available, has the potential to be superior to using SST/SAT from a statistical calibration. *We will reword this to be clearer.*

*Page 11, line 16: With this many data points, I think leaving out a percentage of the proxy data, as was done in Tierney et al. and Osman et al., is a more robust test than leave-one-out. I suggest withholding 20%-25% of the data.*

Leave one out allows a comprehensive test of predictive ability for the scheme as published. The main reason for leaving out a larger random set would be if computational costs precluded testing each data point in isolation, which is not the case here. We did however also test the omission of various combinations of TEA, MARGO and Bartlein et al, which is a far stiffer challenge than using

random subsets as it tests the possibility that these data sets have different characteristics such as coherent biases due to the different approaches of the different groups of researchers. However, as described in the text, we found only modest differences.

### Latitudinal variation of posterior LGM anomaly

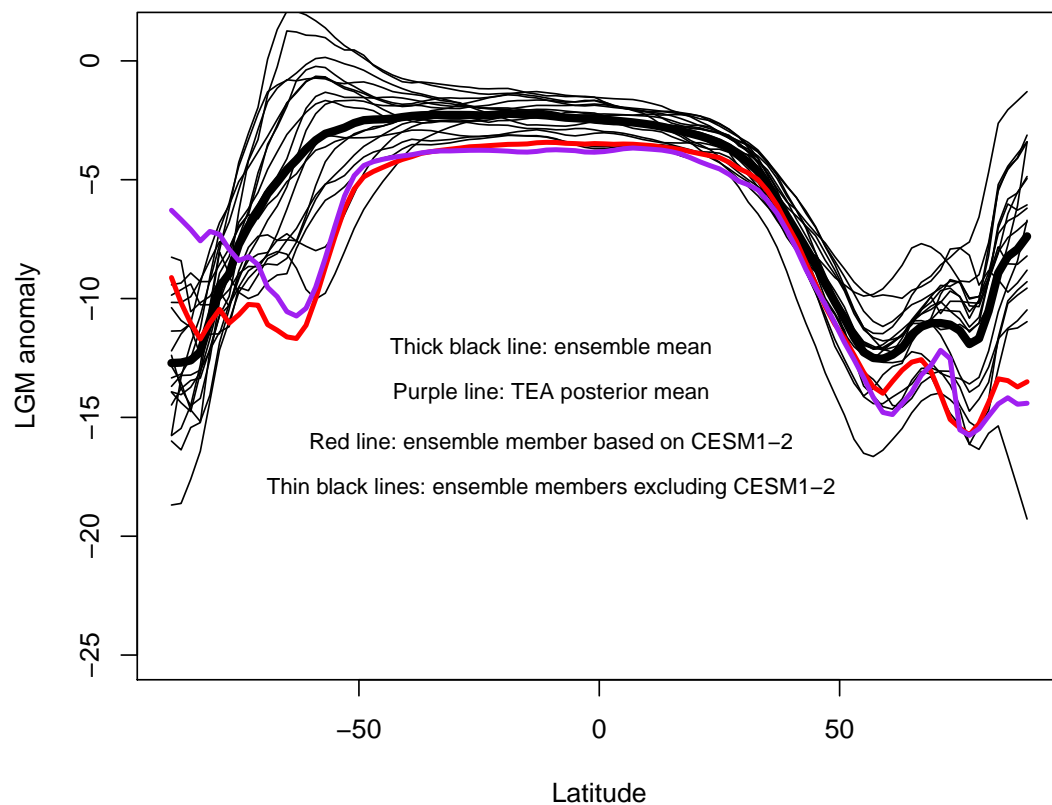


Figure 5: Latitudinal summary of our posterior ensemble. Sample based on CESM1-2 simulation is highlighted in red, with the other ensemble members shown as thin black lines. Thick black line is ensemble mean. Purple line indicates TEA posterior mean.