

Earth Syst. Dynam. Discuss., referee comment RC1
<https://doi.org/10.5194/esd-2022-31-RC1>, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on esd-2022-31

Anonymous Referee #1

Referee 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-RC1>, 2022

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.

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.

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.

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.

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.

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.

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.

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.