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

This paper explores the usefulness of an established model weighting procedure (following Knutti et al. 2017, Sanderson et al. 2017) for incorporating large ensembles of single-model projections into multi-model projections. The model weighting method is shown to produce reasonable results three large ensembles ("SMILEs") are combined with an ensemble of CMIP5 model runs. The paper is generally well written and the results interesting. Some improvements explaining the methods would be useful, but overall I think pending minor revisions the paper should be suitable for publication.

Thank you for taking the time to review our manuscript, we really appreciate your feedback and are happy to hear that you are interested in some of the results. We hope you similarly find our new analysis associated with alternative independence assumptions interesting and that we are able to address in the revised version of the manuscript.

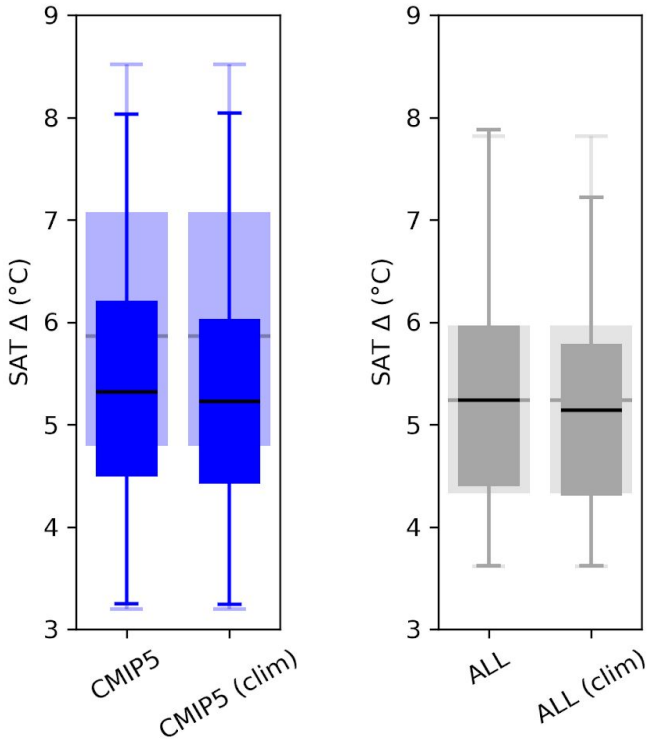
Main comments:

1. The selection of predictors is not completely convincing. I appreciate that the main purpose of the paper is to demonstrate that the weighting method is plausible for the type of ensemble considered, not to explore all possible choices of predictors. Nevertheless Appendix C shows that variability (SLP and SAT standard deviation) shows a weak past/future relation, and Fig 2a suggests weak or no past/future relation for DJF NEU SAT estimated forced trend. Are the results sensitive to exclusion of these predictors?

It is a good point that not all predictors we have chosen have strong emergent relationships, and we will add more discussion about predictor selection, both in the main text and in the appendix. We’ve revised the following in the main text:

“The assumption is that if a model accurately represents an aspect of historical climate, it is likely to realistically represent relevant physical processes and therefore is likely to provide a reliable future projection. If a model is significantly biased with respect to observed climate, its future representation of climate may be cause for concern (Knutti et al. 2017). For these tendencies to hold, a statistical relationship between the historical and future climate feature of interest must exist. In the absence of a strong relationship, predictors serve to add degrees of difference between members which helps to ward against overconfident weighting.”

The tasSTD predictor is considered to be one of the better predictors in terms of correlation with the end-of-century warming target (as in Lorenz et al. 2018) over both periods for the DJF NEU. For the JJA MED case, several members of CMIP5 have more-than-observed SAT and SLP variability, which features we wanted to be reflected in the performance weight as biases can indicate issues with physical processes (i.e., land-atmosphere interactions). Ultimately, though, if a predictor does not have a strong emergent relationship, its inclusion simply adds a bit of noise into the distances but doesn't strongly affect the CMIP5 weighting (below).



For the ALL ensemble, using only climatological predictors further narrows the distribution because the SMILEs tend to have near observed climatology and therefore a higher performance weight than other CMIP5 members. This illustrates why it tends not to be a good practice to use one or few predictors for performance with this method. We try to avoid situations of overconfidence in future results, as no single emergent relationship is really indicative of a model's performance.

However, climatology has proven to be a better indicator of dependence. We've added analysis of this in new versions of Figure 4 and 5.

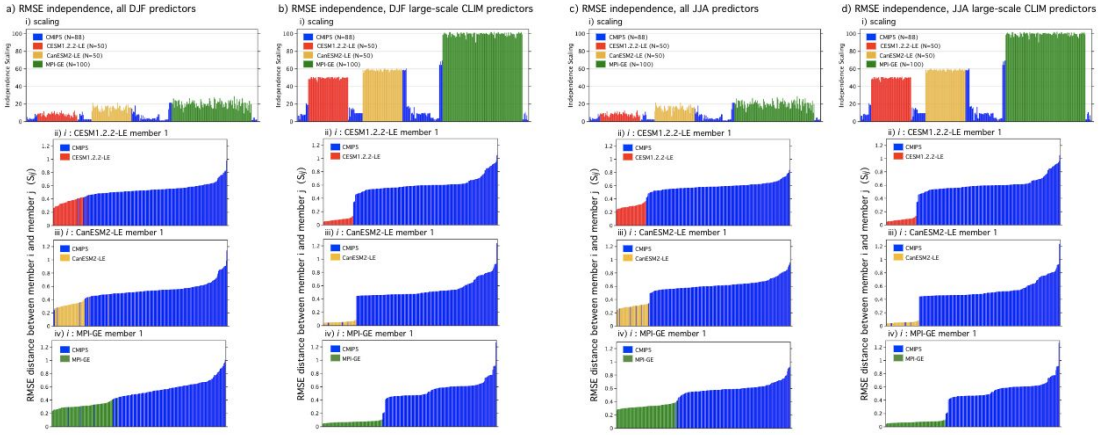


Figure 4. (a-di) The RMSE independence scaling of the SMILEs and CMIP5 ensemble members, shown in the order listed in Table 1. Panels ai and ci show the scaling computed from the 9 predictors used in the original DJF and JJA RMSE distance weighting respectively. Panels bi and di show the scaling computed 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.

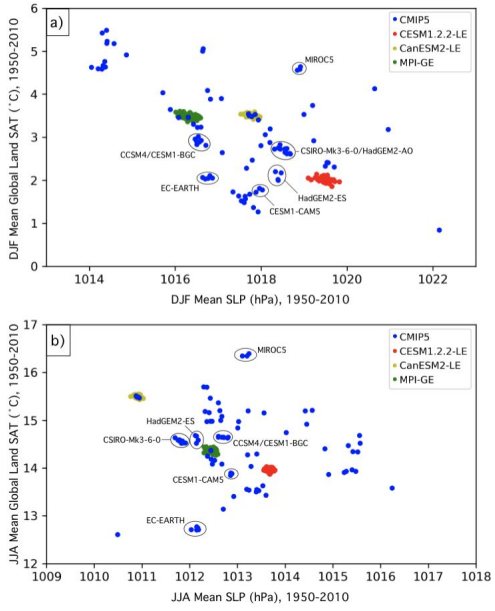


Figure 5. a) Scatter plot showing how ALL ensemble members distribute in the DJF European sector SLP climatology / DJF Global Land SAT climatology predictor space. Select clusters within the CMIP5 ensemble (blue) are labelled by model name. The CESM1.2.2-LE is indicated in red, the CanESM2-LE is indicated in yellow, and the MPI-GE is indicated in green, consistent with other figures. b) As in a), but for the JJA European sector SLP climatology / JJA Global Land SAT climatology predictor space.

2. Terminology of "independence weight": it's confusing that both numerator and denominator of equation (1) are called "weights". A more intuitive use would be that a "weight" is a quantity that's larger when the model run is given more weight, i.e. a stronger influence on the results. The term "weight" is used this way for the overall weight (left hand side of eq. 1) and the "performance weight", but not for the "independence weight". Could a different name be used or if not then could a note on this terminology at least be made clearly in Sec 3 (around l.155-165)?

Thank you for pointing this out, this distinction of using "weight" to refer to the directly proportional performance term and "scaling" to refer to the inversely proportional independence term will definitely help with the clarity and will be henceforth used.

3. The explanation of dynamical adjustment in Appendix A could be clearer. The meaning of Ns, Na and Nr isn't clearly explained. The description refers to the "observational record" but the method is also applied to models, both past and future. Is this appendix meant to be a standalone description of the method or is it assumed the reader is already familiar with the references? I would suggest to improve this description for benefit of completeness and also so that the reader isn't obliged to go to the references to have a basic understanding of the method. It could at least be described how the weights (beta_i) are determined.

We apologize for leaning too heavily on prior works for the basic understanding of the dynamical adjustment methodology. We have revised the section the as follows, an we hope you find the description clearer:

"To obtain estimated forced trends in SAT, a method of dynamical adjustment, based on constructed circulation analogues, is used (Deser et al., 2016; Lehner et al., 2017; Merrifield et al., 2017; Guo et al., 2019). Dynamical adjustment provides an empirically-derived estimate of the SAT trends induced by atmospheric circulation variability; removal of this circulation-driven component from a SAT record thus reveals an estimate of the SAT trend associated with thermodynamic processes and radiative effects. Dynamical adjustment relies on the ability to reconstruct a monthly mean circulation field, which we represent with sea level pressure (SLP) as in Deser et al. (2016), from a large set of analogues. Here, SLP analogues are selected from 60 possible choices (from the period 1950-2010), excluding the target month, and the method is therefore referred to as the "leave-one-out" method of dynamical adjustment.

SLP fields in SMILE members, CMIP5 ensemble members, and the observational estimates ERA-20C and NOAA-20C are constructed in this manner for target months in the 1950-2010 period. For model years 2011-2099, analogues are selected from the entire 1950-2010 period. No notable trends in SLP have been identified over this period in previous dynamical adjustment studies (Deser et al., 2012, 2016; Lehner et al., 2017).

It is important to acknowledge that because of the paucity of analogue choices in leave-one-out dynamical adjustment, the term "analogue" is a bit of a misnomer. The term evokes the idea of a match, though in practice, analogues may not closely resemble the target. For convenience, we will continue to refer to the months used in target SLP construction as "analogues", but we do so with the understanding that target and analogue patterns may differ over the selection domain. A month is determined to be an analogue of the target month if the Euclidean distance between target and analogue SLP is small. Euclidean distance is computed at each grid point and averaged over the European sector domain also used for SLP predictors (25-90˚N, 60˚W-100˚E). This selection metric, therefore, does not require an analogue to match the target month spatially over the whole domain. This is necessary because, with 60 possible options, it is statistically unlikely that a "perfect" analogue will exist for a particular target month. van den Dool (1994) found that it would take on the order of 10 years to find two Northern hemisphere circulation patterns that match within observational uncertainty. With this in mind, a smaller than hemispheric domain and an iterative averaging schemes are employed to make the most of "imperfect" analogues available (Wallace et al., 2012; Deser et al., 2014, 2016). Once the Euclidean distances are determined, 50 closest SLP analogues are chosen, and the iterative process of selecting 30 of 50 SLP analogues and optimally reconstructing a target SLP field X_h commences. The optimal reconstruction of target SLP is mathematically equivalent to multivariate linear regression; each analogue is assigned a weight (β) such that a weighted linear combination of analogues produces a least-squares estimate of the target SLP. β is computed through a singular value decomposition of a column vector matrix X_c containing the 30 selected analogues and can also be estimated using through a Moore-Penrose pseudoinverse:

$$\beta = [(\mathbf{X}_c^T \mathbf{X}_c)^{-1} \mathbf{X}_c^T] \mathbf{X}_h \quad \text{(A1)}$$

The analogue weighting scheme ensures that analogues which are further from (closer to) the target, in a Euclidean distance sense, contribute less (more) to the constructed SLP field.

After the target SLP field is constructed, the **β** values derived for each SLP analogue are applied to their corresponding monthly-averaged SAT fields. Prior to the application of weights, a quadratic trend representing anthropogenic warming is removed from the SAT record at each point in space. The purpose of this detrending is so that months picked from the end of the record do not 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). Detrending strategies are further discussed in Deser et al. (2016). The weighted, detrended SAT fields are then used to construct a dynamic SAT anomaly field for the target month. SLP, which is a representative of low-level atmospheric circulation, and SAT are physically related; SLP-derived weights are applied to SAT to empirically construct that relationship. Conceptually, dynamic SAT anomalies are those that would occur given the attendant circulation pattern. The second through fifth steps of dynamical adjustment (selection of 30 of 50 SLP analogues, optimal reconstruction of target SLP, and construction of dynamic SAT) are then repeated 100 times, following Lehner et al. (2017). The dynamic component of SAT in the target month is the average of the 100 constructions. It is then subtracted from SAT in the target month to find the residual thermodynamic component of SAT, used as an estimate of the regional SAT response to surface processes and radiative forcing. The trend of the residual thermodynamic SAT component is used as a predictor in this study; trend is computed at each land grid point in the predictor domain and subsequently area-averaged.”

Comments & suggestions by line number:

18: "increases linearly": maybe say "changes linearly". It seems unintuitive (at least to me) to describe the weight as increasing when it’s actually the reciprocal of the "independence weight" that gets multiplied by the performance weight. This makes sense after reading Sec. 3, but someone reading just the abstract could find this confusing.

Thank you, corrected.

20: "subsetted ensemble" –> "subsetted ensemble of one model run per model"

Corrected.

45: "more-than-representative uncertainty" - what does this mean? Please clarify and/or give a reference for this concept.

We have revised this sentence to:

"Known biases associated with cloud processes, land-atmosphere interactions, and sea surface temperature (e.g. Boberg and Christensen, 2012; Li and Xie, 2012; Pithan et al., 2014; Merrifield and Xie, 2016) may result in more uncertainty in projections of future climate than is warranted given our understanding of the climate system (Vogel et al. 2018). Using expert judgement to weight or select multi-model ensemble members based on process- or region-specific metrics of performance has been shown to justifiably constrain uncertainty (e.g. Abramowitz et al., 2008; Knutti et al., 2017; Lorenz et al., 2018)."

62: "ensemble, first," –> "ensemble. First,"

Corrected.

97-98: Don't most reanalyses provide both SLP and SAT? 100: ERA-20C doesn't assimilate surface temperature. Do you know that it's suitable for evaluating SAT trends? This could be tested by comparison with an observational dataset (HadCRUT4?).

We have added an additional dataset to better assess observational uncertainty: the Berkeley Earth Surface Temperature (BEST) product and NOAA-20C SLP reanalysis V3. We were unfortunately not able to use HadCRUT4 without decimating all other fields onto its 5˚x5˚ grid. But we hope that the observational datasets we did select serve to establish observational uncertainty/suitability.

109: "representative distribution" - what is this? The distribution is well defined for each model by virtue of the ensemble size. But is this term meant to suggest it's "representative" of the true variability? If not (and I'm not sure how that would be known), suggest change "representative" –> "well defined".

We agree that "well-defined" is a much better choice for describing SMILEs distributions. We have changed the data section quite a bit, but have used this recommendation as follows:

"The multi-model CMIP5 ensemble (Fig.1 blue) has more spread than the single model SMILEs, demonstrating that model uncertainty does rise above well-defined estimates of internal variability in the two European regions and seasons considered."

113: put quotes around "macro" (similar to "micro" at l.117)

Punctuation updated, thank you.

121: "preindustrial" misspelled

Thank you, Corrected.

122: "conditions, " –> "conditions: "

Punctuation updated, thank you.

141: "and model, " –> "and model; "

Thank you for the catch. This paragraph has been heavily revised in the revision and the relevant sentence has been removed.

167: "definition of climate, " –> "definition of climate: "

This sentence has been split up in the revision and now reads:

"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."

170: "fit for purpose" –> "fitness for purpose"

Changed.

178: "trend" –> "estimated forced trend" (and perhaps add that meaning of this will be explained below)

Thank you for the catch. This has been changed as follows:

"...and a 50-year derived SAT trend (estimated residual thermodynamic trend; described in more detail in subsequent paragraphs) for the period of 1960-2009"

191: "idea" –> "assumption"

Changed.

205-206: Perhaps clarify here that the SMILEs reinforce the relationship in the sense that model-mean values (3 data points, one for each SMILE) support the relationship. It's not because the relationship is evident within the SMILEs, which it should not be since the relationship is due to model differences.

Thank you, this sentence has been revised as follows:

"In contrast, a relationship emerges in summer, a season with less midlatitude SAT variability, between 1960-2009 and 2050-2099 European SAT trends. The linear relationship is reinforced by the SMILEs in a model mean sense, i.e., the three new models added to the CMIP5 ensemble support the relationship (Fig.2bi). It is not evident within the SMILEs themselves, which reflects that the relationship is due to model differences not the behavior of individual members."

222: "trends, " –> "trends: "

Changed.

223-225: The positive relation for the SMILEs is only for 3 models, and the CMIP5 relation is very weak. Overall this suggests no relation (across models) between past and future estimate forced trend for DJF NEU.

You are correct, we've revised the sentence to read:

"The addition of the SMILEs then introduces a slightly positive relationship between past and future responses (Fig. 2aii, black line) not apparent in the CMIP5 ensemble (Fig. 2aii, blue line), though no strong relationship emerges from variability in either case."

226: Not it's "bolstered" - perhaps more accurate to say that it's "robust to"? The relationship looks essentially the same for both cases in both panels of Fig 2b.

Thank you, changed.

229: "use is" –> "use them"

Thank you, corrected.

235: "SMILE" –> "SMILEs"

Corrected.

235: Are these distributions over gridpoints? That is, at each gridpoint in the domain of interest, a weighted or unweighted mean over models is computed, and this contributes one member of the distributions shown in the Fig 2 box-whisker plots. Please clarify. If not then I'm not sure what the "weighted distribution" is.

It's a good point that it wasn't entirely clear on what scale the weights were computed on. To address this, we've added the following to the weighting section:

"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) and Sij (Figure B2) for each member."

And the following at the aforementioned location:

"Two ensembles are considered, one comprised solely of CMIP5 members (CMIP5; distribution of 88 values) and one comprised of all available members from CMIP5 and the three SMILEs (ALL; distribution of 288 values."

-

From this point, we have made a complete overhaul of the Results section and the relevant sentences no longer exist. We will take all of the following recommendations forward in the revision, particularly to describe distributional shifts. Thank you for all the specific feedback!

253: "tail broadly" –> "tail is broadly"
288: "function number" –> "function of the number"
300: "a weight" –> "an overall weight"
324-325: Seems an odd way to say this. The distance-based independence measure used in the weighting is a proxy for model structural differences. Consider rephrasing as something like: "models have some independence from one another while members of a SMILE have none (in the sense of model structural uncertainty)".
367-368: Perhaps qualify this by saying it's a modest narrowing (according to Fig 3ai).
371-372: Again, this is a modest shift. Fig 3ai shows the 95th percentile of weighted ALL is only slightly higher than for unweighted ALL.
406: "target month" - meaning the target year, for the month under consideration?
457: I think you mean Figure B3.
488: "domain-averages" –> "domain-averaged"
489: The "emergent constraint" here is that the model's climatological bias is more or less unchanged from past to future. Perhaps useful to also describe it in this simpler way?

Figure B3 caption: perhaps note that the unweighted distributions are the same in every panel, being shown for reference.

To make the figure less busy, we have removed the unweighted distributions from this figure.