Response to: suggestions for revision or reasons for rejection (will be published if the paper is accepted for final publication)

I really appreciate the responses to my comments and the major modifications made to the paper. I think the paper is more interesting now, with a very thorough investigation of issues associated with weighting, that goes beyond the simple incorporation of SMILES in ensemble projections (by the way, I'm not sure that the title is really optimal now). The paper will be suitable for publication after minor revisions.

We would like to thank you for your constructive and thought-provoking reviews both in this round and in the round before. They led to a lot of interesting discussions on our end and hopefully, a paper that will be of interest to a broader audience than before. We do plan to change the title to: An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. Thank you for the recommendation to change the title as well!

The line numbers refer to the version with highlighted changes in the author response.

L4 "SMILES represent internal variability…" The formulation seems strange to me.

Thank you for pointing this out. We've changed the sentence to read:

L3-4: "SMILEs allow for the quantification of internal variability…"

L124 You could say a word on these differences of physics (they are quite limited actually)

Absolutely, we've added the following description:

L124-127: "Additionally, for the GISS-E2-H and GISS-E2-R experiments, NOAA GISS provides members from 3 physics-version ("p") setups that differ in atmospheric composition (AC) and aerosol indirect effects (AIE) (Miller et al., 2014). We treat the 3 setups as follows: p1 (prescribed AC and AIE) and p3 (prognostic AC and partial AIE) members are treated as 2 member IC ensembles and the p2 member (prognostic AC and AIE) is treated as a single member representation (Table 1)."

Figure 1. What is shown for CMIP5? Statistics on all the members, with equal weighting, without taking into account that some models provide several IC members? It should be noted in the

legend as it is not an approach that one would normally used, with maybe a reference to the discussion in section 3.1. Spread: please say in the legend how the spread is calculated in practice. There are many ways to define a spread.

This is a great point; thanks for bringing it up. In line with this recommendation, we've decided to update Figure 1 to show ensemble spread as the 5th-95th percentiles of each distribution (rather than making any sort of gaussian assumption). We've changed the caption of Figure 1 to read:

"Observational estimates (OBS; gray), the CMIP5 ensemble (blue), and the three SMILEs: CESM1.2.2-LE (red), CanESM2-LE (yellow), and MPI-GE (green) evaluated in this study, shown in terms of area- and seasonally-averaged absolute surface air temperature timeseries (SAT; ˚C). The two OBS datasets, ERA-20C Temperature and the Berkeley Earth Surface Temperature (BEST) product, are shown in solid gray and dashed gray respectively. Their average, used to determine member performance, is shown in solid black. For the CMIP5 and three SMILEs, the ensemble means across members are shown in solid color; the shading indicates the 5th-95th percentile of each distribution as a measure of ensemble spread. Note that the CMIP5 ensemble is a multi-model, multi-initial condition member ensemble of 88 members from 40 (named) model setups, not the "one model, one vote" ensemble often used in multi-model ensemble studies. Panel a shows projections for Northern European Winter (DJF NEU) and panel b shows projections for Mediterranean summer (JJA MED) SAT. The number of members in each ensemble is indicated in parenthesis in the legend."

We've also changed the main text to read:

L152-153: "The CMIP5 ensemble and three SMILEs are shown in terms of their respective ensemble means and spreads (represented by the 5th-95th percentile of each distribution) in Figure 1,..."

L218. For wIII. I don't understand the 1/N2 factor. It should be 1/N. Is this just a mistake in the written equation (the title of the subsection is 1/N scaling after all) or are the analyses impacted?

We agree that our choice to write the equation with a $1/N^2$ is unnecessarily confusing. The $1/N^2$ comes from the combination of the 1/N scaling and the 1/N in the average formula applied to the IC member performance weights (such that each IC member receives an

identical performance weight). We've changed the formula wiIII and wiIV to more directly reflect the namesake 1/N scaling.

L236. Does each model get a unique performance weight as in section 3.3 or are different performance weights given to each IC member of a model? It is not clear.

This is an important methodological point. We've revised the paragraph to read:

L234-241: "Finally, the fifth weighting strategy operates under the assumption that independence cannot necessarily be determined by model name, but shared biases in simulating historical climate can give an idea of dependence that comes from differently named models sharing ideas and code. Instead of relying on knowledge of model origin, the RMSE weighting (wiV) initially proposed by Knutti et al. (2017) relies solely on model output to determine a model's overall weight. It features an independence scaling based on RMSE distance metrics in addition to the RMSE-derived performance weights. For results to be compatible with past assessments of this weighting scheme (e.g. Lorenz et al., 2018; Brunner et al., 2019), we assign each member their unique performance weight (as computed in wiII) even if they are IC ensemble members. This puts the RMSE weighting in contrast to the 1/N scaling approaches which ensure IC ensemble members have identical weights."

Line 378-379. It is not totally obvious to me: if a model is almost the truth, is almost perfectly realistic, and all the other models are wrong, why would it be problematic for this model to be over-represented? How do we decide in practice that a model is over-represented or not?

You are right to point out that in order to make the claim a model is over-represented, one would need to do something along the lines of out-of-sample testing. We've rephrased the section to reflect that the outsized contribution of the MPI-GE comes not from it being so much higher performing than other models, but rather from there being 100 high performing members present:

L381-389: "Uncertainty in the DJF NEU ALL ensemble is constrained both by the performance weight diminishing the contribution of CMIP5 members and because MPI is one of the highest performing models based on the chosen DJF predictors. The high performing MPI-GE receives 65.8% of the total ALL ensemble weight, though individual MPI-GE members only receive up to three times more weight than the averaged

assigned weight. The aggregate impact of 100 high performing members, however, is outsized and results in the narrowing of the performance weighted end-of-century warming distribution. The narrowing does not reflect the increased certainty that comes from the agreement of independent entities within the ensemble. Instead, it exemplifies that there is a need for an independence assumption in order to avoid the outsize influence that comes from being both historically realistic and numerously represented in the ensemble."

Line 386-387. I'm not sure I understand that reasoning. The spread could be the same just coincidentally.

It's true they could be the same coincidentally. We hoped to convey that the equally weighted distribution was too narrow because of the majority influence from just three models. Because the performance distribution was as narrow as the equally weighted distribution, we extrapolated that it was likely somewhat over-confident as well. The line of thinking is not crucial to the overall message of the paper, though, and so we've revised the section to read:

L390-394: "For JJA MED SAT change, the performance weight reduces the contribution of the three SMILEs to the ALL distribution in comparison to the equal weighting case, with the largest reduction made to CanESM2-LE contribution (17.4% to 7.4%; Fig.3d). However, the three SMILEs (three independent entities) still receive 51% of the total JJA MED ALL ensemble weight, their contributions again augmented by numerous representations. As in the equal weighting case, the JJA MED ALL performance-weighted ensemble mean is still modestly shifted towards more end-of-century warming than its JJA MED CMIP5 counterpart. This reflects the above CMIP5-average SAT change of the CESM1.2.2-LE and the CanESM2-LE in Mediterranean summer."

L389. "In light of the clear necessity": Not that clear to me.

Thanks for pointing out the need for clarification. We've changed the opening to:

L396: "In an effort to more appropriately handle the mix of models and IC members present in the ALL ensemble,..."

L567: "A weighting scheme, such as the one assessed here, is thus ideal for providing justifiable estimates of uncertainty…" Which weighting scheme? Several weighting schemes are discussed, and for me it is not obvious which one is the best among the last three. This paper actually shows that different weighting schemes that make sense can lead to different results in terms of spread and even response. It is therefore still very difficult to provide justifiable "estimates of uncertainty". And as discussed in Appendix B some ad-hoc choices have to be made: values of sigma d and s etc. I really appreciate the different tests used by the authors to define these parameters, they did a very good job, but some subjectivity remains.

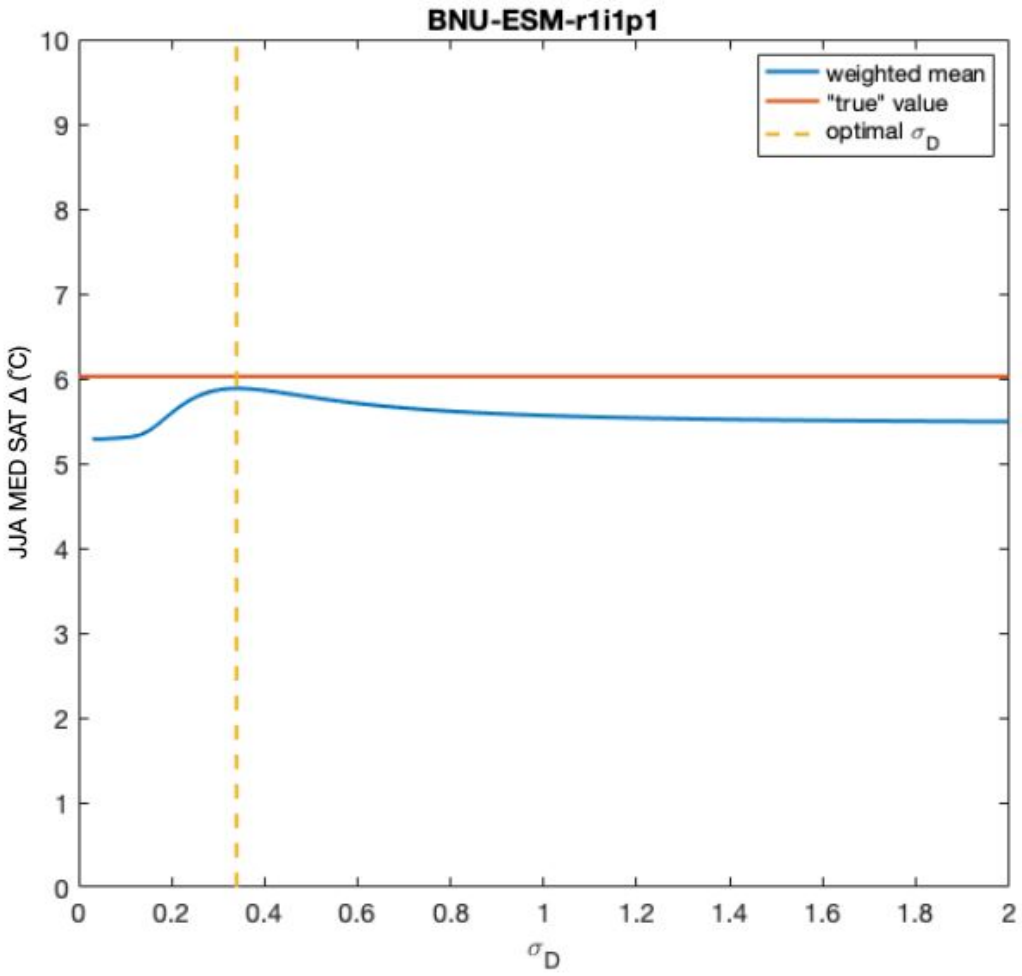
You bring up a really important point that the "justifiable estimate of uncertainty" is still very much a matter of debate. Our closing didn't really reflect that and further, it didn't highlight the main finding (reconciling that the RMSE scaling in its first conception wasn't able to distinguish between IC members and models). In line with this, we've revised the last paragraph to read:

L574-579: "For more conventional multi-model ensembles that may include just a few initial condition ensemble members amongst the models, results may be less sensitive to choices underpinning the independence scaling. When large ensembles are included, however, it becomes clear that an independence scaling, such as the RMSE global predictor scaling presented here, that both scales known dependencies appropriately (i.e., 1/N for IC ensemble members) and assigns a degree of independence to remaining members is necessary. Such an independence scaling will be a useful tool with which to assess uncertainty in the combined multi-model, multi-initial condition ensemble member CMIP6 ensemble."

L625. I'm not sure to understand exactly how it is done. How the prediction intervals corresponding to each truth are calculated in practice?

Thank you for bringing this point up, we were much too brief in our description of the perfect model test. Because the performance weighting was not the primary consideration of this study, we compared a simplified perfect model test to the perfect model test used in previous studies. The two versions are similar (using each model as "truth", then predicting that "truth" using the other models) and are consistent in the σD they give. We hope you find the updated description (and example plot) of the simplified method to be more helpful:

L628-645: “Determining the shape parameters σD and σS is an important step in the RMSE weighting process (Knutti et al. 2017). σD can be set using a perfect model test, as described in Lorenz et al. (2018). Here, a simplified perfect model test is performed on an 47 member ensemble, which includes only the first initial condition member from the SMILEs and each of the CMIP5 models ensembles (40 named models with an additional 4 members from GISS-E2-R and GISS-E2-H physics physics-version ensembles). This is done because having multiple IC members (or a SMILE) in the ensemble could bias the perfect model test, which is based on predicting one member using a weighted distribution of the rest. We use member 1 for each initial condition ensemble because, often, when multiple initial condition members are available, the first member is selected (e.g. Liu et al., 2012; Karlsson and Svensson, 2013; Sillmann et al., 2013). During the perfect model test, each member is assumed to be the "truth" once and a weighting is performed using the remaining members to predict the "true" SAT change. RMSE distances (based on nine predictors) are computed with respect to the truth for the remaining members and used in the performance weighting function (wiII) described in section 3.2. The performance weights are computed for σD values ranging between 0 and 2 (on 0.01 intervals). For each σD, the weighted mean SAT change is computed and compared to the “true” SAT change. The optimal σD for each truth is chosen to be where the difference between the weighted mean SAT change and the true SAT change is minimized. In the few cases when the weighted mean exhibits asymptotic behavior with no clear minimum difference prior to σD = 2, the σD value is selected at the point where the leveling off begins (as determined by the intersection between a threshold value and the weighted mean curve). For the nine predictor RMSE weightings, we set σD values to the mean of the 47 optimal σD values computed during the perfect model test. It is important to note that this choice is ultimately subjective and further parameter sensitivity testing is recommended in studies focused on model performance.”



For each truth, do you calculate the weighted mean and weighted standard deviation (for each value of sigma), which you use to compute the 10-90% prediction interval supposing a Gaussian distribution, and then check whether the truth is within this interval?

Good question. In the more nuanced version of the perfect model test, the prediction interval does not assume a gaussian distribution. The 10 and 90th percentiles are computed as :

For x 1 …. x i and weights w 1 …. w i ,
W is the sum of all weights and s j is the sum of the first j weights.

For the probability p, if pW falls
(1) between s j and s j +1 , the quantile is estimated at x j+1
(2) on s j , the quantile is estimated at ½ (x j + x j +1)

Further details are given here:
https://www.statsmodels.org/dev/generated/statsmodels.stats.weightsta
ts.DescrStatsW.quantile.html#statsmodels.stats.weightstats.DescrStats
W.quantile
https://support.sas.com/documentation/cdl/en/procstat/

I suppose that the basic "equal weighting" approach pass the test, right? Therefore, how do you decide that your weighting scheme is better than the standard democracy? You say that you choose the smallest value of sigma that passes the test, but why the smallest value of sigma should be used? Why is it preferable?

For values of sigma larger than the distances between models and observations (in our case, for sigma values larger than about 1.3), the weighting tends towards a standard democracy. We can often see in the perfect model test that a stronger-than-equal weighting gets you closer to the "true" warming than an equal weighting does, suggesting that down-weighting some models (by setting sigma, the gaussian width, to a value within distribution of distances) does have some added value.

In terms of "better than the standard democracy", we find it helpful to think of the weighting as a way to assign meaning to the distribution. In the multi-model, multi-initial condition ensemble used in this study, the equal weighting approach is misleading; the SMILEs have more weight than other models because they are numerously represented. The "one model, one vote" democracy is simply a binary weighting where much of the available information arbitrarily gets assigned zero weight. The weighting we explore allows us to explicitly state that models that are significantly biased with respect to observations are unlikely to effectively simulate future regional climate and that models that are effectively duplicates of one another don't get the opportunity to "stuff the ballot box".

L657: "… and set sigma s at two standard deviations below the SMILE Sij mean value". I understand the logic, but why this specific choice of two standard deviation below Sij mean value?

This was an attempt to offer guidance on how to best avoid overconfident weighting in the case of a predictor set with too much internal variability to distinguish the IC members from the models. From the sensitivity analysis, it was evident that the sweet spot for σS was somewhere between 0.2 and 0.3 which placed it in the lower tail of the SMILE intermember distance distributions. However, we've emphasized that rather than deal with the scenario where the distribution is sensitive to the shape parameter, it's preferable to instead choose predictors that can distinguish IC members from models:

L679-684: "Another more robust option, as discussed in the main text, is to select a set of independence predictors that explicitly differentiate inter-IC member distances from inter-model distances. In this case, σS should not be set to two standard deviations below the SMILE Sij mean, rather it should be set to a value greater than all IC member Sij but less than inter-model Sij (particularly differently named models). For the large-scale CLIM predictor set explored in Figure 4, σS can be computed based on initial condition member intermember distances as described in Brunner et al. 2019; σS in this instance is 0.22."