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.”

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 the 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.

Selecting a climate model subset to optimise key ensemble properties

Nadja Herger[1], Gab Abramowitz[1], Reto Knutti[2,3], Oliver Angélil[1], Karsten Lehmann[4], and Benjamin M. Sanderson[3]

[1]Climate Change Research Centre and ARC Centre of Excellence for Climate System Science, UNSW Sydney, Sydney NSW 2052, Australia
[2]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[3]National Center for Atmospheric Research, Boulder, Colorado, USA
[4]Satalia, Berlin, Germany

Correspondence to: Nadja Herger (nadja.herger@student.unsw.edu.au)

Abstract. End-users studying impacts and risks caused by human-induced climate change are often presented with large multi-model ensembles of climate projections whose composition and size are arbitrarily determined. An efficient and versatile method that finds a subset which maintains certain key properties from the full ensemble is needed, but very little work has been done in this area. Therefore, users typically make their own somewhat subjective subset choices and commonly use the equally-weighted model mean as a best estimate. However, different climate model simulations cannot necessarily be regarded as independent estimates due to the presence of duplicated code and shared development history.

Here, we present an efficient and flexible tool that makes better use of the ensemble as a whole by finding a subset with improved mean performance compared to the multi-model mean while at the same time maintaining the spread and addressing the problem of model interdependence. Out-of-sample skill and reliability are demonstrated using model-as-truth experiments. This approach is illustrated with one set of optimisation criteria but we also highlight the flexibility of cost functions, depending on the focus of different users. The technique is useful for a range of applications that, for example, minimise present day bias to obtain an accurate ensemble mean, reduce dependence in ensemble spread, maximise future spread, ensure good performance of individual models in an ensemble, reduce the ensemble size while maintaining important ensemble characteristics, or optimize several of these at the same time. As in any calibration exercise, the final ensemble is sensitive to the metric, observational product and pre-processing steps used.

1 Introduction

Multi-model ensembles are an indispensable tool for future climate projection and the quantification of its uncertainty. However, due to a paucity of guidelines in this area, it is unclear how best to utilise the information from climate model ensembles consisting of multiple imperfect models with a varying number of ensemble members from each model. Heuristically, we understand that the aim is to optimise the ensemble performance and reduce the presence of duplicated information. For such an optimisation approach to be successful, metrics that quantify performance and duplication have to be defined. While there are examples of attempts to do this (see below), there is little understanding of the sensitivity of the result of optimisation to the subjective choices a researcher needs to make when optimising.

As an example, the equally-weighted multi-model mean (MMM) is most often used as a "best" estimate for variable averages (Knutti et al., 2010), as evidenced by its ubiquity in the Fifth Assessment Report of the United Nations Intergovernmental Panel on Climate Change (IPCC, 2014). In most cases, the MMM – which can be regarded as an estimate of the forced climate response – performs better than individual simulations. It has increased skill, consistency and reliability (Reichler and Kim, 2008; Gleckler et al., 2008) as errors tend to cancel (Knutti et al., 2010), although part of that effect is the simple geometric argument of averaging (Annan and Hargreaves, 2011). However, model democracy ("one model, one vote") (Knutti, 2010) does not come without limitations. A lack of independence in contributions to the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor et al., 2012) archive (Masson and Knutti, 2011; Knutti et al., 2013), where research organisations simply submit as many simulations as they are able to (thus often referred to as "ensemble of opportunity" (Tebaldi and Knutti, 2007)), means that it is extremely unlikely that the MMM is in any way optimal. Different research groups are known to share sections of code (Pincus et al., 2008), literature, parametrizations in their models, or even whole model components, so that at least heuristically, we understand that individual model runs do not necessarily represent independent projection estimates (Abramowitz, 2010; Abramowitz and Bishop, 2015; Sanderson et al., 2015a). Ignoring the dependence of models might lead to a false model consensus, poor accuracy and poor estimation of uncertainty.

Instead of accounting for this dependence problem, most studies use whatever models and ensembles they can get and solely focus on selecting ensemble members with high individual performance (e.g., Grose et al. (2014)). They assume that if individual members of an ensemble perform well, then the mean of this ensemble will also have high skill. As we demonstrate later, this is not always the case, and can potentially be highly problematic.

Given that climate models developed within a research group are prone to share code and structural similarities, having more than one of those models in an ensemble will likely lead to duplication of information. Institutional democracy as proposed by Leduc et al. (2016) can be regarded as a first proxy to obtain an independent subset. However, in this case dependence essentially reflects an *a priori* definition of dependence that may not be optimal for the particular use case (e.g., variable, region, metric, observational product). There are also a few cases in which a model is shared across institutes and thus this approach would fail (e.g., NorESM is built with key elements of CESM1), or at least need to evolve over time.

Only a few studies have been published that attempt to account for dependence in climate model ensembles. A distinction can be made between approaches that select a discrete ensemble subset and those that assign continuous weights to the ensemble members. For example, Bishop and Abramowitz (2013) proposed a technique in which climate model simulations undergo a linear transformation process to better approximate internal climate system variability, so that models and observations were samples from a common distribution. This weighting and transformation approach was based on a mean square difference adherence to an observed product over time and space within the observational period, with ensemble spread at an instant in time calibrated to estimate internal variability. The same process was also used for future projections, with the danger of over-fitting 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. This resulted in a metric to measure the distance between models, as well as models and observational products, in a projected model space (Abramowitz et al. (2008) is another example of an attempt to do this). Knutti et al. (2017) aims to simplify the approach by Sanderson et al. (2015a), where models which poorly agree with observations are down-weighted, as are very similar models that exist in the ensemble, based on RMSE distance. Projections of the Arctic sea ice and temperatures are provided as a case study. Perhaps not surprisingly, the effect of weighting the projections is substantial, and more pronounced on the model spread than its best estimate.

Sanderson et al. (2015b) proposes a method that finds a diverse and skillful subset of model runs that maximises inter-model distances, using a stepwise model elimination procedure. Similar to Sanderson et al. (2015a), this is done based on uniqueness and model quality weights.

~~?~~ Sanderson et al. (2017) applied a similar continuous weighting scheme to climatological mean state variables and weather extremes in order to constrain climate model projections. Only a moderate influence of model skill and uniqueness weighting on the projected temperature and precipitation changes over North America was found. As under-dispersion of projected future climate is undesirable, only a small reduction in uncertainty was achieved.

In the previous paragraph we discussed approaches that assign continuous weights to model runs. 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. 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. Such an approach is presented in Evans et al. (2013), where independence was identified to be central for creating smaller ensembles.

The problem of defining and accounting for dependence is made more challenging by the fact that there is no uniformly agreed definition of dependence. A canonical statistical definition of independence, that two events A and B are considered to be independent if the occurrence of B does not affect the probability of A, P(A), so that P(A|B)=P(A). As discussed by Annan and Hargreaves (2017), there could, however, be many approaches to applying this definition to the problem of ensemble projection that could potentially yield very different results. An appropriate course of action regarding what to do if two models are identified to be co-dependent does not follow directly from this usual definition of independence.

One disadvantage of many of these studies is that they are technically challenging to implement and therefore discourage frequent use. Further, the sensitivity of each approach to the choice of metrics used, variables included and uncertainties in observational products is largely unexplored. This leads to a lack of clarity and consensus on how best to calibrate an ensemble for a given purpose. Often, out-of-sample performance has not been tested, which we consider essential when looking at ensemble projections.

~~Here, we~~ 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 ~~effective way.~~ feasible way. Flexibility is introduced by an adjustable cost function which is allowing this approach to be applied to a wide range of problems. The meaning of "optimal" ~~can vary~~ varies depending on the aim of the study. As an example, we will choose a subset of the CMIP5 archive that minimises regional biases in present day climatology, based on RMSE over space using a single observational product. The resulting ensemble subset will be optimal in the sense that its ensemble mean will give the lowest possible RMSE against this observational product of any possible combination of model runs in the archive. The more independent estimates we have, the more errors tend to cancel. This results in smaller biases in the present day which reduces the need for bias correction. 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. More precisely, it is the number of zero weight that leads to some models being discarded from the ensemble. Out-of-sample skill of the optimal subset mean and spread is tested using model-as-truth experiments. The distribution of projections using model runs in the optimal subset is then assessed.

We then examine the sensitivity of this type of result to choices of the cost function (by adding additional terms), variable and constraining data set. We argue that optimally selecting ensemble members for a set of criteria of known importance to a given problem is likely to lead to more robust projections for use in impact assessments, adaptation and mitigation of climate change.

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.

In the next section, we introduce the model data and observational products used for this study. Section 3 contains a description of the method used, which includes the pre-processing steps and three sub-sampling strategies, one of which is the novel approach. In section 4 we examine the results by first giving the most basic example of the optimisation problem. We then expand on this example by highlighting the method's flexibility and before applying the novel approach to the future, we test out-of-sample skill with model-as-truth experiments to ensure that our approach it not overfitting on the current present-day state. Finally, section 5 contains the discussions and conclusions.

2 Data

We use 81 CMIP5 model runs from 38 different models and 21 institutes which are available in the historical period (1956–2013; RCP4.5 after 2005) and RCP4.5, RCP8.5 period (2006–2100) (see Table 1 in the Supplementary Information (SI)). We examine gridded monthly surface air temperature (variable: tas) and total monthly precipitation (variable: pr). Results shown here are based on raw model data (absolute values), although repeat experiments using anomalies (by subtracting the global mean climatological value from each grid cell) were also performed (not shown here).

Multiple gridded observation products for each variable were considered with each having different regions of data availability (see Table 2 and additional results in the SI). Model and observation data were remapped using a first order conservative remapping procedure (Jones, 1999), to either 2.5° or 5° spatial resolution, depending on the resolution of the observational product (see SI Table 2). For the projections, the model data was remapped to a resolution of 2.5°. For observational products whose data availability at any grid cell changes with time, a minimal two-dimensional mask (which does not change over time) was used. The remaining regions were masked out for both the observational product and the model output.

3 Method

We first illustrate the technique by considering absolute surface air temperature and total precipitation climatologies (time-means at each grid-cell), based on 1956–2013. The land-only observational product CRUTS, version 3.23 (Harris et al., 2014) is used for both variables and model data is remapped to the same spatial resolution and masked based on data availability in this product.

Next, we select an ensemble subset of size K from the complete pool of 81 CMIP5 simulations, using three different approaches:

Random ensemble: As the name implies, the random selection consists of randomly selected model runs from the pool of 81 without repetition. This procedure is repeated 100 times for each ensemble size in order to gauge sampling uncertainty.

Performance ranking ensemble: This ensemble consists of the “best” performing model runs from the ensemble in terms of their RMSE (based on climatology — time means at each grid-cell). Individual model runs are then ranked according to their performance and only the best K model runs are chosen to be part of the subset.

Optimal ensemble: In this case we find the ensemble subset whose mean minimises RMSE, out of all possible K-member subsets. This is non-trivial – there are $2.12 \cdot 10^{23}$ possible ensembles of size 40, for example, so that a “brute-force” approach is simply not possible. Instead, we use a state-of-the-art mathematical programming solver (Gurobi (2015)). It minimises the MSE between the mean of K model runs and the given observational product, by selecting the appropriate K model runs. Hereinafter we refer to the ensembles (one obtained for each K) derived from this approach as "optimal ensembles" and the optimal ensemble with the overall lowest RMSE as the "optimal subset". ~~The~~ Note, that optimal refers to the specific question at hand that the ensemble is calibrated to. The ensemble would no longer be optimal if the specific application changes. The problem itself is a mixed integer quadratic programming problem because the decisions are binary (that is: model run is in the set or not), the cost function is quadratic (see Eq. (1)), and the constraint is linear. Such a problem is solved using a branch-and-cut algorithm (Mitchell, 2002).

In the following section, we compare these three subsampling strategies with the benchmark, which is the simple unweighted multi-model mean (MMM) of all 81 runs. We then examine the sensitivity of results to the observational product, ~~cost function~~ the cost function (to demonstrate flexibility by optimising more than just the ensemble mean) and other experimental choices.

4 Results

Figure 1 displays the area-weighted root mean square error (RMSE) of the subset mean and RMSE improvement relative to the MMM of all 81 model runs (solid horizontal line) as a function of ensemble size for the three different methods used to select subsets. The RMSE is calculated based on the climatological fields of pre-processed model output and observations. Results based on CRUTS3.23 as the observational product are shown for both surface air temperature (a) and precipitation (b). We focus on panel (a) for now. Each marker represents the RMSE of an ensemble mean, except for ensemble size one, which refers to the single best performing model run in terms of RMSE. Blue markers are used for the random ensemble, with the error bar indicating the 90% confidence interval (from 100 repetitions). The performance ranking ensemble is shown in green. For ensemble sizes one to four, the RMSE of the performance ranking ensemble increases. This is because multiple initial condition ensemble members of the same model (MPI-ESM) are ranked high, and averaging across those leads to higher dependence within the subset and thus less effective cancelling out of regional biases. Interestingly, the performance-based ensemble sometimes even performs worse than the mean of the random ensemble, which can be observed across multiple observational products and across the two variables (see SI). This is a clear example of the potential cost of ignoring the dependence between ensemble simulations. Selecting skillful but similar simulations can actively degrade the present-day climatology of the ensemble mean.

For the optimal ensemble (black circles), RMSE is initially large, the value representative of the single best performing model run (black dot being behind the green one). The RMSE of the ensemble mean rapidly decreases when more model runs are included until it reaches a minimum (red circle indicates the optimal subset over all possible ensemble sizes). That is, the RMSE improvement relative to the MMM (solid horizontal line) is largest at this ensemble size. As more model runs are included in the ensemble, the RMSE increases again. This is expected as worse performing and more dependent model runs are forced to be included. The MMM generally outperforms every individual model run (green, black and blue dots at subset size one being above the solid horizontal line). The optimal ensemble curve in the vicinity of the lowest RMSE is often rather flat, so different ensembles with similarly low RMSE could be chosen instead if, for example, a given model is required to be part of the subset. A flat curve is also of advantage in the case when computational resources are limited and thus a small ensemble size has to be chosen (for example when global model boundary conditions are being chosen for a downscaling experiment). Here, however, we always consider the ensemble with the overall smallest RMSE (red circle) as our optimal subset even if ensembles of similar sizes are not much worse. We will discuss the black triangle markers and other horizontal lines in a later section.

Of the three sub-sampling approaches, it is evident that the optimal ensemble mean is the best performing one for all ensemble sizes if the bias of the model subset average should be minimized – essentially indicating that the solver is working as anticipated. Regional biases in different models cancel out most effectively using this approach. Across different observational products, we observe an improvement in RMSE relative to the MMM of between 10–20% for surface air temperature, and around 12% for total precipitation (see Figure S1 and S2). The size of the optimal subset is significantly smaller than the total number of model runs considered in this study (see red text in Figure 1). For surface air temperature we obtain an optimal ensemble range between six to ten members and for precipitation around twelve members. This suggests that many model runs in the archive are very similar.
We achieve similar RMSE improvement if we exclude closely related model runs *a priori* and start off with a more independent set of model runs (one model run per institute), see Figure S3.

Figure 1 solely looks at the performance of the ensemble mean. A characterisation of the relationship between model simulation similarity and performance in these ensembles is shown in Figure 2. Simulation performance (in terms of RMSE) is plotted against the simulation dependence (expressed as average pairwise error correlation across all possible model pairs in the ensemble) for the three sampling techniques (3 colors). As before, CRUTS3.23 was used as the observational product, but this figure looks very similar across different variables and observational products. Circular markers are used for the average of individual members of the subset ensemble of any given size and diamond markers are used for ensemble mean. The darker the color, the larger the ensemble size. Members of the optimal ensemble (black markers) are more independent than members of other ensembles, at least in terms of pairwise error correlation. Members of the performance-ranking ensemble (green markers) however show high error correlations as closely related model runs are likely to be part of the ensemble. We thus conclude that the optimal ensemble has favourable properties in terms of low ensemble mean RMSE and low pairwise error correlation of their members. We will therefore focus on this the ability of this sampling technique for the remainder of the paper.

4.1 Sensitivity of results

We now develop this optimisation example to highlight the flexibility of the method. In doing so, it should become clear that calibration for performance and dependence is necessarily problem dependent. A graphical representation of the experimental choices we explore is shown in Figure 3. We explore different aspects of this flowchart below.

Choice of observational product. The ensembles in the previous subsection were calibrated on a single observational product (depicted in green in Figure 3). Observational uncertainty can be quite large depending on the variable and can thus result in a different optimal subset. Figure 1 for different observational products (and varying observational data availability) can be found in the supplementary material (Figure S1 and S2). Moreover, observational uncertainty within one observational product (instead of across the products) should also be considered to test the stability of the optimal subset. This has not been done here, but could certainly be investigated in future studies. Lastly, if multiple observational products per variable are available and all equally credible, finding a subset that is optimal using all of them is also a possibility. 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.

Here, we only optimise our ensemble to one observational product at a time and investigate how sensitive the optimal subset is to that choice.

Variable choice. The selection of the variable has a profound influence on the resulting optimal subset. This was already briefly highlighted in Figure 1, where the optimal subsets for surface air temperature (a) and total precipitation (b) consist of rather different ensemble members. Generally, the optimal ensemble size for precipitation tends to be larger. Similar to the discussion above for the sensitivity to observational products, one might consider optimising the subset across multiple variables. This is particularly important if physical consistency across variables needs to be ensured. 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 the fact that the optimal subsets for the individual variables potentially look very different. One might calibrate the ensemble on multiple variables using a ~~pareto~~ Pareto solution set, similar to what has been done in Gupta et al. (1998) for hydrological models and Gupta et al. (1999) for land surface schemes. An important characteristic of such a problem is that it does not have a unique solution, as there is a trade-off between the different and non-commensurate variables. When improving the subset for one variable (i.e., RMSE reduced), we observe a deterioration of the subset calibrated on the other variable.

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

Absolute values vs. anomalies. Results presented in this study are all based on absolute values rather than anomalies. Whether or not bias-correction is required depends on the variable and the aim of the study. To study the Arctic sea ice extent for example, absolute values are a natural choice as there is a clear threshold for near ice-free September conditions. An example of where bias-correction is necessary is in the field of extreme weather. For example, mean biases between datasets must be removed before exceedance probabilities beyond some extreme reference anomaly can be calculated.

Alternatives to climatology. As part of the data pre-processing step, we computed climatologies for the model output and observational dataset. In addition to climatologies (time-means at each grid cell), we will later look at linear trends and 10-year running means (hereafter referred to as “space+time”). Subsection 4.2.1 shows (based on a model-as-truth experiment) how sensitive the ensemble can be to the quantity of a variable (mean, trend, or variability) chosen in pre-processing.

Defining the benchmark. To assess whether our optimal subset has improved skill, we need to define a benchmark. In Figure 1, we used the MMM of 81 model runs as our benchmark (solid line). However, other benchmarks could be used. The three horizontal lines in Figure 1 refer to three different baselines that could be used to compare against subset performance. The solid line is the MMM of all available model runs. For the dashed line, we first aggregated across the ensemble members from each climate model and then average across all 38 models. The dotted line is the ensemble mean when only allowing one run per institute to be part of the ensemble. Interestingly, the dotted line is very often the highest one and the solid line has the lowest RMSE. One likely explanation is that the original CMIP5 archive is indirectly already slightly weighted due to a higher replication of well-performing models (Sanderson et al., 2015b). By eliminating those duplicates, our ensemble mean gets worse because regional biases do not cancel out as effectively. For the model-as-truth experiment described in subsection 4.2.1, our benchmark was also obtained by selecting one model run per institute.

Sensitivity to the underlying cost function. An essential part of the optimisation problem is the cost function. Comparison of all the sensitivities mentioned above is made possible only because our subsets are truly optimal with respect to the prescribed cost function. For the results above the cost function $f(x)$ being minimised by the Gurobi solver was:

$$f(x) = f_1(x) = MSE\Bigg(\Big(\frac{1}{|x|}\sum_{i\in x} m_i\Big), y\Bigg). \qquad (1)$$

Here, x denotes the optimal subset (with $|x|$ being the subset size), y is the pre-processed observational product and m_i is model simulation i. MSE stands for the area-weighted mean squared error 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.

Of course this cost function can and should be adjusted depending on the aim of the study, as long as the expressions are either linear or quadratic. To illustrate this idea, we add two new terms to the cost function above that account for different aspects of model interdependence:

$$f(x) = \frac{f_1(x) - \mu_1}{\sigma_1} + \frac{f_2(x) - \mu_2}{\sigma_2} - \frac{f_3(x) - \mu_3}{\sigma_3} \tag{2}$$

Here, minimising $f(x)$ will involve minimising the first and second terms in Equation (2) and maximising the third term (note the minus sign in front of term 3). To ensure that the three terms all have a similar magnitude and variability, we subtract the mean (μ) and divide by the standard deviation (σ) derived from 100 random ensembles of a given ensemble size.

The function $f_1(x)$ is the same as in Eq. (1). It minimises the MSE between the subset mean of a given size and the observational product y. The second and third terms can be written as follows:

$$f_2(x) = \frac{1}{|x|} \sum_{i \in x} MSE(m_i, y) \tag{3}$$

$$f_3(x) = \frac{2}{|x| \cdot (|x| - 1)} \sum_{i \neq j \in x} \frac{MSE(m_i, m_j)}{\frac{1}{2}\Big(MSE(m_i, y) + MSE(m_j, y)\Big)} \tag{4}$$

The function $f_2(x)$ in the second term ensures that the mean MSE between each ensemble member and the observational product is minimised. So, this term is related to the performance of the individual ensemble members — we want to avoid very poorly performing members being in the final ensemble. It would of course also be possible to make an a priori decision on which models should be considered before starting the optimisation process. The function $f_3(x)$ averages the pairwise MSE distances between all ensemble members and then divides by the mean performance. This should be maximised and helps to avoid clustering by ensuring that the ensemble members are not too close to each other relative to their distance to the observational product. This is a way to address dependence in ensemble spread. ~~?~~ Sanderson et al. (2017) used a similar idea of calculating pairwise area-weighted root mean square differences over the domain to obtain an inter-model distance matrix. This matrix is then normalised by the mean inter-model distance to obtain independence weights as a measure of model similarity.

Based on the climatological metric, Gurobi can solve Eq. (2) within a few seconds for any given subset size. Finding an optimal solution without this solver would have been impossible within a reasonable amount of time. Results show that the RMSE of the optimal ensemble mean based on eq. (2) is almost as low as for eq. (1), see Figure 1 (black circles for eq. (1) and triangles for eq. (2)). However, the members of the optimal ensemble seem to have a better average performance and are slightly more independent. This might be of advantage if end users want to avoid having multiple ensemble members from the same model in the optimal subset. Term 3 in Eq. (2) will take care of that. Moreover, term 2 will make sure that bad performing model runs are excluded from the optimal subset. In other words, explicitly considering single model performance and eliminating obvious duplicates does not significantly penalize the ensemble mean performance. The magnitude of the three terms in eq. (2) as a function of the ensemble size are shown in Figures S8 and S9.

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 et al. (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)).

For those concerned about overconfidence of the ensemble projections (due to the "unknown unknowns"), one could add another term which maximises future spread. This would result in an ensemble which allows to explore the full range of model responses. It is also possible to start weighting the terms of the cost function differently depending on what seems more important.

4.2 Application to the future

4.2.1 Testing out-of-sample skill

The optimal selection approach is clearly successful at cancelling out regional biases in the historical period~~. To~~, where observations are available. We refer to this period as "in-sample". 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~~are conducted~~. This should give an indication of whether sub-selecting in this way is likely to improve future predictability or if we are likely to be overconfident with our subset. 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 exactly where model-as-truth experiments come into play. For this purpose, one simulation per institute is considered to be the "truth" as though it were observations, and then the optimal subset from the remaining 20 runs (one-per-institute) is determined for the in-sample period (1956–2013), based on the cost function in Eq. (1). The optimal ensemble's ability can then be tested in the out-of-sample 21st century, since we now have "observations" for this period. Results are then collated over all possible simulations playing the role of the "truth". In all our model-as-truth experiments, near relatives were excluded as truth, because members from the same model are likely to be much closer to each other than to the real observational product. This subscription to institutional democracy is consistent with what was found by Leduc et al. (2016) to prevent overconfidence in climate change projections. ~~?~~ Sanderson et al. (2017) also removed immediate neighbours of the truth model from the perfect model test when deriving the parameters for their weighting scheme.

Figure 4 shows the results of the model-as-truth experiment for surface air temperature for the climatological field, the linear trend and space + time, as described above. Panel (a) shows global absolute mean temperature time series for the in- and out-of-sample periods. The in-sample period, in which the optimal subset is found for each model as truth is 1956–2013. For the climatological metric and the space + time metric, the same subset was tested out-of-sample in 2071–2100 using the same truth as in the in-sample period. The out-of-sample period for the trend metric is 2006–2100, as 30 years are not long enough to calculate a linear trend at each grid-cell without internal variability potentially playing a big role. Both in- and out-of-sample data undergo the same pre-processing steps. The mask which was used for those calculations is shown in the lower right corner of panel (a).

Figures 4b–d show the RMSE improvement of the optimal subset for a given size relative to the mean of all remaining 20 simulations for each simulation as truth. The black curve is the in-sample improvement and the blue curve is the out-of-sample improvement for RCP8.5 averaged across all truths. The shading represents the spread around the mean. Results for RCP4.5 look very similar and are therefore not shown here.

It is evident that both the climatological metric and the space + time metric have improved skill out-of-sample compared to simply taking the mean of all available runs. We observe an RMSE improvement almost as big as the in-sample improvement, in which we conducted the optimisation. This primarily shows the persistence of the climatological bias. 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.

The trend metric is different, however. To be clear, here we obtain the optimal subset based on a two-dimensional field with linear (58-year) trends at each grid-cell in the in-sample period. We then use this subset trained on trend values to predict the out-of-sample trend field (using the same simulation as "truth" as in the in-sample period). The RMSE improvement presented in panel (d) is calculated from the "true" RCP8.5 trend field and the predicted trend derived from the optimal subset. We see a large in-sample improvement, but out-of-sample this skill quickly disappears. We thus conclude that the magnitude and nature of trends within individual models do not persist into the future and a subset based on this metric will not have any improved skill out-of-sample. Figure S5 shows the very weak correlation between in- and out-of-sample trend very clearly. This highlights the difficulty of finding an appropriate metric which constrains future projections. Results for precipitation can be found in the SI (Figure S6).

Figure 4 shed light on the increased skill of the optimal ensemble compared to the simple MMM, at least for the mean signal. We have not yet investigated the spread of the ensemble, which is as least as important, especially for impact and risk related fields. As an example, the potential danger of having a too narrow ensemble spread (overconfident projections) by neglecting important uncertainties is highlighted in Keller and Nicholas (2015).

Results for the ensemble spread are shown in Figure 5 for surface air temperature. Panel (a) explains how the spread of the ensemble is quantified. We calculate how often the truth lies within the 10th to 90th percentile of the optimal ensemble for a given ensemble size. We derive the percentiles from a normal distribution, whose mean and standard deviation were calculated from the optimal ensemble (for a given truth and ensemble size) during the in-sample, or training period. This is done for every grid cell and each model as truth. The curves shown in Figures 5b–d are the average of the fractions of "truth" values that lie within this range, across all grid cells and truths plotted against the subset size for the climatological field (b), the space + time (c) and linear trend (d). We would expect the truth to lie within the 10th to 90th percentile of the ensemble at least 80% of the time to avoid overconfidence. Black is used for the in-sample fraction and the two shades of blue for RCP4.5 (light blue) and RCP8.5 (dark blue). The fraction for an ensemble consisting of all 20 model runs — the benchmark in this case — is shown with a horizontal line. The ensembles obtained based on the climatological metric and the space + time metric are slightly over-dispersive both in- and out-of-sample, which suggests the optimal ensemble should not result in overconfidence in ensemble spread, relative to the entire ensemble. An ensemble that is overconfident can lead to projections whose uncertainty range is too narrow and thus misleading. This is the case for the trend metric, at least for smaller ensemble sizes.

Such a model-as-truth experiment can also assist with the choice of an optimal subset size for the application to projections. It does not necessarily have to be the same as the in-sample ensemble size, as aspects like mean skill improvement and reduction of the risk of underdispersion have to be considered.

Can a subset calibrated on absolute historical temperature constrain temperature *changes* in the future, as opposed to just minimising bias in the ensemble mean? This anomaly skill in the out-of-sample test is depicted in Figure 6. The setup is similar to Figure 4, but here we are predicting regional temperature change from mean values in 2006–2035 to those in 2071–2100. The optimal subset is still derived using either the climatological (b), space + time (c), or trend metric (d). The only thing that has changed is what is being predicted is now out-of-sample. The curves are the RMSE improvement relative to the MMM of 20 model runs averaged across all truths for RCP4.5 (light blue) and RCP8.5 (dark blue). Shading indicates the spread (one standard deviation) across the different truths. Results for regional precipitation change are shown in Figure S7.

Panel (b) shows that there is very little to be gained by constraining the climatology in terms of out-of-sample skill. Across all metrics and variables, the subsets show hardly any RMSE improvement compared to the MMM of the 20 model runs, which is consistent with ~~?~~Sanderson et al. (2017). This result is partly about the discrepancy between the metric used to derive the optimal ensemble and that used to evaluate it, and reinforces how sensitive this type of calibration exercise is to the somewhat subjective choices faced by a researcher trying to post-processes climate projections. It is an important limitation that should be kept in mind when using this sampling strategy to constrain future projections.

4.2.2 Projections

In earlier sections we presented results based on a single observational product per variable. However, the importance of the choice of product should not be neglected. The influence of obtaining an optimal subset based on different observational products can be visualised with maps. To create Figure 7, the temperature change between the 2081–2100 and 1986–2005 climatologies was calculated for the RCP8.5 scenario using the mean of all 81 model runs. Then, the temperature change of the optimal subset (based on a given observational product), calculated in the same way, was subtracted. The result is a map that shows the difference the optimal sampling makes to projected temperature changes. Maps are shown for the optimal subsets derived from different observations, with grey contours highlighting the area used to derive the subset. The number in brackets refers to the size of the optimal subset. Despite the maps looking quite different, we can identify some regions with consistent changes. The Southern Ocean is consistently warmer in the optimal subset and the Arctic is colder than the MMM (except for BEST, global). Generally, the optimal subset results in a cooler land surface.

Figure 8 shows the same as 7 but for precipitation change based on three different observational products. They all show an increase in precipitation in the equatorial Pacific and the western Indian Ocean and a decrease over Indonesia.

5 Discussion and conclusions

We presented a method that selects a CMIP5 model subset which minimises a given cost function in a computationally efficient way. Such a calibrated smaller ensemble has important advantages compared to the full ensemble of opportunity, in particular reduced computational cost when driving regional models, smaller biases in the present day which reduce the need for bias correction, reduced dependence between the members and sufficient spread in projections. The cost function can be varied depending on the application. The simplest cost function presented here simply minimises biases of the ensemble mean. We have shown that this method accounts to some degree for the model dependence in the ensemble by the way it optimizes the ensemble mean, but closely related models or even initial condition ensemble models of the same models are not penalized and can still occur. This optimal subset performs significantly better than a random ensemble or an ensemble that is solely based on performance. The performance ranking ensemble sometimes even performs worse than the random ensemble in its mean, even though of course the individual models perform better. Depending on the application, one of the other will matter more.

We also illustrated the expansion of the cost function to optimise additional criteria, enabling an optimal subset that minimises the ensemble mean bias, the individual model biases, and the clustering of the members, or any combination thereof. One could also, for example, add a term that maximises the ensemble projection spread to avoid overconfidence. The choice of what is constrained by the cost function clearly depends on the aim of the study (e.g., present day bias, dependence issue, future spread). We highlight the importance of testing the sensitivity to the metric and observational product (incl. varying data availability) used, as they can lead to quite different results.

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).

Model-as-truth experiments were used to investigate the potential for overconfidence, estimate the ensemble spread, and test the robustness of emergent constraints. Based on those experiments we learned that absolute present day values constrain absolute values in the future (due to a persistent bias). However, absolute present day values do not constrain projected changes relative to a present day state.

There were other pertinent questions we did not address, of course. These include the question of how best to create an optimal subset across multiple variables and gridded observational products. This seems especially important if physical consistency across variables should be maintained. Having a pareto set of ensembles (by optimising each variable separately) rather than a single optimal subset is a potential solution, but is clearly more difficult to work with.

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.

Many of the points raised here are also clearly not restricted to global climate models. The same holds for regional climate models, hydrological models or perhaps ecological models. We encourage others to apply the same approach to different kinds of physically based models.

Critically, we wish to reinforce that accounting for dependence is essentially a calibration exercise, whether through continuous or discrete weights, as was the case here. Depending on the cost function, the data pre-processing and the observational product one can end up with a differently calibrated ensemble. Depending on the application, bias-correction of the model output might be appropriate before executing the calibration exercise. We suggest that the approach introduced in this study is an effective and flexible way to obtain an optimal ensemble for a given specified use case.

Future research will help to provide confidence in this method and enable researchers to go beyond model democracy or arbitrary weighting. This is especially important as replication and the use of very large initial condition ensembles will likely become a larger problem in the future global ensemble creation exercises. An approach that attempts to reduce regional biases (and therefore indirectly dependence) offers a more plausible and justifiable projection tool than an approach that simply includes all available ensemble members.

6 Code availability

A simplified and easily-adjustable Python code (based on the Gurobi interface) is accessible on a GitHub repository (https://github.com/nherger/EnsembleSelection/blob/master/Gurobi_MIQP_random.py). Gurobi is available via a free academic license.

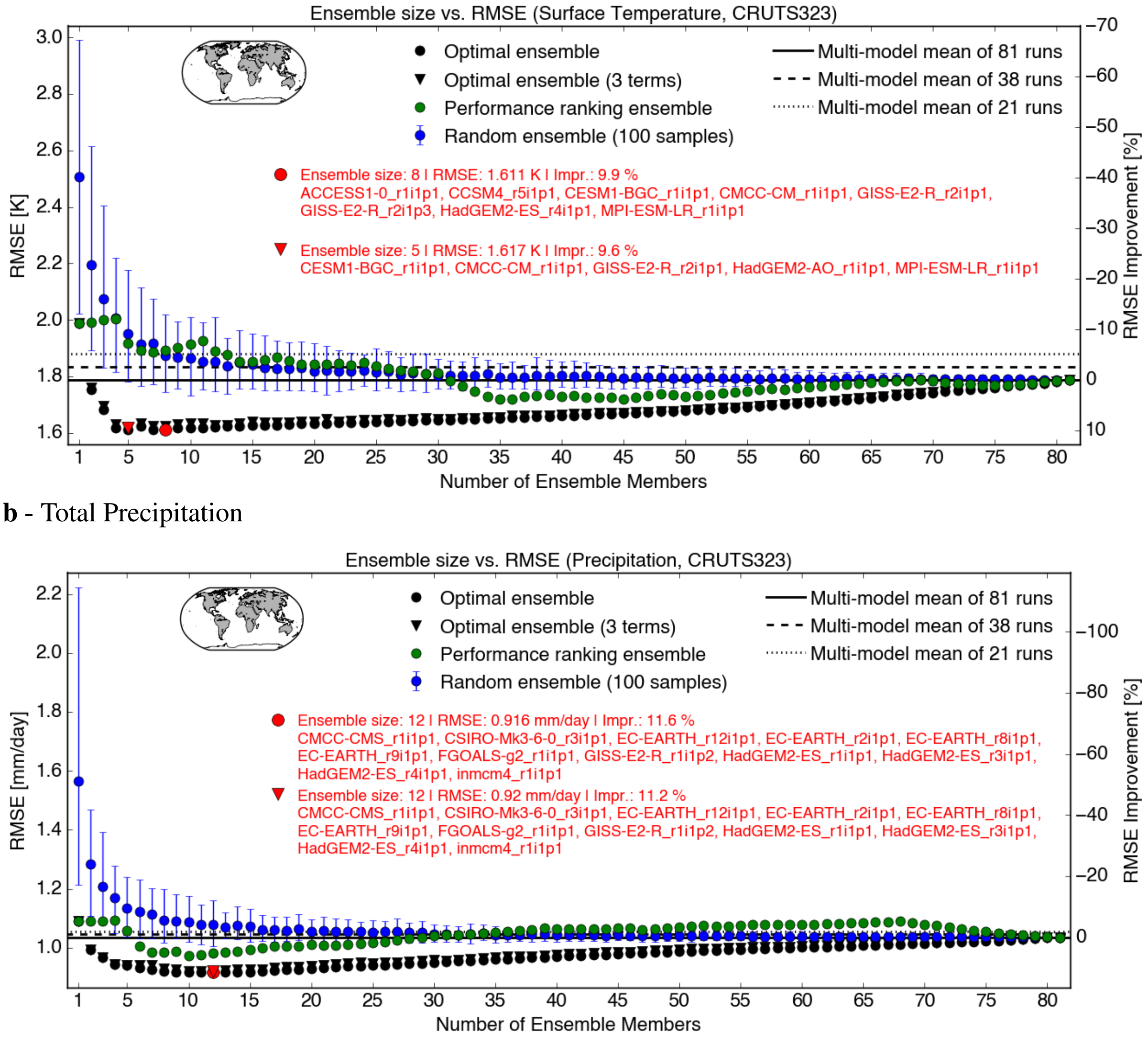
7 Data availability

CMIP5 data can be obtained from http://cmip-pcmdi.llnl.gov/cmip5/.

Author contributions. N. Herger conducted the analysis, produced the figures and prepared the manuscript. G. Abramowitz came up with the core idea of ensemble selection to minimise regional biases, discussed results and helped writing the manuscript. R. Knutti contributed to discussions on the methodology and results and helped writing the manuscript. O. Angélil helped shape the methodology and contributed to the interpretation of results. K. Lehmann provided support while writing the Python code for the mathematical solver Gurobi. B. Sanderson provided useful discussions and feedback which helped shape this work.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We would like to thank Jan Sedláček for providing access to the next generation CMIP5 archive based at ETHZ. We are also grateful to Steve Sherwood and Ruth Lorenz for interesting discussions which helped shape this study.
We acknowledge the support of the Australian Research Council Centre of Excellence for Climate System Science (CE110001028).
The authors acknowledge the support from the H2020 project CRESCENDO "Coordinated Research in Earth Systems and Climate: Experiments, kNowledge, Dissemination and Outreach", which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 641816.
We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table S1 in the SI) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.



a - Surface Air Temperature

b - Total Precipitation

Figure 1. Size of the CMIP5 subset on the horizontal axis and the resulting RMSE of the ensemble mean and its improvement relative to the multi-model mean (MMM) on the vertical axes, for surface air temperature (**a**), total precipitation (**b**) and three different types of ensembles. The RMSE was calculated based on the 1956–2013 climatology of the ensemble mean and the observational product CRUTS3.23. Black dots indicate the values for the optimal ensemble, green dots the ensemble based on performance-ranking of individual members and randomly selected ensembles in blue. For the random ensemble, the dot represents the mean of 100 samples and the error bar is the 90% confidence interval. The red circle indicates the optimal subset size with the overall smallest RMSE compared to the observational product. The model simulations which are part of this optimal subset are listed in red font next to the circle. The black triangles represent the optimal ensembles for a cost function that consists of three terms (see Eq. (2)). The corresponding red triangle is the optimal subset of the black triangle cases. The map shows CRUTS3.23 coverage. The solid horizontal line indicates the RMSE value for the MMM of all available simulations. For the dashed line, we first aggregate over the members of one model and then average over all 38 models. The RMSE of the mean of 21 simulations (1 simulation per institute) is represented with the dotted line.

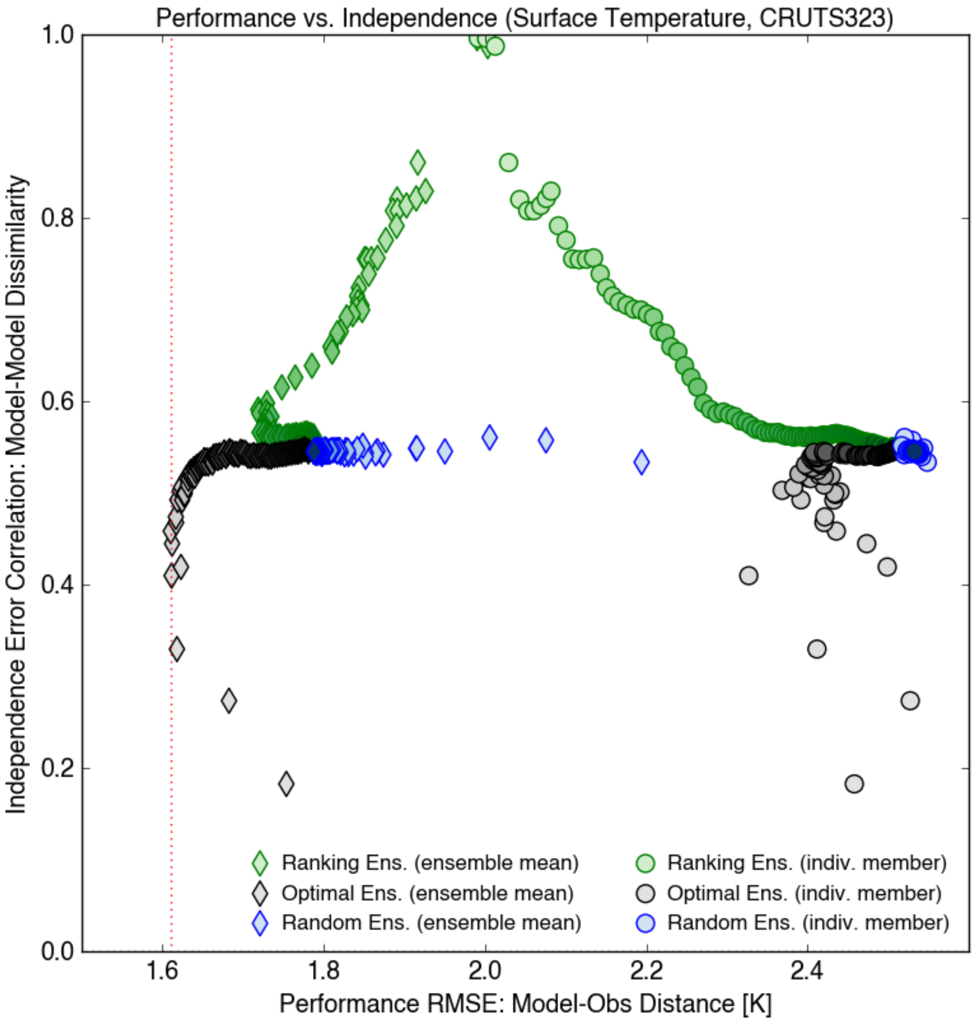
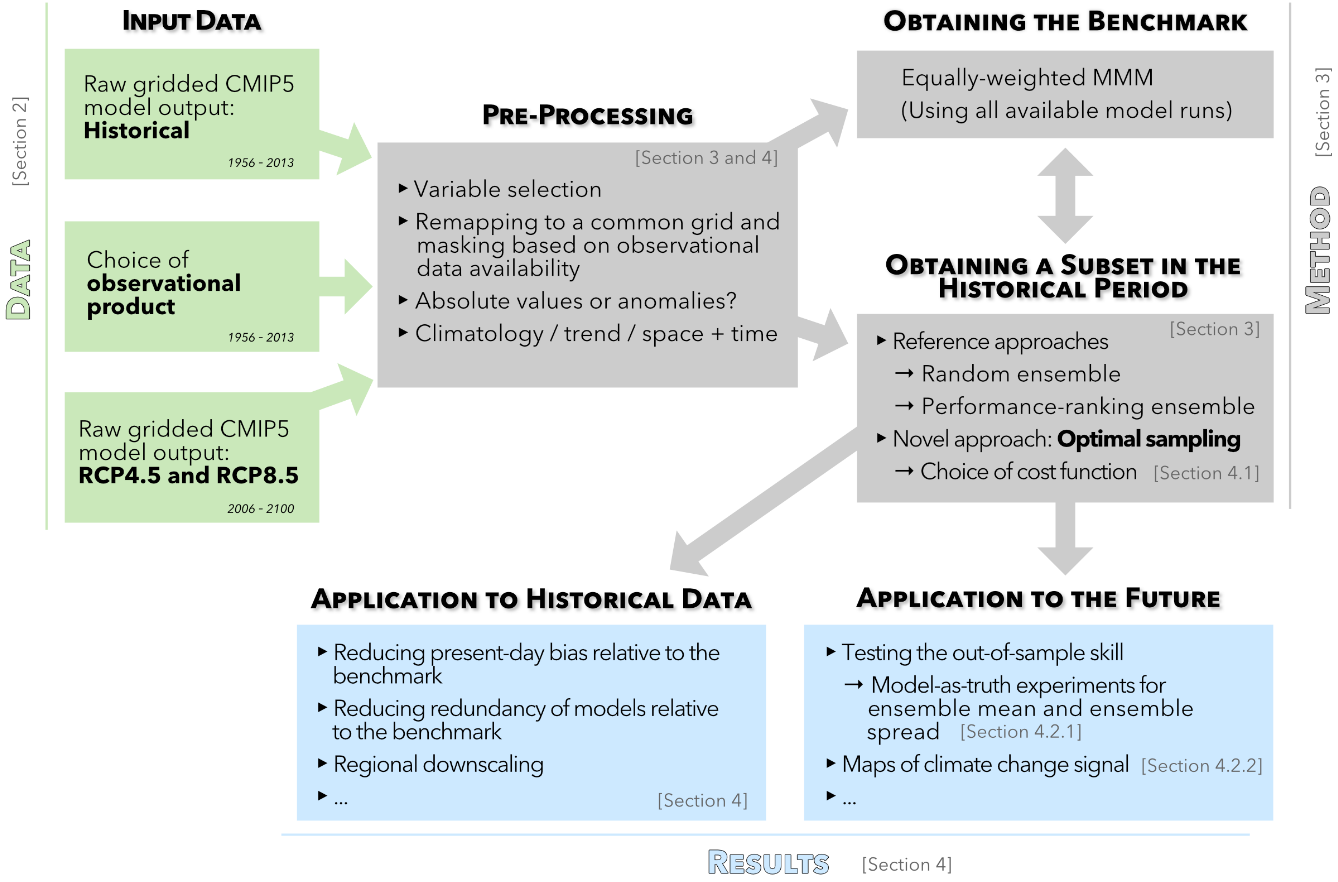


Figure 2. The dependence (in terms of average pairwise error correlation across all possible model pairs in the ensemble) is plotted against the performance (in terms of RMSE) for three different sampling techniques. It is based on surface air temperature and CRUTS3.23 is used as observational product. For the circular markers, the mean of model-observation distances within the ensemble is plotted against the mean of pairwise error correlations for the individual members within an ensemble for a certain ensemble size. The diamonds are used to show the RMSE of the ensemble mean (rather than the mean RMSE of the individual members) compared to the observational product. The values on the vertical axis are the same as for the circular markers. The larger the ensemble size, the darker the fill-color. The red dotted line indicates the lowest RMSE for the optimal ensemble (based on the ensemble mean).



Input Data
[Section 2]
Data
Raw gridded CMIP5 model output: Historical
1956 - 2013
Choice of observational product
1956 - 2013
Raw gridded CMIP5 model output: RCP4.5 and RCP8.5
2006 - 2100
Pre-Processing
[Section 3 and 4]
‣ Variable selection
‣ Remapping to a common grid and masking based on observational data availability
‣ Absolute values or anomalies?
‣ Climatology / trend / space + time
Obtaining the Benchmark
Equally-weighted MMM
(Using all available model runs)
Obtaining a Subset in the Historical Period
[Section 3]
‣ Reference approaches
→ Random ensemble
→ Performance-ranking ensemble
‣ Novel approach: Optimal sampling
→ Choice of cost function [Section 4.1]
Method
[Section 3]
Application to Historical Data
‣ Reducing present-day bias relative to the benchmark
‣ Reducing redundancy of models relative to the benchmark
‣ Regional downscaling
‣ ...
[Section 4]
Application to the Future
‣ Testing the out-of-sample skill
→ Model-as-truth experiments for ensemble mean and ensemble spread [Section 4.2.1]
‣ Maps of climate change signal [Section 4.2.2]
‣ ...
Results [Section 4]

Figure 3. Graphical representation of the method for this study and its flexibility. The different colors are used for three sections in this publication: Data, method and results.

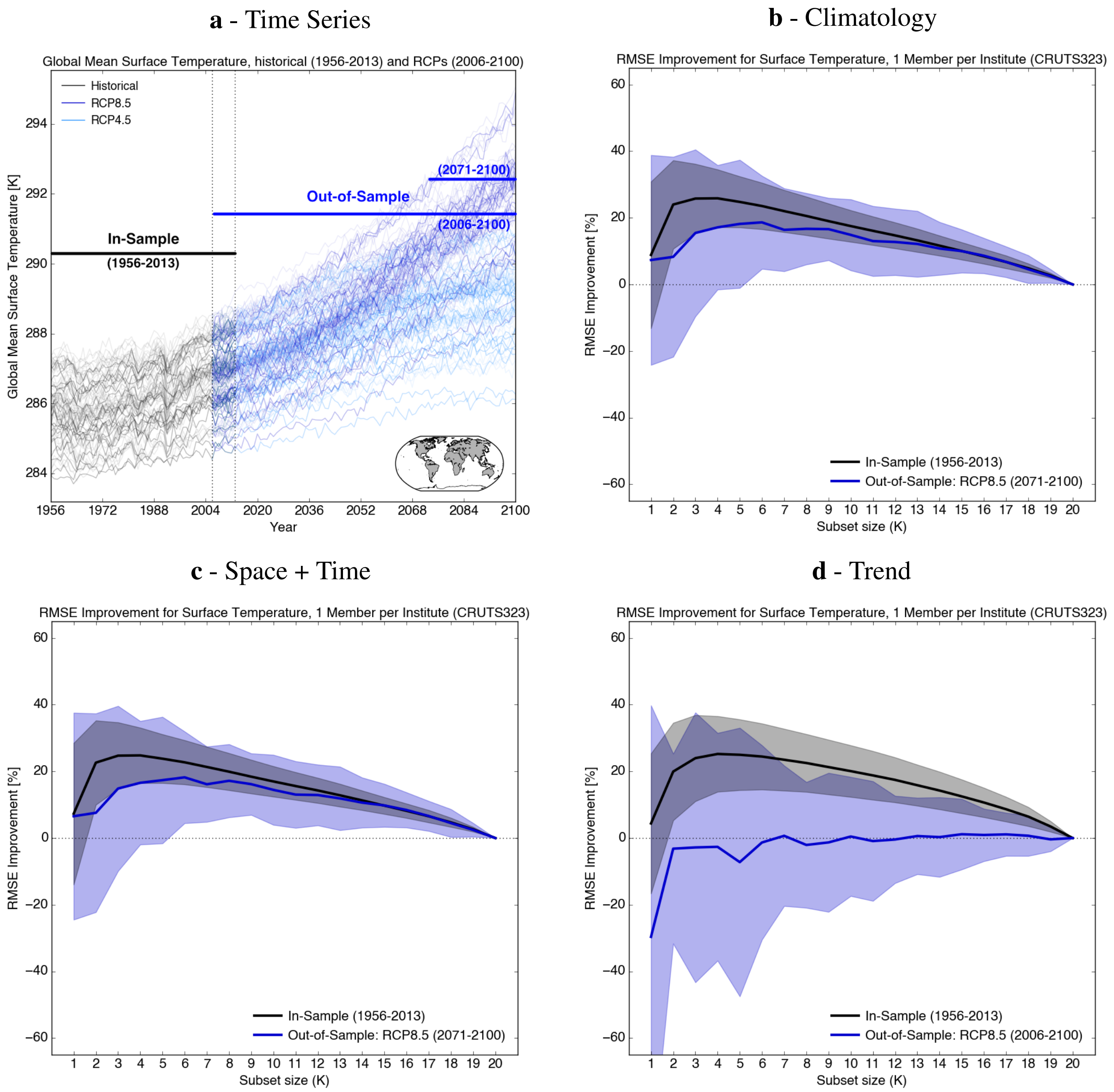


Figure 4. Results of the model-as-truth experiment based on three different metrics (**b-d**) and 21 model simulations (1 simulation per institute). **a**: Time series of surface air temperature averaged over the areas where CRUTS3.23 has data-availability (see map in lower right corner). The time series of the 21 model simulations which are used for the experiment are plotted slightly thicker. 1956–2013 was used as in-sample period, in which the optimal subset is derived and 2006–2100 was used as out-of-sample period for the trend metric and 2071–2100 for the remaining two metrics.

b: The RMSE improvement of the optimal subset relative to the MMM is plotted as a function of the subset size for each model simulation as truth. The subset for each given ensemble size was derived in the in-sample period based on the climatological metric. The curve is the mean improvement across all the 21 model simulations as "truth" and the shading around it represents the spread. Black was used for the historical period and dark blue for RCP8.5. **c** and **d** show the same as **b** but for different metrics.

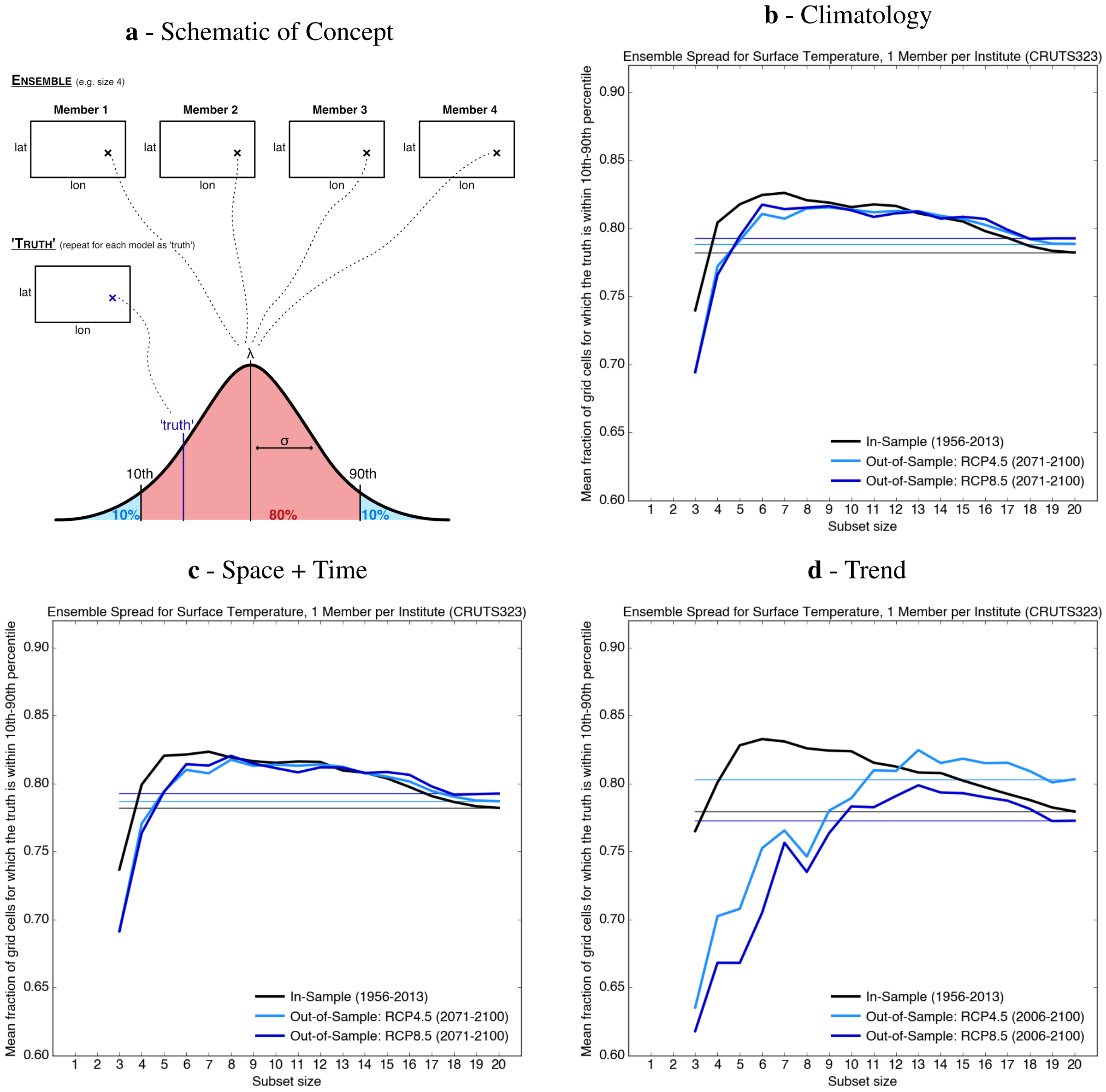


Figure 5. The number of times the 'model-as-truth' is within the 10th-90th percentile of ensemble spread (defined by the optimal subset for a given size) averaged across all 'truths' is plotted against the subset size. **a**: Schematic explaining how the fraction of 'truth' lying in the predicted range is obtained. **b-d**: In- (black) and out-of-sample (blue) curves for three different metrics. Surface air temperature is used as the variable. The horizontal lines refer to the percentage obtained by using all 21 model simulations.

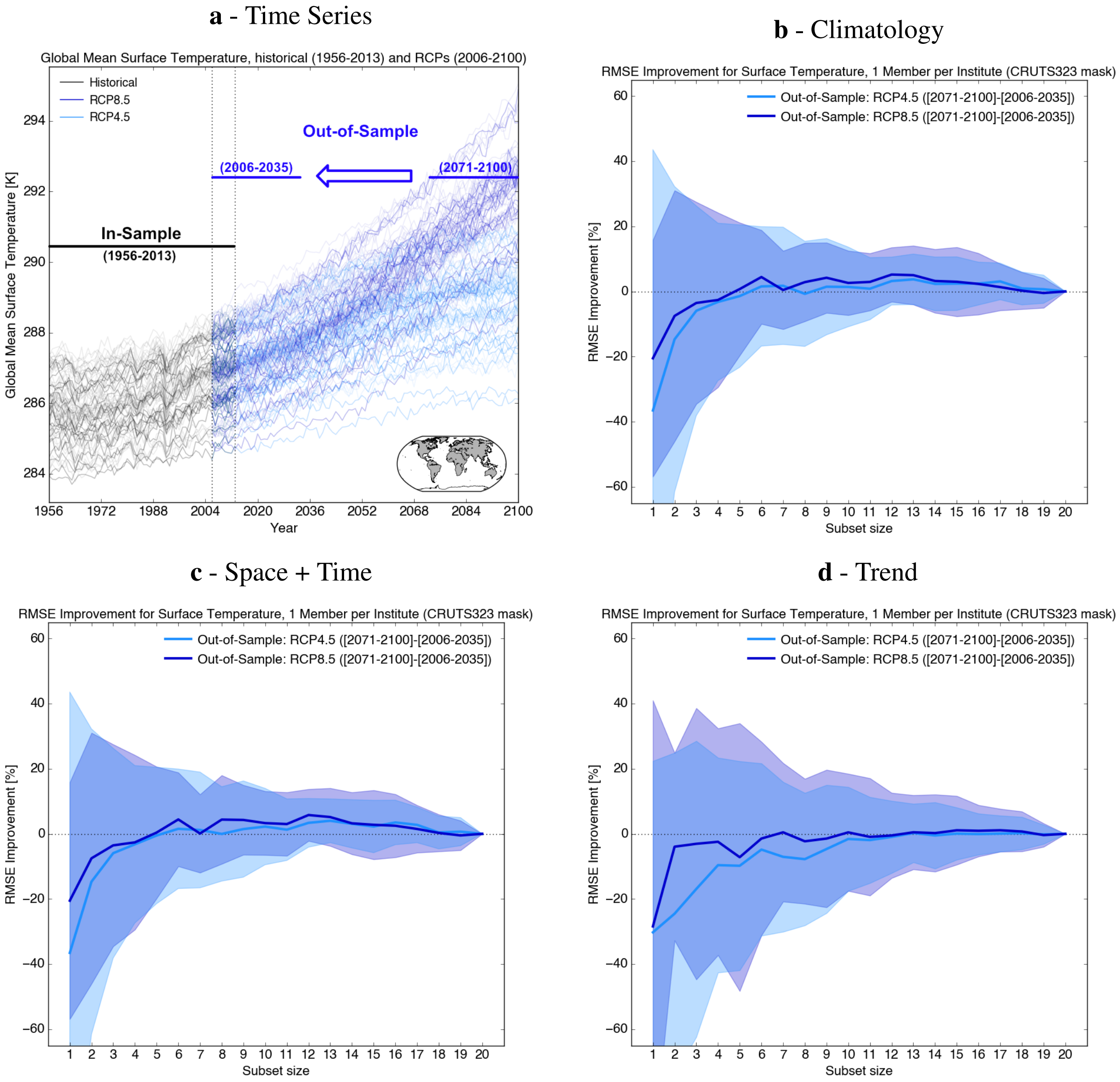


Figure 6. Similar to Figure 4, but here we are trying to predict the [2071–2100]-[2006–2035] temperature change (**a**) based on the optimal subsets obtained with different metrics. For **b**-**d** the optimal ensembles obtained in-sample (1956–2013) are used to predict the surface air temperature change and compared to the "true" temperature change. The same is done with the MMM and then the RMSE improvement of the optimal subset relative to the one of the MMM is calculated for both RCP4.5 and RCP8.5. The curve is the mean across all models as truth and the shading is the spread around it.

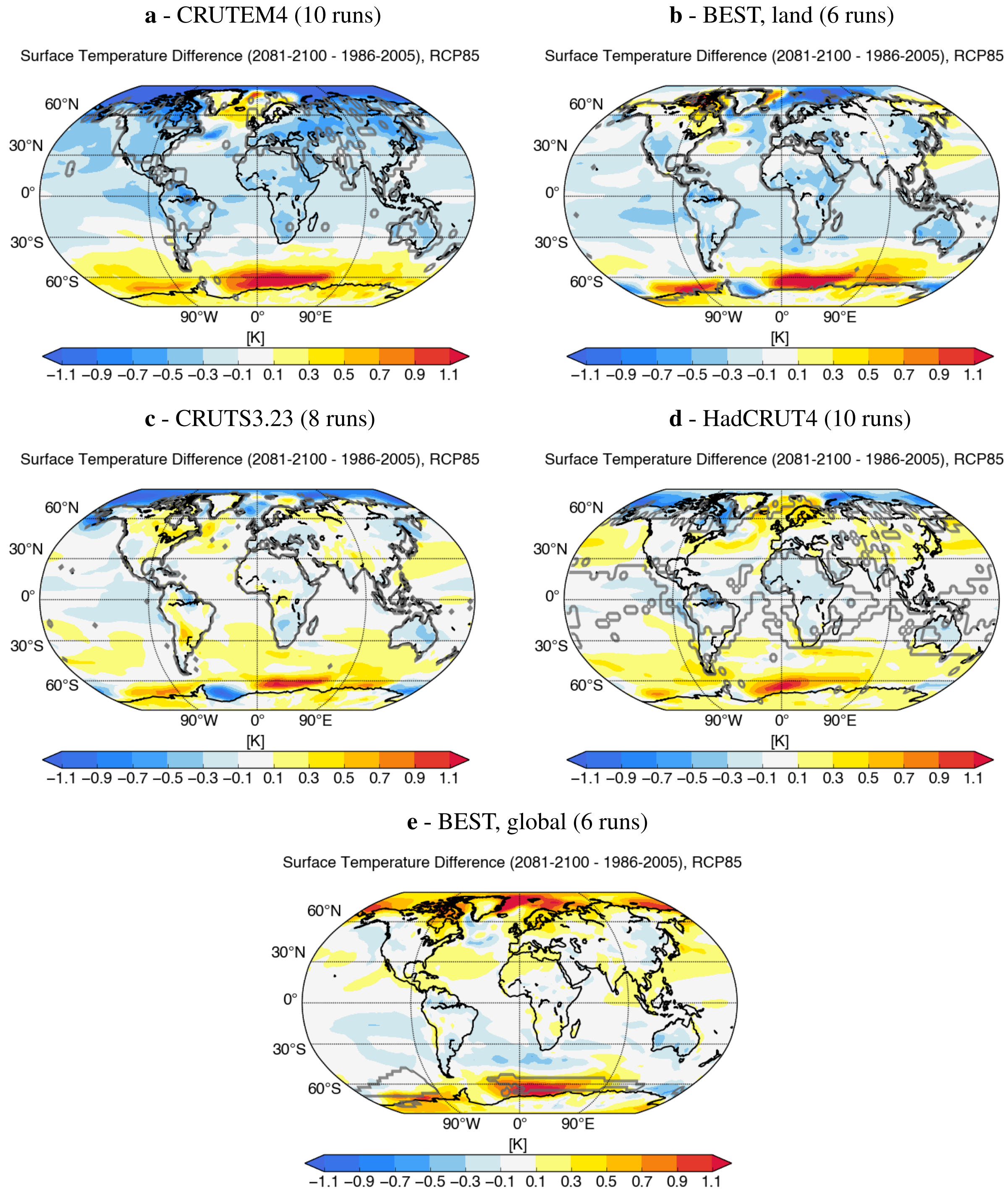
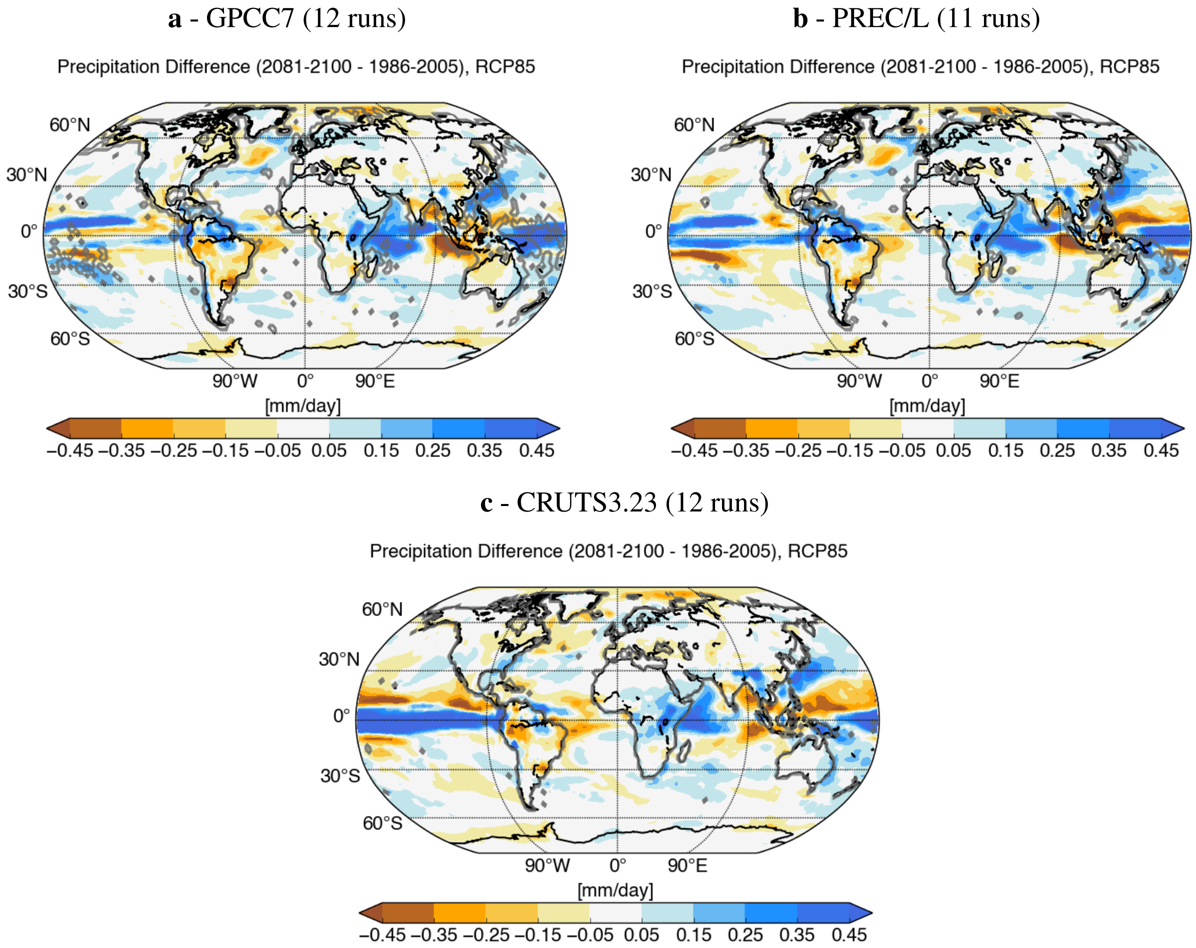


Figure 7. The difference between the multi-model mean (81 runs) and the optimal subset is shown for the RCP8.5 surface air temperature change between [2081–2100] and [1986–2005]. The optimal subset is different depending on which observational product is used. The grey contours outline the region which was used to obtain the optimal subset in the historical period. The optimal ensemble size for each observational product is given in the title of each map.



a - GPCC7 (12 runs)

b - PREC/L (11 runs)

c - CRUTS3.23 (12 runs)

Figure 8. Same as Figure 7, but for precipitation change.

References

Abramowitz, G., and Gupta, H.: Toward a model space and model independence metric, Geophys. Res. Lett., 35(5), doi:10.1029/2007GL032834, 2008.

Abramowitz, G.: Model independence in multi-model ensemble prediction, Australian Meteorological and Oceanographic Journal, 59, 3–6, 2010.

Abramowitz, G. and Bishop, C. H.: Climate model dependence and the ensemble dependence transformation of CMIP projections, J. Climate, 28(6), 2332–2348, doi:10.1175/JCLI-D-14-00364.1, 2015.

Angélil, O., Perkins-Kirkpatrick, S., Alexander, L. V., Stone, D., Donat, M. G., Wehner, M., Shiogama, H., Ciavarellad, A., and Christidis, N.: Comparing regional precipitation and temperature extremes in climate model and reanalysis products, Weather and Climate Extremes, 13, 35–43, doi:10.1016/j.wace.2016.07.001, 2016.

Annan, J. D., and Hargreaves, J. C.: Understanding the CMIP3 multimodel ensemble, J. Climate, 24(16), 4529–4538, doi:0.1175/2011JCLI3873.1, 2011.

Annan, J. D., and Hargreaves, J. C.: On the meaning of independence in climate science, Earth Syst. Dynam., 8, 211–224, doi:10.5194/esd-8-211-2017, 2017.

Baker, N. C., and Taylor, P. C.: A framework for evaluating climate model performance metrics. Journal of Climate, 29(5), 1773–1782, doi:10.1175/JCLI-D-15-0114.1, 2016.

Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Clim. Dyn., 41(3-4), 885–900, doi:10.1007/s00382-012-1610-y, 2013.

Evans, J. P., Ji, F., Abramowitz, G., and Ekström, M.: Optimally choosing small ensemble members to produce robust climate simulations. Environ. Res. Lett., 8(4), 044050, doi:10.1088/1748-9326/8/4/044050, 2013.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res. Atmos., 113(D6), D06104, doi:10.1029/2007JD008972, 2008.

Grose, M. R., Brown, J. N., Narsey, S., Brown, J. R., Murphy, B. F., Langlais, C., Gupta, A. S., Moise, A. F., and Irving, D. B.: Assessment of the CMIP5 global climate model simulations of the western tropical Pacific climate system and comparison to CMIP3, Int. J. Climatol., 34(12), 3382–3399, doi:10.1002/joc.3916, 2014.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 34(4), 751–763, doi:10.1029/97WR03495, 1998.

Gupta, H. V., Bastidas, L. A., Sorooshian, S., Shuttleworth, W. J., and Yang, Z. L.: Parameter estimation of a land surface scheme using multicriteria methods, J. Geophys. Res. Atmos., 104(D16), 19491–19503, doi:10.1029/1999JD900154, 1999.

Gurobi Optimization, Inc., Gurobi Optimizer Reference Manual, http://www.gurobi.com, 2015.

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H.: Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset, Int. J. Climatol., 34, 623–642, doi:10.1002/joc.3711, 2014.

IPCC, 2014: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R. K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp.

Jones, P. W.: First-and second-order conservative remapping schemes for grids in spherical coordinates, Mon. Weather Rev., 127(9), 2204–2210, doi:0.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2, 1999.

Keller, K., and Nicholas, R.: Improving climate projections to better inform climate risk management. In: The Oxford Handbook of the Macroeconomics of Global Warming, Oxford University Press, doi: 10.1093/oxfordhb/9780199856978.013.0002, 2015.

Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P., Hewitson, B., and Mearns, L.: Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections. In: Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections [Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, and P. M. Midgley (eds.)]. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, 2010.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in combining projections from multiple climate models, J. Climate, 23(10), 2739–2758, doi:10.1175/2009JCLI3361.1, 2010.

Knutti, R.: The end of model democracy?, Clim. change, 102(3–4), 395–404, doi:10.1007/s10584-010-9800-2, 2010.

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophys. Res. Lett., 40(6), 1194–1199, doi:10.1002/grl.50256, 2013.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophys. Res. Lett., 44, doi:10.1002/2016GL072012, 2017.

Leduc, M., Laprise, R., de Elía, R., and Šeparović, L.: Is Institutional Democracy a Good Proxy for Model Independence?, J. Climate, 29(23), 8301–8316, doi:10.1175/JCLI-D-15-0761.1, 2016.

Masson, D., and Knutti, R.: Climate model genealogy, Geophys. Res. Lett., 38, L08703, doi:10.1029/2011GL046864, 2011.

Mitchell, J. E.: Branch-and-cut algorithms for combinatorial optimization problems, Handbook of applied optimization, 65–77, 2002.

Pierce, D. W., Barnett, T. P., Santer, B. D., and Gleckler, P. J.: Selecting global climate models for regional climate change studies, Proceedings of the National Academy of Sciences, 106(21), 8441–8446, doi:10.1073/pnas.0900094106, 2009.

Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., and Glecker, P. J.: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, J. Geophys. Res. Atmos., 113(D14), D14209, doi:10.1029/2007JD009334, 2008.

Reichler, T., and Kim, J.: How well do coupled models simulate today's climate?, Bull. Am. Meteorol. Soc., 89(3), 303–311, doi:10.1175/BAMS-89-3-303, 2008.

Sanderson, B. M., Knutti, R., and Caldwell, P.: Addressing interdependency in a multimodel ensemble by interpolation of model properties, J. Climate, 28, 5150–5170, doi:10.1175/JCLI-D-14-00361.1, 2015a.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A representative democracy to reduce interdependency in a multimodel ensemble, J. Climate, 28(13), 5171–5194, doi:10.1175/JCLI-D-14-00362.1, 2015b.

Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, Geosci. Model Dev. Discuss., doi:10.5194/~~gmd-2016-285, in review, 2016.~~ 10.5194/gmd-10-2379-2017, 2017.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.-Atmos., 106, 7183–7192, 2001.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bull. Am. Meteorol. Soc., 93(4), 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.

Tebaldi, C., and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, Phil. Trans. R. Soc. A., 365(1857), 2053–2075, doi:10.1098/rsta.2007.2076, 2007.

Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3), 273–282. doi:10.1111/j.1467-9868.2011.00771.x, 2011.

Xu, Z., Hou, Z. , Han, Y., and Guo, W.: A diagram for evaluating multiple aspects of model performance in simulating vector fields, Geosci. Model Dev., 9(12), 4365–4380, doi:10.5194/gmd-9-4365-2016, 2016.