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

Response to Referee #2

We thank referee #2 for taking the time to review our manuscript. This document outlines our point-by-point responses to the comments made by referee #2 and the improvements we are going to make to the manuscript (*italicised text in quotation marks*).

The paper is generally well written and fits within the scope of ESD. However, the authors present this method as something that is simple to calculate and generally applicable which is by no means the case. In fact, the authors lack to clearly highlight the aspects of their work that go beyond what has already been published. The example given as an application of their method does not seem well suited as a proof of concept to select an optimal ensemble for climate applications as it is too simple. A demonstration of how their method can be applied to multi-variable problems using multiple metrics as it would typically be needed for climate analyses would be more helpful. Another important point that is not discussed sufficiently is how to account for observational uncertainties, which is of key importance when ranking and benchmarking models. Also, even though the term 'model interdependence' is repeatedly used, no attempt is made to define model interdependence or discuss the relevant aspects for determining an optimal ensemble. Further work is required to clarify what we can learn from this study and in which cases this method can be applied, before I can recommend publication in ESD, see details below.

We thank the reviewer for his/her comments. We note the lack of clarity in the Introduction, which when addressed should answer a few of the reviewer's concerns (see below and other responses to reviewer concerns in this document). It is important to highlight that there is no single best approach for ensemble selection available and our approach does not replace any of the other techniques in the literature. Any approach will have to be tailored depending on the specific use-case. Using Gurobi offers the ability to customise the cost function and metrics used for obtaining an optimal subset. This is essential for a given approach to be widely applied. Attempting to find a single best approach is therefore a pointless task; hence our focus on finding an approach that could potentially be applied to a wide range of use-cases.

We explain the range of approaches for model weighting that have recently emerged with the range of applications that such an approach can be applied to. Bishop & Abramowitz (2013) for example focus their approach solely on variance by looking at time series and finding a linear combination of model runs to most accurately represent observational variability. Sanderson et al. (2015) however focus on climatology without considering any time component. Just as there are many ways of addressing model performance, there are many ways of addressing independence.

The text in the Introduction was adjusted to make this clearer:
"[...] The same process was also used for future projections, with the danger of overfitting mitigated through out-of-sample performance in model-as-truth experiments (Abramowitz and Bishop, 2015). In their approach, they solely focus on variance by looking at time

series. Another method also using continuous weights but considering climatologies rather than time series was proposed by Sanderson et al. (2015a). It is based on dimension reduction of the spatial variability of a range of climatologies of different variables. [...]"

The reviewer rightly comments that we did not define model interdependence. This is because the definition of dependence is problem-dependent. Most of the authors on this manuscript attended a workshop last December on exactly this topic where it became evident that a generally agreed-on definition is currently absent.

Rather than testing our approach on multiple variables at a time we did it separately for surface air temperature and total precipitation. Monthly mean temperature is a variable commonly used by the community, and the problem at hand (e.g. one model one vote) has been clearly framed in other work by some authors on this paper.

We discuss the topic of observational uncertainty below in our answer to Q10.

1. What is the aim of this study? Is the aim to (a) present a new method: then please what is new, what are the differences and advantages compared to the other methods that have recently been published (e.g., [Knutti et al., 2017; Sanderson et al., 2015a; b]? Quantitative comparisons would be required. (b) to present a method that is only slightly different but to provide a demonstration that this method can be used for impact studies and other climate applications? The paper fails to convincingly show that this method can be applied for concrete applications, see further comments below. The example given in the manuscript is too simple to provide any helpful insights beyond of what has already been published (see references above).
Currently a mixture of both is presented.

We note the lack of clarity in our framing of the contribution this work makes, and have therefore adjusted the Introduction accordingly:

"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."

"Such an approach with binary (0/1) rather than continuous weights is desired to obtain a smaller subset that can drive regional models for impact studies, as this is otherwise a computationally expensive task."

The aim of this manuscript is mainly the reviewer's (a). We are presenting a new, flexible ensemble selection method that can be applied to impact studies. It is not clear to us that in order to address point (a), a quantitative comparison to previous approaches is required. Comparing existing approaches for a given use-case is certainly something valuable that should be done in the future, but it goes beyond the scope of this study, given that detailing the technique alone has already made this manuscript reasonably long.

We also think that introducing a new approach, as stated in (a) without showing where it could be applied would not be very useful. We therefore also touch on (b) by highlighting that such an approach could be used for impact studies which requires a small number of runs (e.g. for dynamical downscaling). This point has been addressed in the introductory

part of the manuscript, see here:
"Regional dynamical downscaling presents a slightly different problem to the one stated above, as the goal is to find a small subset that reproduces certain statistical characteristics of the full ensemble. In this case the issue of dependence is critical, and binary weights are needed, since computational resources are limited."
As our approach results in a discrete subset, we do not see the need to perform the additional step of using this optimal subset for downscaling and impact assessment. The novelty is to find a discrete optimal subset for a given use-case, and thus using that for impact studies would add little to the literature and goes beyond the scope of this study.

We believe the Introduction already covers the main differences between this approach and existing approaches.

Theoretically, a (c) could be added to our aim: making it clear that asking 'which of the existing approaches is the best' is not a well framed question. It is equivalent to asking 'which climate model is the best?', without specifying the application. Only when calibrated to a given use-case it is useful to compare existing approaches of ensemble selection, or definitions of model dependence.

2. The paper could expand on recommendations of pre-selection in an ensemble. The statement on p6, l.34 that similar improvements can be made if closely related model runs are a priori removed from the ensemble to start off with a more independent ensemble could be such a recommendation.

One conclusion that emerged from the workshop on model dependence in multi-model climate ensembles, held in December 2016, was the idea to write a review paper on this topic. The participants are currently working on a review of the current literature around this topic and are trying to give recommendations on how to use multi-model ensembles whose members are not independent.

Pre-selection in the ensemble will always be somewhat subjective and case-dependent. Giving general recommendations of pre-selection in an ensemble is thus not straightforward. We have, however, added the following sentence regarding the possibility of filtering out certain model runs before starting the optimization process (Section 4.1, "Sensitivity to the underlying cost function"):

"It would of course also be possible to make an a priori decision on which models should be considered before starting the optimisation process."

3. It is quite confusing that within a short time this is the forth (?) recommendation for a method that should be applied for model weighting considering both model performance and interdependence (with two of the authors of this paper being also authors on all the previous papers). Yet the authors do not show the differences between this newly presented method and the previous ones. Neither they give a recommendation whether this method now supersedes the previous ones nor do they provide a sophisticated comparison of the published methods for a concrete example. For example, how would the results on sea ice extent weighting from Knutti et al. [2017] change if this method instead of the Knutti et al. [2017] method was applied and what are the policy and stakeholder relevant implications when analyzing model ensembles?

We hope that our answer to the reviewer's first comment already addresses some of those

concerns. As mentioned before, there is no single best approach. Which approach to choose depends on the the specific use case. In some cases (e.g., when simply computing a mean and range across a set of GCMs), continuous weights are sufficient. In others, having a discrete subset of models is appropriate, e.g., for subsequent downscaling, because dynamical downscaling is computationally expensive and can thus only be applied to a small subset of model runs.
The reviewer mentioned the Arctic sea ice extent weighting from Knutti et al. (2017). This is an example where the benefit of model weighting (compared to simply taking the equally-weighted multi-model mean) is expected to be very large as some models are not even able to capture the present day state properly. Global mean temperatures are usually captured more accurately by models than sea ice extent and if we see improvement in the ensemble mean in this case, we regard this as a stronger proof of concept. We therefore do not see the need to apply our method to this exact use-case.

The introduction states the main differences between the existing approaches. However, for clarity we have added a few sentences to the Introduction to make this clearer (see also below Q4):

"This approach is not meant to replace or supersede any of the existing approaches in the literature. Just as there is no single best climate model, there is no universally best model weighting approach. Whether an approach is useful depends on the criteria that are relevant for the application in question. Only once the various ensemble selection approaches have been tailored to a specific use-case, can a fair comparison be made. Flexibility in ensemble calibration by defining an appropriate cost function that is being minimised and metric used is key for this process."

4. Related to the above: if the authors can't convincingly show what is different to the above methods, then it is also not clear what is new.

The main difference between this approach and most of the existing ones is the use of binary (zero or one) weights rather than continuous weights. Having a zero weight leads to a discrete subset which can subsequently be used for regional downscaling (and used for impact studies) — desirable as computational cost is then reduced compared to if one would use the full ensemble. Note, that the stepwise model elimination procedure described in Sanderson et al. (2015) can also be considered to be an approach with binary weights. It is different from what we did as the focus is on joint projections of multiple variables and is arguable more technically challenging to implement.

Apart from having a discrete subset, the method allows for changes in the cost function being optimised and the metric used. Different from most other approaches, out-of-sample performance has been tested to avoid overfitting of the ensemble to the present-day state. Also, by providing the code, we see no reason why it would be much of a hurdle to implement. Other published approaches are considerably more technically challenging (e.g. Sanderson et al. (2015), Bishop and Abramowitz (2013)).

To make this clearer, we added a few sentences to the Introduction of the manuscript (see above).
"This approach is not meant to replace or supersede any of the existing approaches in the literature. Just as there is no single best climate model, there is no universally best model weighting approach. Only once the various ensemble selection approaches have been tailored to a specific use-case, can a fair comparison be made. Flexibility in ensemble

calibration by defining an appropriate cost function that is being minimised and metric used is key for this process."

We have also added the following paragraph to 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)). 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."

5. Climate change is not a single, but a multi-variable problem. Using RMSE as only metric does not always seem appropriate, more comprehensive metrics are available (see for example Xu et al. [2016]). The authors show that the optimal ensemble is performing best if the bias of the model subset average should be minimized - essentially indicating that the solver is working as anticipated (p6, l24). However, if a bias correction with climatological mean temperature would be the answer for an optimal ensemble, one could for example tune the models accordingly. There are good reasons why one might not want to do so (see for example Mauritsen et al. [2012]). Why would an ensemble that captures mean temperature be better than another one? The multi-variable issue is mentioned on p7,l29 but it would be good if the authors could expand their analysis to explore this further and if possible give advice to the reader.

The reviewer is correct that climate change cannot be fully addressed by solely looking at one variable or metric. However, this is not what we are trying to accomplish with this work. Note that we optimize spatial fields not global means, and the former cannot really be tuned in a GCM.
To introduce this novel approach, we separately applied it to surface temperature and total precipitation, using RMSE as a metric. It can of course be applied to more variables, as long as reliable observations are available, and once suitable scaling factors are chosen to aggregate different units. As long as it can be implemented into the solver Gurobi, almost any other metric of interest is possible. For example, we have begun working on a related project using the Kolmogorov–Smirnov test statistic instead of RMSE (reducing distribution biases which is for example relevant for event attribution). We expanded the paragraph with the following text where we talk about the multi-variable issue to make the flexibility of this approach clearer (Section 4.1, "Sensitivity to the underlying cost function"):

"The cost function presented in this study solely uses MSE as a performance metric. There are of course many more metrics available (e.g. Xu et al. (2016), Taylor (2001), Gleckler et al. (2008), Baker and Taylor (2016)) that we might choose to implement in this system for different applications. So as not to confuse this choice with the workings of the ensemble selection approach, however, we illustrate it with RMSE alone, as this is what most existing approaches in this field use to define their performance weights (e.g. Knutti et al. (2017), Sanderson et al. (2017), Abramowitz and Bishop (2015))."

We note that when comparing panels (a) and (b) in Figure 1, depending on the chosen variable, we end up with a different optimal ensemble size, different ensemble members and different performance gains. This is best framed as a calibration exercise since one can only obtain an optimal subset for a clearly defined use-case (given the variable, metric, region, observational product etc.).

If the goal is to obtain a single optimal subset across multiple variables, one could preprocess the model output in a way Sanderson et al. (2015) did in their Journal of Climate paper (see their Figure 1). Gridded model output is normalized and concatenated into a long multi-variable vector which is then used for further analysis where a single cost function is optimized. We added a few sentences to our manuscript highlighting the possibility of doing the same (see below). Even though this will result in a single optimal subset across all variables, it is sensitive to how the variables were normalized and it also conceals the fact that the optimal subset for the individual variables might look very different. In many cases it is therefore useful to employ the calibration exercise on each variable separately to see how the optimal subset varies instead of first combining all the variables and then finding a single optimal subset. Additionally, if only one variable is of interest for a particular case, one can only gain from selecting a subset based on only that variable. The following text has been added (Section 4.1, “Variable choice”):

“This could most simply be done using a single cost function that consists of a sum of standardised terms for different variables. This is similar to what has been done in Sanderson et al. (2015a) (see their Figure 1). However, this might conceal that fact that the optimal subsets for the individual variables potentially look very different. ”

Alternatively, a Pareto solution set of ensembles is possible, which is often used in multicriteria calibration papers for hydrological models. For example: Gupta et al. (1998): “Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information”.

6. The physical consistency is mentioned yet the authors are not evaluating the optimal ensemble whether it captures other important climate features including modes of variability. This strongly limits the applications of this method and generalizations of the application like the one on p4,l10 (’We argue optimally selecting ensemble members for a set of criteria of known importance to a given problem is likely to lead to more robust projections’) should be avoided.

Given that we are not assigning continuous weights to the CMIP5 ensemble member, our subset is as physically consistent as the original ensemble. While a model average may not show physically plausible behaviour, each single model run should (to the degree that it represents the real world), and using each individually for impact analysis or downscaling will preserve as much of the physical consistency as possible.

It is true that the cost functions we have used to illustrate the technique are simple, not comprehensive, and in particular not focused on modes of climate variability. This work is only a first step, being the introduction of a new method. There are a myriad of modes of variability that one could attempt to calibrate an ensemble towards. Which ones are important? Again, it comes down to the specific use-case (e.g. region, variable, question.)

7. Related to the above: what about model tuning? A model could be tuned towards a correct present-day temperature climatology but it might still not be the best model to

project climate? What about climate sensitivity?

We agree with the reviewer that a model which is very closely tuned toward the present-day state won't necessarily be skillful for projections. The model-as-truth experiment in our paper is intended to check for the possibility of overfitting/"over-tuning" of the ensemble members to the present state, although it could perhaps be better explained. We added the following to better motivate the use of the model-as-truth experiment (Section 4.2.1):

"Rigid model tuning for example could cause the ensemble to be heavily calibrated on the present-day state. An optimal subset derived from such an ensemble would not necessarily be skillful for future climate prediction as we are dealing with overfitting and we are not calibrating to biases that persist into the future. This is where model-as-truth experiments come into play."

Note that while global mean temperature can be tuned to some degree, the spatial fields of climatology cannot (otherwise the current GCMs would not have such large persistent climatological biases). Regarding climate sensitivity, climate models which are biased high (in terms of temperature for example) in present day, are often at the higher end of the distribution in the projections. In our approach, we make use of this persistent bias. Improvement of our optimal subset relative to the ensemble mean (of 1 run per institute) is expected to decrease with increasing time/forcing, as the climate system will reach a state it has never experienced before. At that point, calibrating a subset on the present day might not lead to any improvement. However, this is a problem for any weighting or calibration approach, and one way to check for this is to use model-as-truth experiments to show where we have a breakdown of predictability. This is certainly something worth exploring in a future study.

However, for what we have used it for, there seems to be some predictability in the system and using the optimal subset out-of-sample is likely to have advantages over simply using the equally-weighted multi-model mean. For clarity, we have added the following text (Section 4.2.1):

"Climate models which are biased high (in terms of temperature for example) in the present day, are often at the higher end of the distribution in the projections. This is related to climate sensitivity and our approach is able to make use of this persistent bias."

In Section 5:
"Using model-as-truth experiments, we observed that the skill of the optimal subset relative to the unweighted ensemble mean decreases the further out-of-sample we were testing it. This breakdown of predictability is not unexpected as the climate system reached a state it has never experienced before. This is certainly an interesting aspect which should be investigated in more depth in a future study."

8. Can process-oriented diagnostics be used? This might be an interesting option to avoid selecting models that get the right results for the wrong reasons.

This is an interesting point, which also came up in discussions among the authors of this manuscript. Depending on the application, process-oriented diagnostics can potentially improve the ensemble selection by giving us more confidence of selecting the subset for the right reasons. We decided to focus on global temperature and precipitation as this

manuscript is a proof of concept, and introducing the ensemble selection approach for another specific example might be confusing. Also, multiple observational products exist for those two variables and sensitivity to the chosen product could be tested.
The metric used for this approach can take any form which makes it very flexible. A few sentences have been added to the manuscript to highlight the possibility of using process-oriented diagnostics (Section 4.1, “Variable choice”):

“The presented approach can obtain an optimal subset for any given variable, as long as it is available across all model runs and credible observational products exist. One might even consider using process-oriented diagnostics to provide greater confidence when selecting a subset for the right physical reasons.”

9. The study is motivated by the need of the impact and user community who need concrete guidance on how to use the large zoo of model output available in the CMIP ensemble (e.g. first sentence in abstract). While this is true, the paper needs to improve on giving concrete guidance. It either needs to provide realworld examples or avoid generalizations of the applicability of the method. It mathematically works fine, but whether or not it should be applied depends on whether the diagnostics chosen for the benchmark are actually relevant for the specific application. Finding these diagnostics remains a challenge.

Given that our approach results in a discrete subset, using it subsequently for regional downscaling and then impact assessments is certainly an application we had in mind. However, we do not agree with the reviewer that it should be within the scope of this study to give an impacts-replated example. As the reviewer states, it “mathematically works fine” and this is what we wanted to demonstrate here (proof of concept). The novel part is finding a discrete subset of model runs and the subsequent steps needed for impact assessments are unchanged and thus do not need to be discussed here in detail.

We have adjusted the manuscript to highlight the importance of tailoring the cost function and metric to the problem at hand to avoid generalizations, as noted above.

10. The authors show that different observational products lead to different ensembles (Figure 1 and S1). But given there is observational uncertainty, some choices would need to be made. It would be good if the authors could expand on this topic and give a recommendation how observational uncertainty can be considered in the method, the formulas presented in section 4.1 and the code.

We agree with the reviewer that text could be added discussing the problem of observational uncertainty. However, no single best solution exists for this problem. Before starting the calibration (i.e. ensemble selection) exercise, one should first identify which observational products can be trusted (for the specific region, variable, time period in mind).
The discussion is actually similar to the reviewer’s question 5 (for multiple observational products instead of variables). In this study, we presented a different optimal subset for each chosen observational product. Alternatively, one could of course put multiple observational products into a single cost function and end up with a single optimal subset. However, when using ensembles for inference, then a lot can be learned about predictability from the differences between using different observational products. This additional uncertainty added by observations is ignored if all the products are combined in a single cost function.

As for Q5, one could also end up with a pareto front across different products, where we have a whole range of subsets rather than a single best one. This is something that is worth investigating in a future study. The following text has been added (Section 4.1, “Choice of observational product”):

“This could be done by putting multiple observational products into a single cost function. However, when using ensembles for inference, a lot can be learned from the spread across observational products. This additional uncertainty added by observations is ignored if all the products are combined in a single cost function.”

11. Section 4.2 applies the method to the future, keeping the limited sample of weighting the ensemble based on temperature means / trends. A model could simulate a correct present-day climatology but why would it be a good model to project future climate? One of the authors convincingly shows that there is hardly any correlation between present-day and future temperature patterns [Knutti et al., 2010]. Climate change is non-linear. Could the authors choose a multivariate and preferably process-oriented diagnostic approach? Otherwise, please limit general statements for the applicability of this method to improve projections (see above).

In order to test if the subset has skill in the future (we call it out-of-sample, as we do not have observations), we conducted model-as-truth experiments. From that we learned that our optimal subset does not always improve projections relative to the simple multi-model mean, especially when optimizing for the trend (Fig. 4d). When optimizing for the climatology however, we observe an improvement of more than 10% out-of-sample. This suggests that we are not simply fitting noise, but actually gaining from the subset selection. If there was no signal in the present-day climatology, we would not have obtained an improvement out-of-sample.
We agree with the reviewer that correlations between present-day and future temperature patterns are weak (see also our supplementary figure S5). Finding a good emergent constraint is exactly what is needed to find an optimal subset with skill out-of-sample. Regional biases seem to persist, which is why we found improved out-of-sample skill in some cases.

We commented on the idea of using process-oriented diagnostics at Q8 (above).

We added a few sentences at the beginning of Section 4.2.1 to better motivate the need for model-as-truth experiments.

“Is a model that correctly simulates the present-day climatology automatically a good model for future climate projections? To answer this question, we need to investigate if regional biases persist into the future, and determine whether the approach is fitting short term variability. This is done by conducting model-as-truth experiments.”

Minor Comment: There seems to be a mistake how papers are cited as they are missing ‘et al.’

We have adjusted the bibliography so that the “et al.” are now shown in the revised version of the manuscript.