Earth System Dynamics Discussions

Earth Syst. Dynam. Discuss., author comment AC1
https://doi.org/10.5194/esd-2022-31-AC1, 2022

Comment on esd-2022-31

Tamzin Emily Palmer et al.

Author comment on "Performance based sub-selection of CMIP6 models for impact assessments in Europe" by Tamzin Emily Palmer et al., Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2022-31-AC1, 2022

Initial response reviewer #1 from authors

Peer review of "Performance based sub-selection of CMIP6 models for impact assessments in Europe" by Palmer et al. (ESD).

This paper presents a performance assessment of CMIP6 simulations for Europe and selects a subset of models for regional climate impact studies. The performance criteria include large-scale processes such as storm tracks, circulation patterns, and temperature biases. The selection of models is based primarily on subjective assignment of each model into three categories for each performance criterion. The authors highlight that there is a strong tendency for the models with high regional performance to have higher global climate sensitivity. While the causes of this relationship is left for future investigation, the authors note that this relationship creates a tension between selecting for high regional performance and selecting an ensemble consistent with observational constraints on global ECS.

This paper is thoughtful and well executed. It will be useful for European climate impact assessments, and also as a template/benchmark for performance assessments in other regions. While the paper is acceptable with minor technical corrections, I have added some optional suggestions for improvement. The most important of these suggestions is for an assessment of the role of internal variability in the performance evaluation.

We thank the reviewer for their overall positive and very constructive response. Along with their helpful suggestions for improving the manuscript. Our initial response is given below.

Corrections required:

There are many spelling and grammar mistakes. I noted typos in lines 10, 52, 69, 82, 114, 137, 188, 202, 211, 238, 245, 279, 319, 330, 426, 432, 436, 444, 464, 467, and 471 and in the spelling of "conceptulization."

Table 1. ACCESS-CM2 Is missing from the left column. Also, since the right column is a subset of the left, couldn't this table be replaced with a (less space-consuming) list, with selected models highlighted in bold?

Table 2. The selected model in each cluster needs to be identified. This info isn't available from figure 7 or anywhere else in the main text.

We thank the reviewer for noting these errors. The final manuscript will be proof-read, and the typos noted above corrected. Table 1 will also be replaced with a list as suggested here.

The selected model from each cluster will be identified in table 2, either in bold or by an alternative method. All the models in figure 7 will be identified by numbering the points in the supplementary material (Fig. S4).

Suggestions for improvement (optional):

Models are evaluated on the basis of a single realization each. To what extent does internal variability affect the assessments? The paper would be more solid if it included an analysis of the robustness of the performance criteria to multiple realizations of at least one model.

We agree with the reviewer that the assessment would be more robust with an understanding of the importance of internal variability. We have had a provisional look at other realisations for some of the models (e.g., MIROC6, CAMS-CSM1-0, CanESM5), that are excluded (have a red flag), due to temperature bias and/or circulation errors. This provisional investigation indicates that the assessment would not be altered for these models by using the 2nd or 3rd instead of the 1st realisation (for example). We are still considering how to respond, and many of the team are currently away from work at the moment. However, we anticipate using a larger assessment of internal variability perhaps with CanESM5 model, which has a large number of available realisations, for a number of criteria which are currently used to exclude models. These results would be added to the supplementary material for the manuscript.

This paper's strength is in the process evaluations, which will be a useful reference for analysts creating bespoke ensembles. The 3x3 matrix of examples of models in the three subjective categories is a nice way of presenting the results in the main paper and the appendix. however, many analysts would benefit from a supplementary section showing the maps for the full set of assessed models, so they can make their own subjective assessments and better understand figures 4 and 5.

We agree with the reviewer that it would be beneficial to users to make the maps for the full set of assessed models available. An accessible github repository is currently under construction, that will be linked to from this paper. This will include the maps used in the assessment (at a minimum including temperature, SST and large-scale circulation) and plots for the precipitation annual cycle. Including this in a github repository would enable this to potentially be maintained as a living document, that can be added to as more models or diagnostics become available.

In addition, this repository also includes a spreadsheet of all assessments carried out for the CMIP6 models to date. The sample of 31 models included in this study from this spreadsheet were selected because they had both a minimum number of assessed criteria, and ssp585 future projection data for Tas and Precipitation available.

The finding that many of the high-skill models are outside the IPCC assessed ECS range is interesting and important. However, this tension between regional skill and global climate

sensitivity seems somewhat overstated. There are a couple of solutions that partially resolve this tension. First, there is the option of presenting analyses relative to global warming levels instead of time, as widely practiced in the literature and advocated by Hausfather et al. (2022). While the GWL approach doesn't fully resolve the tension (time does matter to many studies), it warrants some discussion here. Indeed, the results of this paper add further weight to the importance of the GWL approach. Second, the IPCC's very likely ECS range is a more inclusive and defensible (66% is a high bar, given the observational uncertainties on the upper tail of ECS) criterion that would only exclude three independent models (CanESM, UKESM/HadGEM, and CESM2). Discussion of these nuances would give more direction to the reader in the face of the tension that this paper highlights.

We agree with the reviewer that the tension between the IPCC assessed climate sensitivity range, and the regional skill of the models is not an issue if the GWL method is a suitable approach. Some discussion of this is warranted in the manuscript. We intend to provide a more nuance, revised discussion in the text.

There are however cases where the GWL method is not suitable, such as, where the distribution of the ensemble is used as a measure of likelihood. As shown in our results, the distribution of the filtered models with greater regional skill is skewed towards higher climate sensitivity. In this case the tension between the regional skill and climate sensitivity then becomes relevant. It is particularly important where assessments are made by a risk adverse user, that is interested in a high impact, low likely hood (but plausible) temperature change within a given time frame (e.g., 2030 or 2040).

The very likely IPCC range for ECS will be added to figure 6 for reference.

The completeness of scenario experiments by each model is an important consideration in ensemble selection that doesn't receive any attention here. For example, HadGEM3-GC3.1 provides only one simulation of SSP126 and no simulations of SSP370 (https://pcmdi.llnl.gov/CMIP6/ArchiveStatistics/esgf_data_holdings/ScenarioMIP/index.html), and as a result may not be viable for some study designs. The paper could benefit from some documentation and/or discussion of this and other practical considerations that will affect the utility of the recommended ensemble

We agree that the completeness of the scenario experiments is likely to be a consideration for users. The focus of the paper is on the process-based assessment, rather than attempting to address some of the wider potential considerations for selecting representative models, for downscaling and impact assessments. However, we agree that some relevant links to the documentation, or a table of the filtered models with the simulations available would be a useful addition to the supplementary information.

The exclusion of UKESM1 based on orange flags comes across as a bit haphazard and arbitrary, especially given that analysis of storm track performance is not available for this model. While I noted the discussion on the confluence of reasons for excluding UKESM1, the paper would benefit from a more systematic documentation of the interaction of criteria leading to model exclusion. Perhaps also there is a role for a "marginal" category of models for which exclusion wasn't clear-cut.

This is useful feedback. We agree that the decision to remove models due to a certain percentage of orange flags is somewhat arbitrary. The decision as to how many orange flags warrants the removal of a model has a degree of subjectivity. We suggest adding an alternative filtered sub-set that only removes models with a red (inadequate) flag and therefore includes both UKESM1 and TaiESM1 as 'marginal' or less preferred models in addition to our current filtered example.

Another approach that a user may wish to consider, is to follow the method of McSweeney et al., (2018), where 'marginal' or less preferred models (with larger numbers of orange flags) are removed if the projection range is not reduced by the removal of a model. If the aim is maintaining the full range (for marginal models) the UKESM1 model is an important consideration as an outlier that otherwise does well at many of the criteria (other than a substantial winter cold temperature bias), as its removal reduces the upper tail of the projected temperature range for Europe.

Minor comments:

Line 225. Some more detail on the reanalysis/observational data would be helpful

Lines 427-8. "The retention of higher sensitivity models is an emergent consequence of assessment of skill at reproducing regional processes." This wording implies some functional relationship between regional skill and model sensitivity that hasn't been established (as duly noted in the conclusion). Simpler wording would reduce the chance of misinterpretation by the reader.

Lines 459-60. Shiogama (2021) excluded models based on a criterion of high recent warming relative to observations, rather than based on ECS or TCR as implied here. Mahony (2022) (DOI:10.1002/joc.7566) would be a more direct example of ensemble selection based on the IPCC assessed ECS range.

Thank you for these points, these will be addressed in the manuscript.

McSweeney, C. et al. (2018) Selection of CMIP5 members to augment a perturbed–parameter ensemble of global realisations of future climate for the UKCP18 scenarios.