Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing

Authors' response to Anonymous Referee #2

> *The authors provide a very timely and valuable review of recent studies that have tried to cope with the issues related to model dependence and performance in multi-model climate ensembles. They make clear that the climate community needs to go beyond model democracy "one model, one vote" in the future use of CMIP projects for climate and impact studies (and I agree very much with that statement). They also provide some guidance on the testing and evaluation of methods that are commonly used to move beyond equally weighted ensembles.*
> *I fully support the publication of the paper pending some minor revisions described below.*

We are pleased that the reviewer finds the work valuable and timely. We address each of their comments and suggestions below.

> *Section 1: this section is an important one as it seeks to define the different types of uncertainty involved in climate projections. I find the current text a bit confusing. For instance, the paper begins with a list of uncertainty sources (six of them are mentioned) and then proposes to classify them in the second paragraph in two main classes: epistemic and aleatory. They define epistemic uncertainty as our incomplete "knowledge and understanding of the climate system" implicitly meaning that it is reducible with more information and knowledge. Yet the first source of uncertainty listed in the first paragraph is the uncertainty due to the lack of predictability of human behavior and 21st century history (an uncertainty that includes future GHG emissions but that is indeed much larger). This is usually considered as a third class of uncertainty and is dealt with the use of multiple and plausible scenarios with no explicit credibility ranking among them. Including it in epistemic uncertainty would imply that epistemic uncertainty cannot be reduced to the climate system (as written) but that it has to include the interaction between the climate system and human (social, political and economic) behavior.*

Yes, this is a good point.
Proposed revision: Add the third class of uncertainty, as suggested, and describe the existing descriptions of epistemic and aleatory uncertainty as conditional upon a particular scenario.

> *Another remark is that the limitations in observations are only mentioned as "required for accurate model initialization". Clearly, observations also play a central role in model development and evaluation. Furthermore, the text does not explicitly mention the uncertainty related to the history of past external forcings. Yet this is a major source of uncertainty, in particular at regional scale. As the authors do not begin by their classification, they have to use other terms that are not precisely defined (like structural, page 2, line 3). My suggestion would be to start with the classification right from the beginning (by introducing the three classical types of uncertainty) and then use the same words throughout the text to define the different uncertainty sources.*

As the reviewer notes above, we have not comprehensively accounted for all types of uncertainty in our introduction. In addition to the changes described above, we will be more comprehensive in our description. However, the aim of this work is not to provide a comprehensive account of uncertainty, but rather highlight the kinds of uncertainty that contribute to epistemic and aleatory uncertainty, as these two require a very different treatment in defining and accounting for model dependence.

Next, we disagree that the concept of model structural uncertainty is ill-defined, noting that many pieces of work in this field have been dedicated to discussing it (e.g. Tebaldi and Knutti, 2007; Gupta et al, 2012). We feel most readers familiar with model evaluation would be comfortable with this term, and further, be much more familiar with it and initial condition uncertainty than they would be with epistemic and aleatory uncertainty. We appreciate that the latter are much more general concepts, but we feel starting with familiar concepts is a better pathway for the audience.

Proposed revision: add a more detailed discussion of uncertainty, but most importantly, be clear that the reason we discuss those two broad categories is because they require a different treatment of model dependence.

> *The last paragraph of section 1 states the objectives of the paper but lacks to mention the fact that the sampling GCM strategy is only one piece, albeit an important one, in the complex workflow that goes from emission scenario to climate projection to impact studies. For instance, the independence issue also applies to the design of the GCM- RCMs matrix.*

Yes, this is a good point. We hope that the concepts explored here could be applied equally to the different pieces of this workflow, but nevertheless agree making this point explicit is a good idea.
Proposed revision: add a sentence making it clear that while we have discussed dependence in the context of GCM sampling, there are many more links in the chain to impacts prediction, and that these issues apply equally to other steps in the process, referring to Clark et al (2016) as an example of uncertainty exploration in this chain.

> *Finally, I find a bit strange the last sentence of the section (line 33) as we live in a world where resources for model development are becoming scarce to say the least while there still are many unresolved modelling issues.*

We're happy to rephrase this.
Proposed revision: rephrase as "the use of limited resources on model development that could essentially duplicate the information that other models provide"

> *Section 2: Page 3, line 9: there is a need to clearly differentiate throughout the text multi-model ensembles from initial condition ensembles (single-model ensemble with many members differing by their perturbed initial conditions). I suggest avoiding the use of the word member when referring to the former.*

We do not see a logical reason for the request only to refer to the constituent simulations of an ensemble as 'members' when the ensemble is generated using perturbed initial conditions. In both cases, multiple simulations form an ensemble, so each is a member of the ensemble. We agree more broadly with the reviewer's suggestion that internal variability ensembles (typically explored through initial conditions perturbation) should be treated as categorically different to those representing model structural uncertainty (typically through multi-model ensembles), which is why we have gone to such great lengths to delineate between aleatory and epistemic uncertainty in the way we have.
Proposed revision: none.

> *Section 3: First paragraph: the authors make an interesting distinction between climate model components/processes where we do not expect epistemic departures from the true physical system (where we expect to have strong dependency among models) and those where we expect to see such departures (and where it would be needed to have independent representations of processes). Yet, I wonder if this distinction is really useful in practice. These components are often tightly coupled (think of the atmospheric dynamics/physics pair of components) in a GCM meaning that errors in the latter would lead to biases in the former (for instance biases in cloud microphysics could lead to biases in temperature gradients that would in turn affect atmospheric circulation). Disentangling the exact origin of the biases in a fully coupled system is a rather difficult task. It would be interesting to propose and discuss the hierarchy of models and experiments that would allow a clean separation between these*

> *different types of components/processes. A simple example is the use of SST-forced experiment in addition to a fully coupled one to assess the origin of atmospheric biases.*

Yes, we completely agree with the reviewer that in practice this is likely to be very difficult, for precisely the reasons they describe. In the manuscript we refer to this problem as epistemological holism (line 41, page 3), and point out the detailed exploration of these ideas in Lenhard and Winsberg (2010). Nevertheless, being explicit that independence in models is not as straightforward as wanting all models to be different is important. We want them all to approximate the climate system, insofar as it can be defined through measurements - in that sense we want dependence. We appreciate that the reviewer finds this point interesting, and interpret this point as a comment.
Proposed revision: none.

> *Section 4: Page 4, lines 16-26: as it currently stands, the text seems to imply that component democracy (instead of model democracy) is too difficult to implement "beyond categorical inclusion or exclusion". I would argue that this exact sentence also applies to the institutional democracy that the authors are advocating for. There is certainly some subjectivity in using version numbers to make claims about independence, but I think there is as much subjectivity in using the modelling center names. If a modelling center has 4 different model versions that differ by physics and/or resolution, the final choice of just one version will also be subjective. Finally, there will also be cases where two modelling centers share most of their components leading to potential strong dependency between their simulations. In fact, the authors in their conclusions (page 13, lines 26-31) recognize that some additional work is deeply needed to efficiently use institutional democracy, this extra-work being more or less related to the Boé (2018) type of analysis of the available model documentation and meta-data.*

Yes, we agree with this point. We are only advocating approximate institutional democracy as a naive strategy in the absence of any other strategy that can actually use observational data to define dependence statistically. We agree that component democracy might indeed be similar to (or perhaps even better than) institutional democracy.
Proposed revision: in discussion/conclusions note that where model component information is available, this provides an equally good 'naive' strategy as institutional democracy.

> *Section 8: Page 9, line 17: one could also cite Boé and Terray (2015) "Can metric-based approaches really improve multi-model climate projections? the case of summer temperature change in France. Climate Dynamics, vol. 45, iss. 7, pp. 1913-1928" which discuss the sensitivity of weighting strategy results to a large range of methodological choices.*

Thanks for pointing us to this publication.
Proposed revision: add this citation to the existing list.

> *Page 10, lines 4-31: what is discussed here has close and strong links with the emergent constraint literature and its recent developments (see for instance Nijsse, F. J. M. M. and Dijkstra, H. A.: A mathematical approach to understanding emergent constraints, Earth Syst. Dynam., 9, 999-1012, https://doi.org/10.5194/esd-9-999-2018, 2018). Yet, there is no mention of it and almost none of the relevant papers is cited.*

Yes, agree.
Proposed revision: include a paragraph detailing the relevance and relationship of what's here to emergent constraints.

> *The authors could also discuss (or at least mention) the possible caveats in using regression analysis for the weighting problem: adequacy of the assumed linear model between the predictor and*

> *predictand, standard use of OLS instead of TLS (with error- in-variable), sensitivity of the results and selection bias in data pre-processing (like spatial averaging) . . .*

We agree, as noted above, we do not intend to provide a comprehensive list of sources of uncertainty in this process, but discussing more where appropriate is a good idea.
Proposed revision: rework Section 8 to be more explicit about some of these additional sources of uncertainty.

> *Page 10, lines 33-35: the authors could also cite some recent references, for instance:*
> *D. Maraun: Bias Correcting Climate Change Simulations - a Critical Review, Curr. Clim. Change Rep. 2:2011-220, 2016.*
> *J.M. Gutiérrez, D. Maraun, M. Widmann, R. Huth, E. Hertig, R. Benestad, O . Roessler, J. Wibig, R. Wilcke, S. Kotlarski, D. San Martín, S. Herrera, J . Bedia, A. Casanueva, R. Manzanas, M. Iturbide, M. Vrac, M. Dubrovsky, J . Ribalaygua, J. Pórtoles, O. Räty, J. Räisänen, B. Hingray, D. Raynaud, M.J. Casado, P. Ramos, T. Zerenner, M. Turco, T. Bosshard, P. Šteˇpánek, J. Bartholy, R. Pongracz, D.E. Keller, A.M. Fischer, R.M. Cardoso, P.M.M. Soares, B. Czernecki, C. Pagé. An intercomparison of a large ensemble of statistical downscaling methods over Europe: results from the VALUE perfect predictorcrossâA ˘Rˇvalidationexperiment.Int.J.Climatol.,onlinefirst,2018.*
> *E. Hertig, D. Maraun, J. Bartholy, R. Pongracz, M. Vrac, I. Mares, J.M. Gutierrez, J. Wibig, A. Casanueva and P.M.M. Soares: Comparison of statistical downscaling methods with respect to extreme events over Europe: Validation results from the perfect predictor experiment of the COST Action VALUE. Int. J. Climatol., online first, 2018.*
> *D. Maraun, R. Huth, J.M. Gutierrez, D. San Martin, M. Dubrovsky, A. Fischer, E. Hertig, P.M. Soares, J. Bartholy, R. Pongracz, M. Widmann, M.J. Casado, P. Ramos and J. Bedia: The VALUE perfect predictor experiment: evaluation of temporal variability, Int. J. Climatol., online first, DOI: 10.1002/joc.5222, 2017.*

These are useful references.
Proposed revision: add some of the above references to the section noting that these issues apply more broadly to the climate impacts chain.

> *Section 9: Page 11, lines 9-16: see comment on section 4 that also applies here.*

Proposed revision: Yes, as noted above, we will address this by also suggesting naive selection by model component information, if it’s available.

> *Page 11, line 19: Institutions also often co-develop (instead of “copy”) models and/or components (such as the NEMO ocean engine in Europe).*

Yes. Thanks for pointing this out.
Proposed revision: amend this sentence to include the potential for model co-development.

> *Page 11, lines 22-28: some of statements in this paragraph are just claims with no supporting evidence (“. . . quickly become difficult and time consuming . . .”, “Using this information . . . seems the only option”). I think that the issues with regard to component and institutional democracy are quite similar.*

We agree that many of these issues apply equally to institutional democracy.
Proposed revision: amend this section to highlight that many of these issues apply equally to institutional democracy.

> *Page 12, lines 4-13: the authors might also want to discuss and cite: Borodina, A., E.M. Fischer, and R. Knutti, 2017: Emergent Constraints in Climate Projections: A Case Study of Changes in High-Latitude Temperature Variability. J. Climate,30, 3655– 3670, https://doi.org/10.1175/JCLI-D-16-0662.1*

It’s not entirely clear why this paper should be in the section suggested here, other than it using more than a single constraint. Borodina et al. (2017) considered multiple predictors but combined them in a single cost function rather than conducting multi-objective optimisation. We interpret this as a comment.
Proposed revision: none.

References:

Gupta HV; Clark MP; Vrugt JA; Abramowitz G; Ye M, 2012, 'Towards a comprehensive assessment of model structural adequacy', Water Resources Research, vol. 48, http://dx.doi.org/10.1029/2011WR011044

Clark, M.P., Wilby, R.L., Gutmann, E.D. et al. Curr Clim Change Rep (2016) 2: 55. https://doi.org/10.1007/s40641-016-0034-x

Tebaldi, C., and R. Knutti (2007), The use of the multimodel ensemble in probabilistic climate projections, Philosophical Transactions of the Royal Society, Series A, 365, 2053–2075, doi:10.1098/rsta.2007.2076.