Reviewer 1

Summary

The authors present a methodology for weighting CMIP6 models based on several performance metrics as well as on their independence from each other. This provides narrower bounds on future global mean temperature changes than in the unweighted ensemble, primarily by down-weighting the highly sensitive models that happen to have poor performance with respect to two reanalysis products and/or are closely related to other models. I found the paper to be nicely motivated, well organized and supported, and a useful contribution to the literature. There are a few areas that I think need to be clarified, and so I recommend minor revisions.

We thank the reviewer for the positive assessment and for the comments on our paper. Please find our answers to the comments highlighted in bold below. We have attached the current draft of your manuscript and we refer to it as 'revised manuscript' . Note that this version of the manuscript might still be updated before the official re-submission.

Major Comments

* Figure 1 and the discussion around lines 241-242: the terminology of 0% to 100%trend-based seems too ambiguous to me and should just be written out explicitly. Couldn't the terms that are included just be stated explicitly in the figure? The figure doesn't really stand on its own, since one has to refer to these lines to know what exactly is meant by these. Additionally, it is not clear what the intermediate values (33%,50%, 66%) correspond to exactly. Upon multiple readings, I still cannot understand what is meant by these percentages at all, and I'm not completely sure what is actually meant by "50% tasTREND and 50% anomaly- and variance-based diagnostics" that forms the basis of the remaining analysis. Please clarify.

Thank you for pointing this out. The reviewer is correct, our notation in the original manuscript was ambiguous. What we are doing in our analysis is splitting 5 diagnostics into two parts: 1) tasTREND, 2) tasANOM, tasSTD, pslANOM, pslSTD. Each of the categories in figure 1 relates to the relative importance of tasTREND compared to the other diagnostics, i.e.:

- **0% tasTREND + (25% tasANOM + 25% tasSTD + 25% pslANOM + 25% pslSTD) [termed 'not-trend based' in the manuscript]**
- **33% tasTREND + (17% tasANOM + 17% tasSTD + 17% pslANOM + 17% pslSTD)**
- **50% tasTREND + (13% tasANOM + 13% tasSTD + 13% pslANOM + 13% pslSTD)**
- **66% tasTREND + (8% tasANOM + 8% tasSTD + 8% pslANOM + 8% pslSTD)**
- **100% tasTREND + (0% tasANOM + 0% tasSTD + 0% pslANOM + 0% pslSTD) [termed 'only tasTREND based' in the manuscript]**

(values not summing up to 100% is due to rounding)

We have adjusted the paragraph in question as well as figure 1 in order to make this clearer (see figure 1 and line 259f in the revised manuscript).

* Discussion of Figure 2 around line 270: Should one have intuitively expected this from the math? I cannot seem to rationalize why using a model that is close to the CMIP6 MME to weigh CMIP6 would pull the CMIP6 MME mean away from the pseudo-observational "truth". This seems like a deficiency in the weighting. Shouldn't the weighting be resilient to this and do very little "harm" in this case?

Again, thank you for pointing this out. We did not mean to say that cases in which the perfect model is close to the unweighted MME *necessarily* lead to a decrease in skill and there are several examples where this is not the case (e.g., for pseudo observations from CanESM2 or IPSL-CM5A-MR; see figure S2 in the revised manuscript). It is crucial, however, to point out that when we write 'close to the truth' we mean close to the truth in the evaluation periods (2041-60 or 2081-00). These periods are not used to inform the weighting and it is possible (in a pure model world as well as in the real world) that the information drawn from the past does not lead to a skill increase in the future if the constraint from the past is unrelated to the future projection. We have adapted our discussion of this topic to be clearer (see lines 291-308 in the revised manuscript).

In addition, skill might be dependent on the emission path. Looking at the time series plots using IPSL-CM5A-LR as pseudo-observations (figure S2 in the revised manuscript), for example, we see a slight downward shift of the distributions for SSP1-2.6 as well as SSP5-8.5. For the former, this leads to an increase in skill while it reduces skill for the latter. We have added a short discussion on this topic to the revised manuscript in lines 309-313 .

We have also added additional information about the skill for each CMIP5 model used as pseudo-observation to figure S2 in the revised manuscript. Finally, we note that figures 2, 4, S2, and S4 have been updated in accordance with a comment from reviewer 2 (see last paragraph of our answer to their comment 10). For each CMIP5 pseudo-observation we now exclude the direct CMIP6 predecessors (if existing) from the calculation (see line 236-237 and table S5 in the revised manuscript).

* Figure 4: The combined and performance-only weights are shown, but not the in-dependence weights. Is there a reason for this? Is it worth also showing the ECS or TCR from these models on this plot, so that one could see that higher ECS/TCR models tend to be down-weighted? I assume this is correct, to the extent that models that warm the most over the 21st Century have high ECS/TCR, but I don't recall the authors coming out and saying it. Modifying this figure in this way could be a compact way of making that point.

We had originally decided against showing independence weights to avoid the readers being overwhelmed by the figure (and because they could be inferred from the difference between combined weights and performance weights). Also, in the original figure we had shown the weights relative to the median weight, so that the distance of a model with, e.g., twice the equal weight would show at the same distance from '1' (equal weighting) as a model with ½ of the weight (see also your last minor comment). However, we realise that this might be slightly harder to interpret so we have changed it in the revised manuscript.

We now show normalised weights for all three cases: independence, performance, and combined. In addition we now indicate TCR by coloring the labels accordingly (Figure 4 in the revised manuscript) and we have added a table containing all values to the supplement (Table S2).

* Figure 4: I'm surprised to see several well-regarded models having relatively low performance weights (UKESM, HadGEM, CanESM, CESM), whereas some models that are typically poor performers seem to do well here (GISS, FGOALS, INM-CM). Any comment? Is it possible that your performance metrics are too restrictive (just involving tas and psl, two fields that may not adequately discriminate models with good vs bad moist physics that governs feedback and ECS), allowing poor performing models to get high weights?

The reviewer is right, several typically well-regarded models receive rather low weights in our scheme. However, we point out that most of the models mentioned as examples have very high TCR. Based on our analysis (and other studies, see, e.g., Tokarska et al., 2020, Nijsse et al., 2020) these very high warming models are less likely and therefore they are down-weighted. In some cases (UKESM, HadGEM, CanESM) the main reason is the obvious

mismatch between the observed and simulated warming over the course of the 20th century, which the modeling groups acknowledge in their technical description papers of the models.

It is indeed possible that our particular diagnostics choice leads to typically less well-regarded models receiving relatively high weights. This means that according to our chosen diagnostics they are performing well compared to other models. It is possible that we would need to include more or other diagnostics to downweight models which have, e.g., bad moist physics, since the weighting method does not include knowledge about specific parameterizations. This point highlights the importance of careful diagnostics choices and the fact that the weighting is always aimed at a particular target and diagnostics choice. The weighting is not supposed to tease out which model is best in every case, and depending on the target and diagnostics choice the models receiving the highest or lowest weights will be different. This does not mean models receiving low weights in this case are bad models in general, as the reviewer realized some low weight models in our case are well regarded models and considered good models in general. But it means that based on their performance in simulating historical warming trends they are considered less likely here.

Minor Comments

*line 61: should be “model’s”*line 78: should be “method’s”

Done.

*Line 250: I don’t see where the 10-20% statement comes from. By my eye, the medians range from near 0% to slightly larger than 25%.

The reviewer is correct, we changed this.

*Figure 1: titles should be “leave-one-out”

We changed the caption so this is no longer applicable.

*Figure 2 caption: should be “which”

Done.

*Figure 2: To clarify, the similarity between pseudo-obs and MME is only assessed over the “Diagnostic period” right? (Side-note: “diagnostic period” only appears in the figure and is not discussed in the text.) By my eye, MPI looks closer to the MME than does CanESM, so I’m a bit confused here. Is the reason because similarity in the evolution of GMST only one of the several metrics employed, and MPI does worse in the ones that cannot be gleaned from this figure?

We now introduce the terms diagnostic period in the main text of the revised manuscript (lines 215). Regarding the second point: the reviewer is correct in assuming that the performance of the models in the diagnostics that inform the weighting can not be inferred from figure 2 in general. We have added a sentence to the caption of figure 2 to make that clear.

*Line 309: “allows us” or “allows one”; also, it seems like some reference to all the performance metrics work done by Gleckler et al seems appropriate here. I believe they also advocate for comparing against multiple observational datasets.

This sentence does no longer exist but we have added a reference to Gleckler et al. (2008) in line 108 in the revised manuscript, where we motivate the usage of more than one observational dataset.

*Line 314: I don't see the motivation for these 3 groupings. Is it in any way objective?

This paragraph no longer exists in the revised manuscript.

*Figure 6: too small to read, suggest stacking the two panels vertically rather than placing them next to each other horizontally

Done.

*Line 334: should be "model's"

Done.

*I don't think the average reader should be expected to know how to interpret a figure like Figure 5. Only the meaning of the colors are explained in the caption. What does the rest signify?

We have added additional description to figure 5 and now provide a more detailed description of the clustering approach in the supplement (section S5 in the revised manuscript).

*Line 391 "The weighting also largely reconciles CMIP6 with 5": what is this referring to specifically, and is there a figure in particular being referenced?

We were referring to the fact that the constrained CMIP6 TCR is closer to the CMIP5 TCR range from, e.g., the IPCC AR5 (1°C-2.5°C). However, this sentence was slightly misplaced here and is no longer included in the revised manuscript.

*Figure 4: Are all weights less than or equal to 1 in absolute units, and only exceed when expressed relative to equal weighting as is done in the figure? Otherwise I'm a little confused about why a model would have a weight in excess of 1. How exactly is wi used? weighted avg of X = sum(wi*Xi)/sum(wi)?

We now show normalised weights for all three cases: independence, performance, and combined. See also our answer to your major point regarding figure 4 above.