Response to “Review of "A weighting scheme to incorporate large ensembles in multi-model ensemble projections" by Merrifield et al.”

In this paper, the authors describe the extension of a weighting scheme for multi-model climate projections described in previous works to incorporate single model initial condition large ensembles (SMILE). This weighting scheme uses a performance metric, based on the similarity of a simulation with observations and an independence metric, based on the similarity between simulations. Several properties of two variables (surface air temperature and sea level pressure) in the present climate are used to measure similarity. The authors intend to demonstrate the applicability and the usefulness of this weighting scheme with SMILEs, focusing on surface air temperature change over Northern Europe and the Mediterranean. They also discuss different properties of the weights and some practical issues that may arise in such applications.

The subject of the paper is interesting and important, and there are some interesting analyses in this paper. It is well written and generally easy to follow. But I also think that the use of the proposed weighting scheme with SMILEs raises fundamental questions that are not addressed. As the incorporation of SMILEs in the weighting scheme is the novelty of the paper compared to previous works, these issues must be properly dealt with before the publication of the paper could be considered.

I am not sure that the authors can address these issues properly, as they are really intrinsic to the chosen approach, but I want to give them the opportunity to prove me wrong. I therefore recommend major revisions to the paper, but I may still recommend rejection of the paper at the next round.

Thank you for your comprehensive and thoughtful review of our paper; We really appreciate you taking the time to interrogate the underlying aspects of the weighting method, as your review has generated a lot of interesting discussion and new ideas for a path forward. We hope to be able to assuage some of your concerns in this response and in the subsequent revision of the manuscript.

Major comments

The notion of "independence" is perfectly defined in statistics and probability theory, but it is very ill defined when applied to climate models (which is not really acknowledged and discussed by the authors). In this paper, as in previous works, two models are considered more or less independent depending on the similarity of their results. Two models are considered "weakly" independent if their results are very similar and "strongly" independent if their results are very different. This is a hypothesis, and it should be discussed.

We agree that we should include a more comprehensive discussion about our hypothesis of independence as a measure of whether a model or simulation provides additional information by having a

distinguishable representation of historical climate (which is indeed not the formal statistical definition of independence). The reasoning behind defining independence as the RMSE distance between models, based on a collection of historical features, is to identify and reduce the influence of shared model biases on the uncertainty distribution of interest. An example is temperature mean state: models that share a warm bias in the Mediterannean may have dried into a land-atmosphere feedback regime different from what has been observed, which then amplifies warming. Additionally, the choice to use RMSE distance allows a measure of independence that doesn't rely on *a priori* knowledge of code sharing, branched development, etc., which, to some extent, renders models with different names not independent as well.

Following this suggestion, we have added a discussion about independence assumptions to the paper to accompany the new analysis on how different independence assumptions affect the weighting (new versions of Figure 3 and 4).

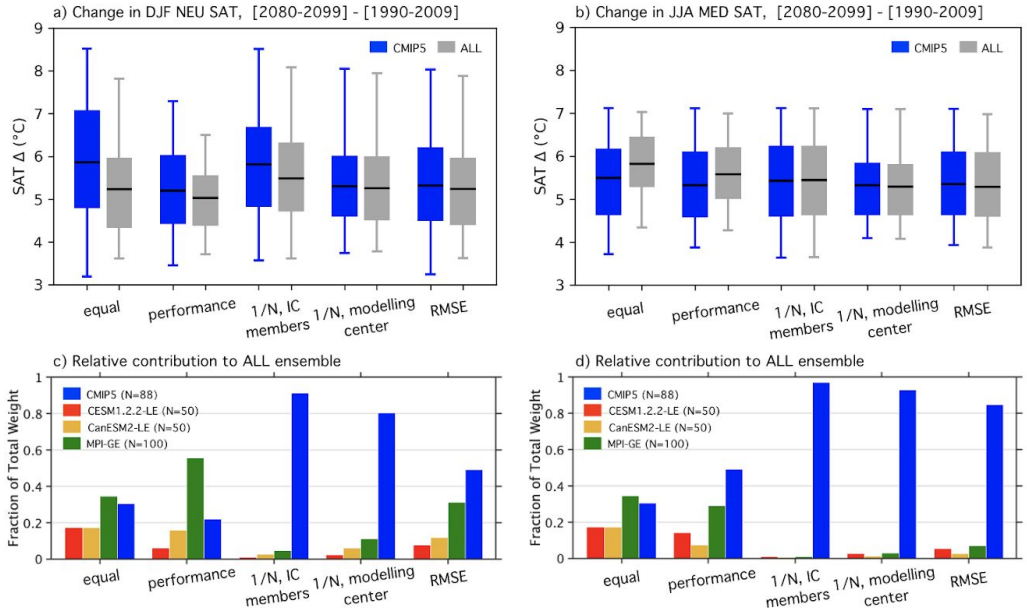


Figure 3. (a) Box-and-whisker plot showing how the five weighting strategies effect the distributions of DJF NEU SAT change (Δ, [2080-2099]-[1990-2009]) for the CMIP5 ensemble (blue) and ALL ensemble (CMIP5 with the 3 SMILEs; gray). The box element spans the 25th to 75th percentile of the distribution; mean SAT change is indicated by the horizontal line within the box. The whisker element spans the 5th to 95th percentile. b) As in a), but for JJA MED SAT change. c) The contribution of SMILE and CMIP5 members to the DJF NEU ALL ensemble under different weighting strategies, in terms of fraction of total weight. d) As in c), but for the JJA MED ALL ensemble.

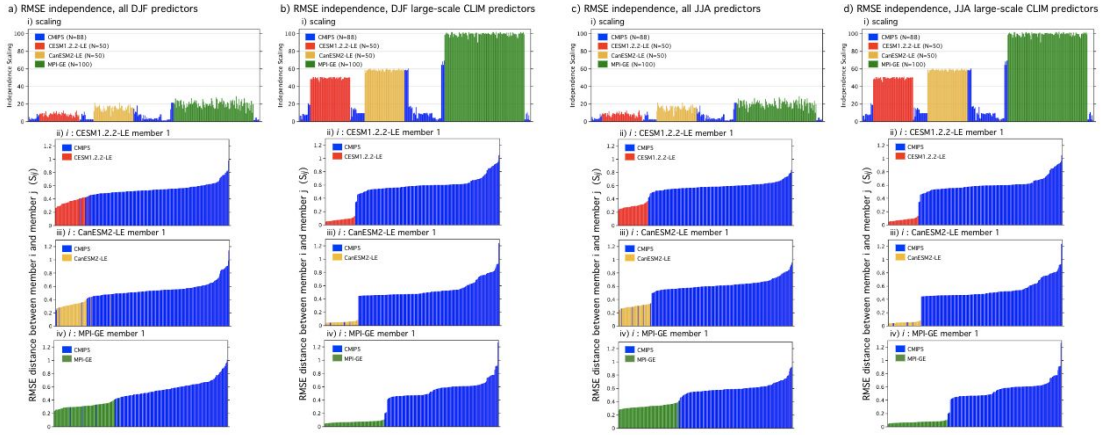


Figure 4. (a-di) The RMSE independence scaling of the SMILEs and CMIP5 ensemble members, shown in the order listed in Table 1. Panels ai and ci show the scaling computed from the 9 predictors used in the original DJF and JJA RMSE distance weighting respectively. Panels bi and di show the scaling computed from 2 predictors: global land SAT and European sector SLP climatology over the 1950-2010 period for DJF and JJA respectively. (a-d ii) The sorted RMSE distance between member 1 of the CESM1.2.2-LE and all other members of the ALL ensemble. (a-d iii) As in ii, but for CanESM2-LE member 1. (a-d iv) As in ii and iii, but for MPI-GE member 1.

The results of two "independent models" cannot be similar? Two independent models cannot converge towards the truth (if the models are close to the truth they will also be close from each other)? Overall, to my opinion, this hypothesis can make sense when dealing with multiple different models, and in any case there is no perfect theoretical and practical way to characterize model independence.

While it's not (yet) explicitly emphasized in the paper and is a bit of a subtle point, members are not penalized for matching results. In this case, that would be reducing the weight of members that warm the same amount by end-of-century. Two models that independently simulate climate (in so much as you can independently simulate the same governing equations, parameterized physics, etc.) that then warm the same amount suggests robustness/increased certainty in the outcome. Those two models may match historically in one predictor (particularly if they are tuned towards observations in a region), but are unlikely to match over nine predictors. This is demonstrated by the fact that even initial condition ensemble members don't match over the nine predictors we use to determine the independence scaling (which we will also subsequently address by finding predictors optimal for determining independence).

But I'm really bothered with this approach when dealing with members from the same model (only differing by initial conditions). I think that the attempt to use this weighting scheme with SMILEs illustrates some difficulties of the definition of independence in terms of similarity. The members of a SMILE are independent in the statistical sense of the term, the only sense of independence that is well defined. But they are not independent in the approach proposed by the authors, and they can be more or less "independent" according the similarity of their results. For me, it is very problematic. If you roll a dice two times, you don't decide that two outcomes are "more independent" if you get a 4 and a 6 than if you get two 3. But it is basically what is done in the proposed method with SMILEs.

This gets at the complexity of the interplay between uncertainty associated with internal variability and uncertainty associated with model differences and whether or not both should be represented in uncertainty estimates. In this study, we do not enforce skill as a property of a model because on the continuum of climate model independence, it's not entirely clear what the cut off should be to define a model. It could be based solely on name (i.e. all initial condition ensemble members have the same skill), or modelling agency (i.e. runs from an institution with different resolutions, vegetation nodes etc. have the same skill), or even grouping models with shared components (i.e. all models with adapted CESM components). The hope

was that by treating assignment of weights as a continuum as well would naturally group the weights of large ensemble members.

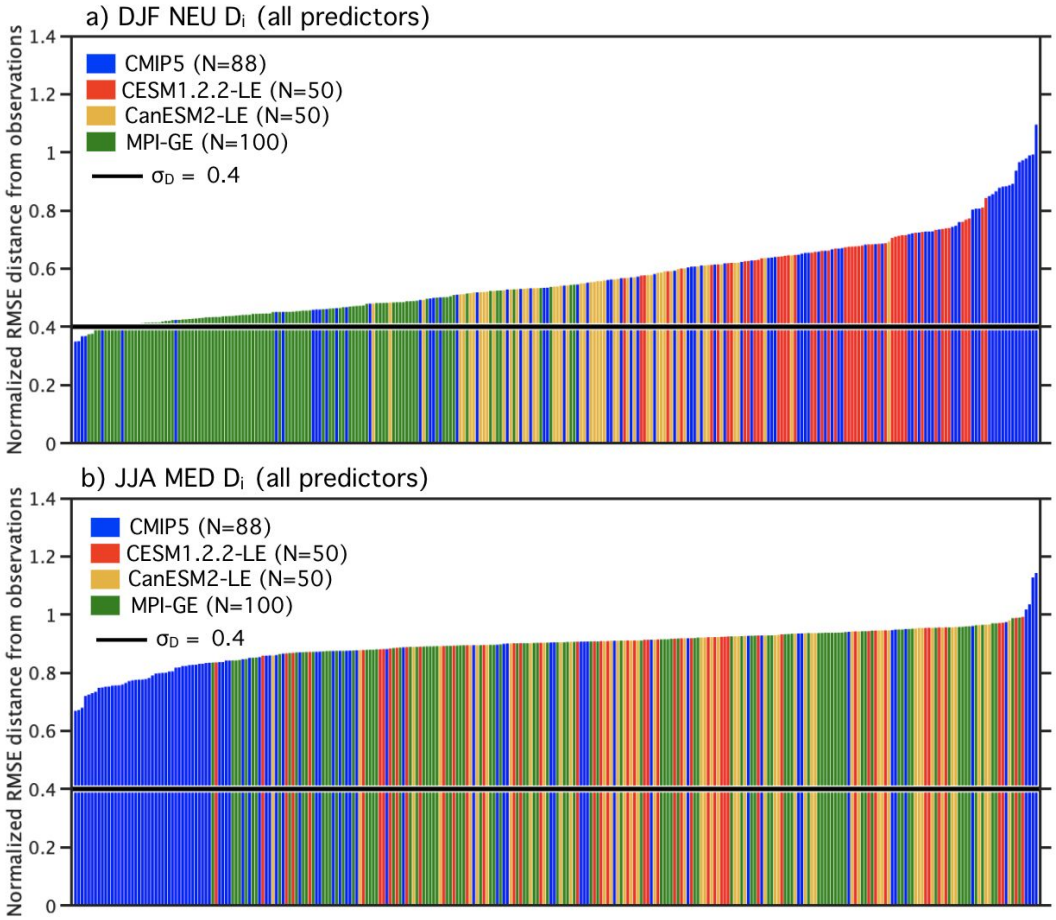


Figure B1. RMSE distance D_i, derived from 9 predictors, between observations and the 288 members of the ALL ensemble (CMIP5 (blue) + CESM1.2.2 (red) + CanESM2 (yellow) + MPI-GE (green)). DJF NEU distances are shown in panel a and JJA MED distances are shown in panel b.

However, that is not what occurred in the current version of the method and we felt it important to be transparent about that. We feel that basing the weighting on multiple predictors, including standard deviation and trend predictors, is important to ward against overconfidence in the performance weight. We want performance to be representative of a member's ability to holistically represent aspects of climate relevant to the target change, rather than just one aspect (like climatology) which would result in strong downweighting of models that likely give valid estimates of future change. The drawback of this choice is that it introduces spread into the RMSE distances of SMILE members (new Figure B1; below) and therefore distinguishes them in terms of weight.

We are working on updates to the method to address this issue; mainly in that we plan to handle predictors differently for performance weights and independence scalings. Steps are being taken to ensure that the perfect model test and independence scalings remain consistent and that we use the information we have about independence i.e., that initial condition members are dependent entities. But most importantly, we have added new analysis to compare the weighting when different independence assumptions are made in Figure 3 and have found a set of predictors on which to base the RMSE independence scaling that differentiates between models and initial condition ensemble members.

As an illustration of this issue: Imagine the particular case where we only have a single SMILE, and that we are interested by the distribution. Using the weighting scheme described in this paper is not correct in this case, right? We know that the SMILES members are independent and that each member should receive the same weight. It is what is done is all the studies based on a single SMILE. But the weighting scheme described in the paper would give different weights to different members. I think that the weighting scheme proposed by the authors (any weighting scheme) should hold seamlessly in a particular case like this one.

The method can be adapted to hold seamlessly in this case by setting the performance and independence shape parameters to be larger than the RMSE distance to observations and between members such that all realizations of the SMILE receive approximately equal weight.

-Giving different performance weights to different members of the same climate model is also problematic, at a fundamental level, I think. The skill is intrinsic to the model, and not specific to a member of the model (once the memory due to initial conditions has disappeared). Whether a particular member of a SMILE is closer to the observations than another is purely accidental and says absolutely nothing on the realism of this particular member in the future climate. The baseline approach to which the weighting scheme is compared in this paper consists in giving an equal weight to all the members of the multi-member, multi-model ensemble (independently of the existence of other members of the same model in the ensemble). Obviously, it is a very bad approach, and nobody would do that, I think.

This was initially our intuition as well, that members of the same model would be indistinguishable from one another and be distinguishable from other models. However, this is not how the RMSE metric works in practice in the current version of the method, and we

feel that it's important to document that this is the case. Internal variability in this instance is large enough such that SMILE members are not necessarily closer to one another than "independent" models are to each other. There are several ways this will be handled going forward; in the new analysis of continuum scaling assumptions, there are two cases where we impose the "model skill, not member skill" assumption. In the first case, we presume initial condition (IC) ensemble members are dependent and we impose a mean performance weight (average of the performance weight of the IC members), scaled by the number of IC members N, to each IC member. Each member of the MPI-GE, for example, receives the same weight, (mean performance weight)/100. In the second case, we extend the same assumption to all output from a model development/modelling center.

While there isn't a settled, optimal way to handle "ensembles of opportunity" like CMIP and many groups are providing large ensembles to CMIP6, we also hope that no one would assign equal weight to all available information. We investigate the implications of doing so in the new version of Figure 3 as well.

If we consider the models as independent and equally skilful, SMILE members can be easily incorporated in a multi-model ensemble, as it has been done for years, by giving a weight to each member of a given model inversely proportional to the number of members of this model in the full ensemble. This approach is perfectly justified from a statistical standpoint (within the hypotheses made). (i) I think that the authors should use this approach as a baseline, to which they can compare their weighting scheme, and show the results obtained with this approach for example in Figure 3. (ii) Logically, the weights of an appropriate weighting scheme should tend towards the ones described above when the "hypothesis" of inter-model dependence and unequal realism is relaxed, I think. It is not the case with the weighting scheme described in the paper.

This is a great recommendation that we have incorporated into a new version of Figure 3. A question arose of how to decide on a performance weight of a SMILE, given how different members weight the SMILE when they serve as the sole representative of it (original Figure 5). Further, this approach requires an *a priori* understanding of what is considered a "model". Therefore, we've decided to compare different weighting approaches for both CMIP5 and the CMIP5-ALL ensemble:

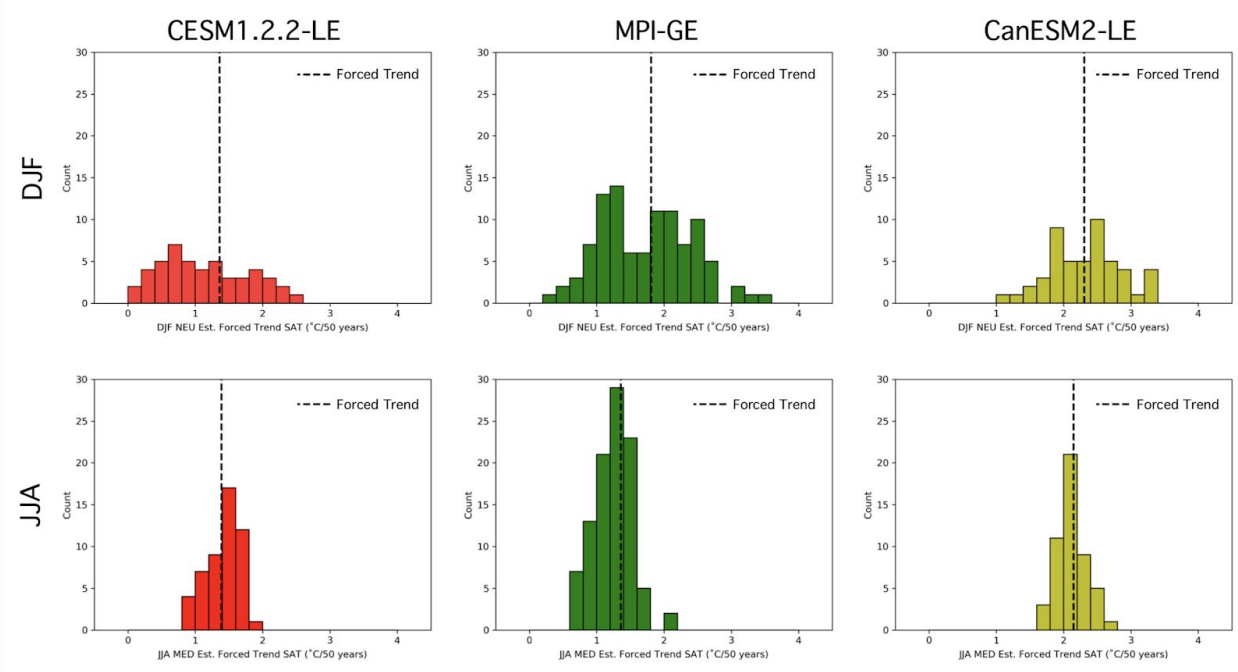
1. **Equal weighting;** assumption that all members are independent. Each member receives a weight of 1.

2. **Performance weighting**; assumption that all members are independent, but some are more realistic than others. Performance weight computed as RMSE distance from observations.
3. **Initial condition ensembles receive a single weight**; assumption that initial condition members are dependent. Weights are comprised of the average of the performance weights of the ensemble (as computed by RMSE) scaled by N (the number of ensemble members). Initial condition members all receive the same weight, reflecting that they all represent equally likely outcomes.
4. **"Models" receive a single weight**; assumption that all members provided by a modelling center or in a development stream are dependent. Weights are comprised of the average of the performance weights of all the members provided by a modelling center or in a known development chain (as computed by RMSE), scaled by N (the number of members). Modelling center contributions all receive the same weight, reflecting that they are one model.
5. **Current version of the weighting**; assumption that independence cannot necessarily be determined by model name, but shared biases in simulating historical climate can give an idea of dependence. Each member receives its own performance and independence weight, based on RMSE distance, regardless of what model it came from.

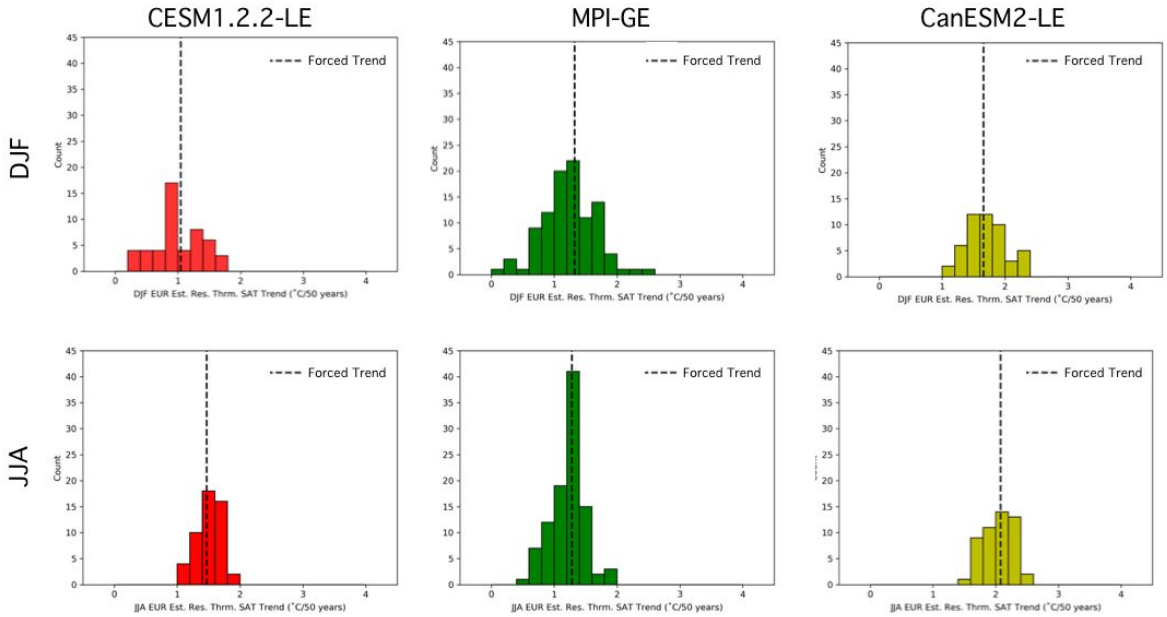
-I disagree with the interpretation of the results of dynamical adjustment in the paper. It is not possible to extract the "forced trend", even the "estimated forced trend" or the "radiatively-forced trend" with dynamical adjustment. Dynamical adjustment only allows separating the part of the trend that is due to large-scale atmospheric circulation from the part of the trend that is not due to large-scale atmospheric circulation. The "part of the trend that is due to atmospheric circulation" is not a correct estimation of the impact of internal variability, except in some particular cases. The variations in atmospheric circulation indeed can be forced, they are not necessarily of internal origin. There are quite a few papers on the detection and attribution of anthropogenic influences on large-scale atmospheric circulation, and there is a clear forced component (in the real sense of the term) in future circulation changes in many models. For this reason, the "part of the trend that is not due to large-scale atmospheric circulation" should not be named "forced trend", even "estimated forced trend". Additionally, it can bear the imprint of internal oceanic dynamics. It is mainly a vocabulary issue here, as the interpretation of the results of dynamical adjustment does no really matter for the results discussed in the paper. Still, it is important to be correct.

We see how the language we've used around dynamical adjustment is too simplistic, thank you for bringing it to our attention. In terms of

the decision to use the terminology "estimated forced trend", we find that for the trends in regional surface air temperature we estimate using dynamical adjustment are distributed around the forced trend computed from the ensemble mean (though it is true that they distribute further from this "true" value in Northern European Winter than in Mediterranean Summer as shown below).



This follows from previous studies that find that internal atmospheric circulation variability is closely associated with midlatitude SAT variability and that the internally generated component of SAT trends are largely induced by dynamics. Once this component is removed, the remaining trend can be described as thermodynamically controlled (either via radiative effects or indirectly through surface influences). To address this language concern, though, we will change "estimated forced trend" to "estimated residual thermodynamic trend" throughout to be more consistent with the dynamical adjustment literature. We will also use a larger predictor domain, Europe, in both seasons rather than the respective NEU and MED SREX regions. The European estimated residual thermodynamic trends are similarly distributed around the forced trend in summer and more narrowly distributed in winter than their SREX counterparts (below, note ordinate value differences).



In regards to potential forced circulation change playing a role in our estimates, we select analogues from the historical period over which there is no significant trend in SLP (Deser et al. 2012,2016). The temperatures fields that correspond to the SLP analogues and are used to construct the circulation-driven component of SAT for each month are detrended with a quadratic high pass filter as in Deser et al. 2016. This step is necessary, since otherwise months picked from the end of the record will contribute higher SAT anomalies simply because of the anthropogenically forced warmer background climate, even if the SLP patterns are the same (Lehner et al. 2017). We have improved the description in Appendix A to clarify some of these subtleties

Minor comments

l37. Parameterized processes are not the only reason for model uncertainty, I think. The dynamical cores can also be important in that context.

This is true; following this comment, we have change the sentence to read:

"Model uncertainty accounts for differences in how models simulate climate, from how the equations governing flow in the atmosphere are numerically solved to how processes in the climate system that are not otherwise captured on the spatial and temporal resolution of global climate models are parameterized."

l53. As said in the major comments, dynamical adjustment cannot be used to quantify the impact of internal variability on climate variables. It can only be used to estimate the part of variability that is not driven by large-scale atmospheric circulation. It is completely different.

We see how the language we used was imprecise. While internal variability in surface air temperature does, in part, arise from internal atmospheric variability, we have now made it clear that dynamics are not the sole driver of internal SAT variability:

"Internal variability manifests itself in climate variables, such as regional surface air temperature (SAT), through a complex set of controlling influences, chief among them being variability in the attendant atmospheric circulation (Wallace et al., 1995, 2015; Branstator and Teng, 2017). The influence of internal atmospheric variability on SAT can be quantified and accounted in projections of future climate using dynamical adjustment methods (e.g. Deser et al., 2016; Sippel et al., 2019)"

l55-56. You mean single "model" initial condition large ensemble and not single "member", right?

Thank you, corrected.

l85, data section I think it would be more logical to introduce the climate simulations (e.g. Table 1 etc.) before describing their result (Figure 1 etc.)

Following this suggestion, we have changed the order of the data section to introduce Table 1 prior to discussing Figure 1.

l96. ERA20C should not be used as observational reference for temperature. Only SLP and winds are assimilated in ERA20C, which leads to a sub-optimal representation of temperature variability. Not surprisingly, issues in regional temperature trends and low frequency variation exist in ERA20C. There are much better datasets to use for temperature. There is no need to use SLP and TAS from the same dataset to "assure consistency". Use the best dataset for each variable: normally, good observations from different sources are consistent. I also think that multiple observational datasets should be used in order to assess the impact of observational uncertainties.

The initial choice to use ERA-20C was made due the length of record and the ease of dynamical adjustment (a method that does in general require physical consistency SLP and SAT that is largely present within model-based reanalysis products). We agree, however, that ERA-20C is certainly not the ideal observational estimate and have thus adapted the method to base performance on ERA-20C SAT and SLP alongside Berkeley Earth Surface Temperature (BEST) and NOAA-20C SLP reanalysis version 3 to better represent observational uncertainty as recommended.

l144. "that adds independent information". What is meant exactly by "independent information" (or "new" information, at some places)? It should be discussed, from a theoretical point of view.

As well as improving the discussion of independence assumptions throughout, we have revised this section to reflect the assumption made by the RMSE weighting that "distinguishable" information (in terms of RMSE distance from other members) contributes to the ensemble distribution. The concept will no longer be framed as independent or "new" information.

l176. What "fit for purpose" means obviously depends on the purpose. I think it would be useful to state the purpose very precisely at this point (even if it can be inferred from other parts of the paper).

We have revised the section to read:

"Both the performance weight used in weighting strategies 2-5 and the independence scaling used in strategy 5 are based on a chosen definition of climate. A model's performance is based on its ability to reproduce observed climate and a member's independence is based on how much its climate differs from the climate in other members. When defining climate, the aim is to optimize the "fitness for purpose", which should include choosing predictors that are relevant to realistically simulating the target and will indicate if a model is biased in a way that is a cause for concern. For example, in Knutti et al. (2017), aspects of climate relevant for September sea ice extent, such as the climatological mean and trend in hemispheric mean September Arctic sea ice extent, gridded climatological mean and standard deviation in SAT for each month, were chosen. The chosen predictors reflected that if models that had either almost no sea ice in the present day or significantly more sea ice in the future than presently observed, they were less suitable for the task of projecting changes in sea ice extent. It is

also good practice to avoid using a single predictor to define climate to avoid an over-confident uncertainty estimate. No one model property can comprehensively reflect if the model is "good" for a particular purpose, and it is dangerous to constrain uncertainty by dismissing models that don't match observations for a particular statistical definition for ones that happen to be tuned to match that definition. Lorenz et al. (2018) discusses a more holistic strategy for choosing predictors and ultimately selected from a set of 24 predictors deemed relevant for projecting North American maximum temperature, based on known physical relationships, predictor-target correlations, and variance inflation considerations

Here "fitness for purpose" is a relatively simple and straight-forward definition of climate within which the sensitivity of the weighting scheme can be interrogated. We base the performance weighting and the RMSE independence scaling on 9 predictors: the climatology and interannual variability (represented by standard deviation) of SAT and SLP during the periods of 1950-1969 and 1990-2009 and a 50-year derived SAT trend (estimated residual thermodynamic trend; described in more detail in subsequent paragraphs) for the period of 1960-2009. We chose predictors to be aspects of regional temperature and pressure in a domain that encompasses modes of atmospheric circulation variability relevant to European climate, because they are (1) physically associated with the target (end-of-century warming) and (2) fields that may reflect model biases which would affect realistic simulation of future climate. For example, a model with a warmer-than-observed mean state in the Mediterannean may experience an enhanced land-atmosphere feedback mechanism that amplifies drying and warming of the region (e.g. Christensen and Boberg, 2012; Mueller and Seneviratne, 2014; Vogel et al., 2018). SAT and SLP have also been found to be highly relevant predictors by earlier studies (Brunner et al., 2019) and are among the most comprehensively measured atmospheric fields prior to the satellite era (Trenberth and Paolino, 1980). In terms of spatial domain, SAT climatology and variability predictors are computed over their corresponding ocean-masked SREX regions (i.e. NEU for DJF and MED for JJA) and SLP climatology and variability predictors are computed over a larger European sector domain which includes the North Atlantic (25–90˚N and 60˚W–100˚E). The derived SAT trend, estimated residual thermodynamic trend, is computed over the ocean-masked continental European domain (EUR; 30–76.25˚N and 10˚W–39˚E)."

l185. I don't really understand how the RMSE distances are computed. You say that they are computed at each point before area averaging. RMSEs are not computed over space but time? How do you compute the RMSE for a climatology at a given point? Please give the equations, it will be clearer.

Thank you for pointing this out, we had misworded the sentence about RMSE distance computation. We have revised the section to read:

"To compute the aggregate distance metrics from 9 predictors, all predictor and observational fields are bilinearly interpolated to a shared 2.5˚x 2.5˚ latitude-longitude grid. The predictors are then time-aggregated, with the mean or standard deviation computed over the periods 1950-1969 and 1990-2009, and the estimated residual thermodynamic trend computed over the period 1960-2009. For each time-aggregated predictor, the differences between the observed mean value and member value (or member value and member value) are computed at each grid point and subsequently squared. The squared differences are then area-averaged over the predictor domain and square-rooted to obtain an RMSE distance for observed-member and member-member pairs. For each predictor, the resulting distributions of observed-member and member-member RMSEs are then normalized by their mid-range value ([maximum + minimum]/2), such that the distance for each of the nine predictors are on the same order of magnitude and can be combined into a single Di (Figure B1) andSij (Figure B2) for each member."

l192-194. It is a reasonable idea when you consider two different models, but not when you consider two members of the same models. And in this paper two members of the same model are dealt with in the same way as two different models.

We agree and have revised the paper accordingly to reflect the "model skill, not member skill" issues with the previous draft. In particular, the new Figure 4 and 5 reflects a set of predictor choices that reconciles the RMSE-based weighting with an understanding of known dependent information.

l200. There is no i (and ii) in Figure 2a. Please add the complete numbering of the sub-figures.

We have moved the i and ii labels from within the panel to the panel titles for clarity.

l207-213. The fact that the "estimated" forced trends are so different between members of the same model clearly shows that one should not talk of forced trends for the results of dynamical adjustment, preceded or not by "estimated". But I agree that independently of its name, it can be an interesting performance metric.

In line with this suggestion, we have revised "estimated forced trend" to "estimated residual thermodynamic trend" throughout.

l214. "Internal variability": no, not necessarily (see major comments).

We have revised this to read: "The influence of large-scale atmospheric circulation contributes to the amplification of the observed.."

l228. Can you clarify what is meant by ""fair""?

We have revised the idea to: "Because estimated residual thermodynamic SAT trends in the broader European region are more comparable between members and observations due to the removal of an estimate of the influence of atmospheric variability that manifests on multi-decadal time-scales,..."

l235 and Figure 3. I'm missing something: I don't understand how the weighted distributions (box-and-whiskers plots) are obtained, based on the weighting scheme described in the paper. It is not directly straightforward I think. Is it a parametric distribution, using the weighted variances and means and a Gaussian hypothesis? It does not seem to be the case as the whiskers are not symmetrical. Please explain how the percentiles are computed when using the weighting scheme.

Thank you for bringing this up, we definitely will clarify this in the paper. The weighted mean is computed as:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

The weighted percentiles are calculated:

For $x_1 \ldots x_i$ and weights $w_1 \ldots w_i$,

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 information can be found at the following: https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.DescrStatsW.quantile.html#statsmodels.stats.weightstats.DescrStatsW.quantile

https://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_univariate_sect028.htm

l245. It would be interesting to add the results of the "classical" weighting scheme generally used when mixing SMILEs and multiple models (see major comments), that makes the hypothesis that the models are independent and equally skilful. It is a much better starting point for the comparison. Nobody in his right mind would add 200 members of the same climate model to the CMIP5 ensemble and compute the distribution without some basic weighting, right? l254-255. This is rather obvious: see the previous comment.

We also sincerely hope not. We have added this analysis to the paper in the revised version of Figure 3 and subsequent discussion.

l281-282. What criterion do you use to judge that the weighting is suitable? What is a suitable weighting scheme? It should be better discussed.

We hope you find the additional discussion associated with the new version of Figure 3 begins to better address the idea of suitability. We agree that the current discussion is not very convincing.

l285-320. I don't think that this analysis is that interesting. More important (and interesting) analyses are in Appendix.

We have replaced this analysis with an analysis of how using different predictors for independence scalings can bring the RMSE-based weighting closer to the "classical" weighting scheme you have described previously.

l456-457. I don't see a test of the sensitivity to "sigma s" in Figure B2. You mean Figure B3? Should you not describe Figure B2 first?

You are absolutely right, we have switched Figures B2 and B3.