Selecting A Climate Model Subset To Optimise Key Ensemble Properties - Herger et al. (2017)

Response to Referee #1

We thank referee #1 for his/her valuable feedback. This document outlines our point-by-point responses to the comments made by referee #1 and the improvements we are going to make to the manuscript (*italicised text in quotation marks*).

One weakness (which is shared with many papers) is the limited discussion of principles underlying the selection of the sub-ensemble. Having a good ensemble mean is one possible property that we might like an ensemble to have, but it's not clear whether/when/why it is important. To illustrate, if we consider a simple one-dimensional case where truth is known to take the value 5, is it better to use an ensemble of two models which take the values 3.5 and 4, or a pair with the values 0 and 9, or yet another pair with the values -10 and +20? The ensemble mean improves (relative to truth) across these three sets, but the models themselves are getting worse, which may be a concern. Another distinction between these ensembles is that both members of the first pair share a bias in sign whereas the other two ensembles bound reality which is close to (at) the 50th percentile. I don't think these questions are easily answered but they do seem fundamental to the whole concept of how and why we use ensembles, so I think they ought to be discussed a bit more fully in the manuscript. Do the authors actually have a good argument why they would like to find an ensemble with a good mean? The analysis does also consider the issue of ensemble spread (both in model selection and assessment of the predictions) to some extent but this isn't really placed in any coherent mathematical framework. For example, the extended cost function on page 9 provides one route to distinguishing more clearly between the three different types of sub-ensembles in my example, but there does not seem to be any structured reasoning behind any particular choice.

We agree with the reviewer that there will never be a method that can deal with ensemble selection for all possible applications. Depending on the application, the ensemble we desire will have different properties. In some cases, finding an ensemble whose mean is close to observations might be the highest priority. This might be the case for downscaling approaches, where we want a discrete subset of models which are centered on the "truth". For impacts for example, maximising the spread in the ensemble is also desirable, as we are interested in the full spread of climate outcomes (in this case, -10 and +20 might be what we want, with models being on either side of observations). The purpose of the extended cost function is to cater for the issue of poor performing models being part of the optimal subset (see next paragraph). We added a sentence to the manuscript (Section 4.1) to address the problem of the optimal subset including poor performing models if we solely focus on optimising the RMSE between the ensemble mean and the observations: "*Also, solely focusing on the ensemble mean could potentially lead to poorer performing individual models as part of the optimal subset despite getting the mean closer to observations.* "

In other cases, we might only want the best-performing models to be part of the ensemble (here we would choose 3.5 and 4). For fields that depend on distribution shapes being representative of observations (event attribution; projections of extremes), the cost

function to minimise could be the Kolmogorov–Smirnov test statistic (this is something we are currently applying this method to).

We have identified the need for ensemble selection to be case dependent. This is why we introduce the flexibility of our cost function in the manuscript. Terms can be added to the cost function to account for different desired ensemble properties. For example, we added an additional term (Term 2 in Section 4.1, "Sensitivity to the underlying cost function") to make sure that the optimal subset does not include poor performing model runs (e.g., models of Venus or Mars will be eliminated). It is possible to add additional terms to Equation (2) to e.g., ensure that the ensemble spread is maximised, if that is an important feature of the desired subset.
We are not trying to advertise the idea of solely focussing on the mean when selecting an ensemble. This point has been discussed amongst the authors of this paper at length and we have thus tried to highlight it more prominently in the revised manuscript. We decided to show the results based on optimizing the ensemble mean because it is a conceptually simple way to illustrate this new approach. Adjusting the cost function (as done on page 9) demonstrates the flexibility of this approach for ensemble selection. For clarity, we have added the following paragraph to the section with the extended cost function:

In Section 4.1 ("Sensitivity to the underlying cost function"):
"Reasons to use ensembles of climate models are manifold, which goes hand in hand with the need for an ensemble selection approach with an adjustable cost function. Note, that we do not consider the MSE of the ensemble mean as the only appropriate optimisation target for all applications. Even though it has been shown that the multi-model average of present day climate is closer to the observations than any of the individual model runs (e.g., Gleckler et al. (2008); Reichler and Kim (2008); Pierce et al. (2009)), it has also been shown that its variance is significantly reduced relative to observations (e.g., Knutti et al. (2010)). Also, solely focusing on the ensemble mean could potentially lead to poorer performing individual models as part of the optimal subset despite getting the mean closer to observations. Errors are expected to cancel out in the multi-model average if they are random or not correlated across models. Finding a subset whose mean cancels out those errors most effectively is therefore a good proxy for finding an independent subset, at least with respect to this metric, and is sufficient as a proof of concept for this novel approach."

In the Introduction:
"The aim of this study is to present a novel and flexible approach that selects an optimal subset from a larger ensemble archive in a computationally feasible way. Flexibility is introduced by an adjustable cost function which is allowing this approach to be applied to a wide range of problems."

Section 3:
"We then examine the sensitivity of results to observational product, cost function (to demonstrate flexibility by optimising more than just the ensemble mean) and other experimental choices."

The method of ordering by model performance seems to have some superficial similarities with Bayesian Model Averaging principles, albeit with 0-1 rather than continuous weights (and implicitly a uniform prior even when initial condition ensembles are present). It might be worth mentioning the link though I don't suppose the conclusions drawn here will be directly applicable to BMA due to the methodological differences. In particular the implied

uniform prior even when IC ensembles are present would probably be considered inappropriate for any more formal implementation of BMA. On the other hand this similarity does highlight the major issue with the method, which is why the RMSE of the ensemble mean is considered to be an appropriate optimisation target in the first place. For BMA (which, at least in many artificial idealised cases, is basically the correct solution to ensemble calibration and weighting) the ensemble mean is not optimised in any meaningful sense even though it will tend to be moved towards observations in the posterior.

The idea of BMA is certainly similar as it also tries to solve the problem of model selection and combined estimation. However, the difference between 0/1 and continuous weights is central in this case. Also, the model weights in BMA would be derived from performance (model's capability to accurately describe the data) only. As we have shown in our study, solely accounting for performance in ensemble selection is not recommended and can even be worse than a random ensemble. This probably has implications for the BMA approach.

As noted above, optimising for the ensemble mean as presented in the manuscript was a conceptually simple approach to illustrate the technique and we do not claim that solely focussing on the mean is desirable for all cases. Hence the addition of the extended cost function. Hopefully our additional text has clarified this a little.

"binary": this word appears 3 times, it might be worth explaining this more fully at the outset as meaning weights of 0 or 1 (and why this restricted choice is significant/beneficial). Actually, I believe the issue is not so much the contrast of binary (or even discrete) weights with continuous, but rather more precisely the number of zero weights, since this is what allows some models to be discarded, thereby reducing computational effort. See for example the lasso approach to regression which might have been a plausible alternative to the 0/1 methods used here. However I'm not suggesting that the authors need to investigate this as part of this piece of work.

We agree with the reviewer that we should be clearer on the meaning of the word "binary" in this context. We have added a sentence to the main text, also highlighting that we contrast binary with continuous weights mostly because of the zero weight. For clarity, we have added the following two sentences to the Introduction:
"With binary we refer to the weights being either zero or one, and thus a model run is either discarded or part of the subset."

"More precisely, it is the number of zero weights that leads to some models being discarded from the ensemble."

The Lasso approach is certainly an interesting option due to assigning weights of 0 to some model runs, however it is, to our understanding, not possible to adjust the cost function that is being minimised (by default: RMSE). The optimizer we are using in our work allows us to define constraints and cost functions depending on the use-case. We regard the flexibility of being able to adjust the cost function depending on the aim of the study as an important strength of our method. We added the following text (Section 5):

"The lasso regression analysis method (Tibshirani, 2011) often used in the field of machine learning tries to select a subset of features (in our case: model simulations) to improve prediction accuracy. It is similar to the presented approach in a way that it also

selects a subset of models by applying weights of zero. However, contrary to the method presented here, it is to our knowledge not possible to customise the cost function that is being minimised (by default: RMSE).”
Fig 1: The red triangles are not explained in the caption, though presumably they represent the optima from the black triangle cases.
The reviewer is correct. We have now added the following explanation of the red triangles to the caption of Figure 1: *“The corresponding red triangle is the optimal subset of the black triangle cases.”*