We thank the reviewers for their insightful comments. Below, their comments are written in grey and our answers are in blue.

Reviewer 1:

General comments

- The reason why I suggested that the 50ppm run together with the 1/8xCO_2 run are the most insightful runs is the following: The "true" inception temperature of this model must be somewhere between the inception temperature derived from the 1/8xCO_2 run and the final equilibrium temperature in the 50 ppm run. That is, these two temperatures set the upper and lower bound for the inception temperature of the model. Therefore, I think that taking the mid point of these two temperatures is potentially a better estimate of the model's inception temperature than simply taking the derived inception temperature of the 50ppm run. From Fig. 4A) it seems like that would give an inception temperature of around 1°C, assuming that the value of -14.7°C temperature change translates to -0.3°C in absolute temperature. My recommendation would be to add this information in a sentence, but since I haven't made this clearer in my earlier reviews, I will not insist.

We agree with the reviewer than the "true" inception temperature probably lies between 1/8xCO_2 and the 50 ppm run, however, we cannot currently calculate a midpoint between both simulation as the reviewer suggests, because the 50 ppm did not reach an equilibrium state. We had to stop the 50 ppm run before it reached an equilibrium, as it is much slower than 1/8xCO_2 (and consequently much more expensive to run). This is specified in the methods section but we have now added this as well in the caption of Fig.4. Comparing the shape of 50 ppm and 1/8xCO_2 (the fact there is a small dip, then a plateau), both runs are strikingly similar, so it is more likely than both transition temperatures would be almost the same. Ideally we would test any CO_2 concentrations between 35 and 50 ppm, however we cannot recommend this in our paper as this represents thousands of years of simulation and therefore months of running time for most models.

Specific comments

- The inception temperature anomaly of -14.7 K compared to PI would translate to an absolute inception temperature of -1.05°C, when taking a standard PI temperature of 13.65°C. How do the authors arrive at the -0.3°C value mentioned in their response? Is it because the PI of MPI-ESM1.2 has a warmer PI mean temperature (i.e. 14.4°C)? Please clarify in the manuscript. It is fine to stick with "close to 0°C " for the rest of the manuscript, but I believe the actual numbers should be clarified at least once. This would also give a better reference for other modelling center that might try to reproduce the runs.

Yes, the PI temperature of MPI-ESM1.2-CR is at 14.4°C. We believe it is common for models to tune the temperature of their PI towards 14.4°C (it is also the case for CESM1.2, which is also presented in our paper). However, the reviewer is right that it is not necessarily obvious, and it could be that some models have a PI temperature which is lower than that. We specified in the methods section the temperature of the PI of MPI-ESM1.2. We wish to highlight that in principle this should not affect the results: as shown when comparing our PI and LGM transitions, the starting temperature does not influence the transition temperature, it only influences how fast the transition is reached. Climate models that have a PI state at 13.6°C could run less years to reach a transition, but the transition should still be expected close to 0°C, and not necessarily a temperature much colder than that.

- Table 1: The added column for the inception temperature has values given as anomalies with respect to pre-industrial conditions, but neither the column header, nor the table caption specify this, making it seem like these are absolute temperatures. This is especially confusing, as later in the text and e.g. in Fig.1 the inception temperature is not referred to as an anomaly, but as an absolute temperature. Please make it clear in this table that this is a temperature anomaly, or (even better) give the values in absolute temperatures here as well

We thank the reviewer for pointing out this oversight. We have added the fact that the temperatures are anomaly in the caption of the table.

Reviewer 2:

Complete Report in attached pdf.

Unfortunately we could not find the report which should include the reviewer's code and screenshots, but only a copy of the following comment as an attached PDF. We will answer this comment based on the description of the reviewer. In this answer we call our figures using letters, such as "Fig.A", to differentiate with the figures using numbers taken from other published papers.

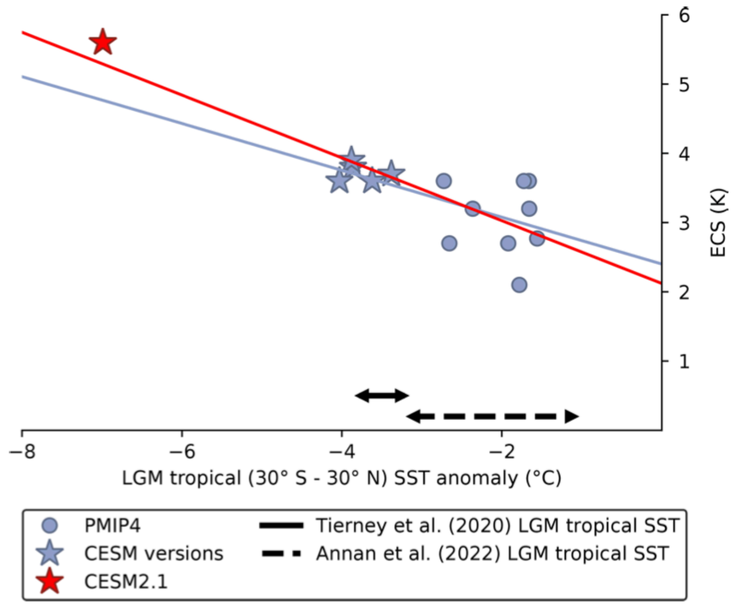
We would like to highlight first that the comment of the reviewer is regarding the use of PMIP4 as a dataset in the method, but it is not criticising the validity of the method itself. The method is an emergent constraint-based approach, which essentially describes that models with higher climate sensitivity are expected to produce colder LGM states. This relies on the simple fact that long-term quasi-equilibrium temperature is expected to be mainly driven by radiative forcing. There have been several older studies which relied on this simple fact because there is no real reason to believe that ECS, a CO2-based metric would not be correlated to some extent to the temperature change at the LGM (e.g. Crucifix et al., 2006; Hargreaves et al., 2012; Renoult et al., 2020).

However, there can be several reasons for the relationship to be more or less significant: for instance, during PMIP2, the relationship was significant in particular over the tropics, as variations in climate sensitivity were dominated by tropical feedbacks (Hargreaves et al., 2012) or during PMIP4, the relationship could be particularly robust if it was based on single-model ensembles, or model-family ensembles, as this could drastically reduce the noise from uncertain ice sheet forcing across different models. During PMIP3, the relationship was the least significant, which also shows that model advancement and technology does not necessarily affect the robustness of the relationship.

Therefore, the method is not invalid, and has the potential to become better in the future: either because climate models used for the upcoming PMIP5 could have a much reduced uncertainty on their LGM states, or because we could find better ways of properly taking into account the characteristics of the LGM in statistical methods, for instance by considering surface temperature pattern effects (Cooper et al., 2024).

That is said, we will now answer the reviewer's comments and show that PMIP4 is a valid dataset to use in our case.

The point cloud in Fig. 7 without the CESM model looks so scattered to me, that I decided to read out the data with an online tool (plotdigitizer.com) and perform the regression myself. When excluding the upper left point in Fig. 7 I find an r-value of only 0.12, and a p-value of 0.74, contradicting the authors'



claim that the relationship is statistically significant without including the outlier model. From this quick analysis, I infer that LGM TS and ECS are uncorrelated or only very weakly correlated among models when excluding the CESM model, in contrast to what the authors asserted. Only when including the CESM model does the relationship pass basic tests of significance (r=-0.76, and p=0.01). Furthermore, the slope changes drastically when including (slope=-0.29 K/K) vs. excluding (slope=-0.05) the outlier. Finally, the data seems to be different from what is shown in Fig. 13 of Renoult et al. 2023, where the statistically significant relationship in PMIP4 is obvious to me.

The data shown in Fig.13 of Renoult et al. (2023) are indeed slightly different, as for this manuscript we used the temperature values reported by Kageyama et al. (2021), which was the lead paper of the PMIP4 multi-model comparisons, where the data are more readily available (and often come from the published papers which are referred within Kageyama et al., 2021). There can be differences with the data used in Renoult et al., (2023), because of 1) approximations, 2) some temperatures being reported either as surface or air-surface temperatures, or 3) differences in climate sensitivity computations either in the computation itself or if they are based on air-surface or surface temperatures. We apologise as we should have indeed specified in the methods that we are using the temperatures of Kageyama et al. (2021), which we have added now.

However this leads to very minor differences that are in fact negligible in the robustness of the relationship, so perhaps the reviewer is referring to the difference between Fig.14 and our manuscript. In this case, Fig.14 is using tropical sea-surface temperatures instead of our manuscript which is using global surface (land + ocean) temperatures. In Fig.14 of Renoult et al. (2013), the PMIP4 ensemble was extended with a CESM-"family" small ensemble, composed of versions of CESM1.2, CESM1.3, and low-ECS versions of CESM2. The idea was to strengthen the PMIP4 ensemble with those models which had a slightly higher ECS than the core of the PMIP4 ensemble, and also avoid the outlier effect of the original CESM2 which had a very high ECS.

In Fig.A, we copied the Fig.14 of Renoult et al. (2023). When comparing the blue relationship (PMIP4 without original CESM2 + CESM-family) and the red relationship (PMIP4 + CESM-family + CESM2), we notice that the slope and intercept are relatively similar and in the case of predicted ECS, whether the high-ECS CESM2 was included or not would only slightly change the median of the inferred ECS. The p-value or r2 coefficient are not shown on the figure of Renoult et al. (2023), however the data points come from published and public data so we can add them now. In the case of the blue relationship, the r2 is -0.60 and the p-value is 0.022; for the red relationship, the r2 is -0.84 and the p-value is 1e-04. As one could expect, including the high-ECS CESM2 strengthen the relationship, however both relationships are significant. In this particular case, the significance using PMIP4 does not require CESM2, which appears as an outlier.
Fig.A: LGM tropical SST anomaly versus ECS, as shown in Renoult et al. (2023), which compares the influence of CESM2 on the robustness of the relationship.

Fig.7 of our manuscript shows that the relationship is also significant without the CESM-"family" small ensemble, therefore we could conclude by combining Renoult et al. (2023) and Fig.7 that CESM2 and the CESM-"family" ensemble are exchangeable. As we do not have access to the code of the reviewer, we can only guess that the reviewer compared p-value between a case where only the PMIP4 ensemble was present (without CESM2 and the CESM-family), and a case where the ensemble took into account CESM2. If testing this, we indeed find the same conclusion as the reviewer (as well as similar p-value and r-value): the relationship is non-significant without CESM2. However, this is not what we attempted to describe in our manuscript, as we wished to describe that CESM2 and the CESM-family ensemble are exchangeable, meaning the relationship is significant as long as one of the two is present; therefore CESM2 is not really an outlier, but the relationship only requires to have a wider range of ECS. We have clarified this in our methods.

However, we did an oversight here that requires us to add a figure as appendix. Fig.A uses tropical SSTs, while Fig.7 of our manuscript global values. While there should be a relationship between ECS and global temperatures (there is no reason to believe global temperatures would not be correlated to CO2 change when tropical temperatures are), the significance could be affected by extra-tropical, uncertain forcing at the LGM, and unfortunately the global SSTs were not included in Renoult et al. (2023). Therefore, on a global scale we cannot directly conclude that the CESM-family ensemble and the original, PMIP4-CESM2 are exchangeable using only Renoult et al. (2023).

In the same style as Fig.A, we have produced Fig.B which is using global SSTs and the same models as the original figure.

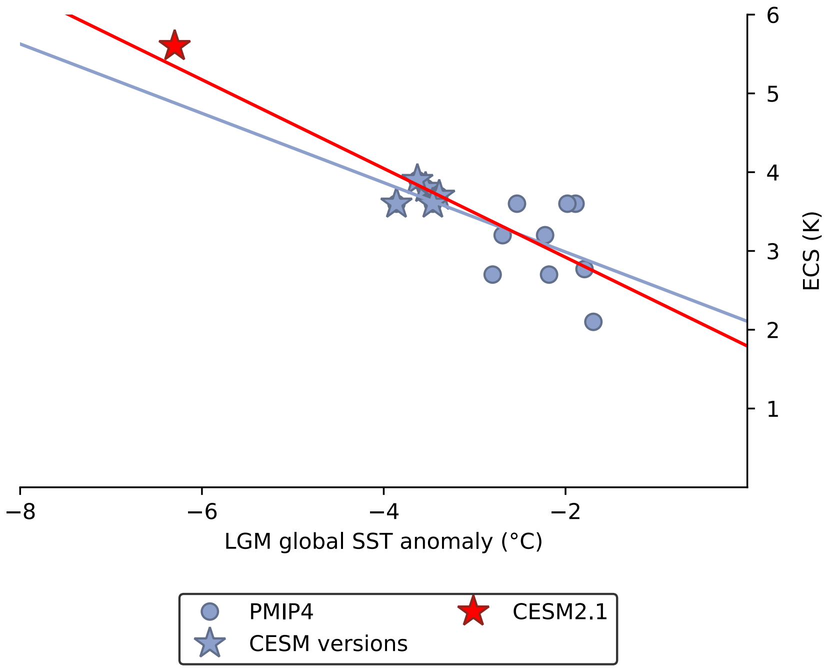


Fig.B: Same as Fig.A, but using global SST anomaly.

In Fig.B, the p-values are the following

- For the blue relationship (everything but high-ECS CESM2): 0.016
- The red relationship (all models): 6e-05
- PMIP4 models and the high-ECS CESM2, without CESM-family: 0.0018

In all three cases the relationship is significant, which means that to attain at least significance, CESM2 and the CESM-family are exchangeable, which was the goal here as we simply wanted to use the original PMIP4 ensemble. We have added Fig.B in Appendix, as well as more details in the methods section. We hope this explanation answers the reviewer's questions.