Review of „Climate Model Selection by Independence, Performance and Spread (ClimSIPS) for regional applications“ by Merrifield et al. (2023) by Swen Brands

General comment: The authors provide a comprehensive description of the ClimSIPS tool for weighting Global Climate Models according to the three criteria mentioned in the title. The study provides a detailed introduction to the concepts and methods used in this research field and comes with a detailed bibliography, so that it is almost a review article. The manuscript is well written, timely and relevant and I recommend a minor version in which the following points should be addressed:

Dr. Brands, thank you so much for taking the time to review our manuscript. That you find it almost a review article is a high compliment, especially in comparison to your well-cited studies. We've worked to carefully address your review point-by-point and feel that it has improved the manuscript. We hope you feel the same.

1. Since the model independence results obtained in your study are very similar to those obtained in Brands (2022), the authors might wish to cite this study in the revised manuscript. Particularly the „one family one vote standard“ (e.g. line 454) was also adopted in Brands (2022), where GCM clusters were built by combining the a priori criterion „use of the same AGCM family“ based on Brands et al. (2023) with the a posteriori criterion „error pattern correlation coefficient > +0.65“ (the Boé 2018 nomenclature is followed here). Albeit another predictor was used to measure a posteriori model dependence, the outcome is similar to yours (compare Figure 3 in Brands 2022 to your Figure 3). This shows that the results are robust to changes in the applied methodology.

   It is fantastic to see that our approach is more or less consistent with field standards! (We've been working to develop the ClimWIP independence definition to be consistent with known dependencies for several years.) We've added a few references to Brands 2022 throughout the paper including:

   L382–384: "We also anticipated three "extended" families based on an analysis of model metadata, summarized in Sup. Tabs. S1 and S2, and the work of Brands (2022b), which grouped models in CMIP5 and CMIP6 via shared atmospheric circulation error patterns."

   L866–867: "We were able to support all family designations with model descriptions and metadata and found our designations to be broadly consistent with other model output and metadata-based dependence definitions (Brands, 2022)."

2. Lines 53-56: „modeling centers often contribute several versions of their base model under different names as well (Leduc et al., 2016); these variants differ by, for example, the spatial resolution of some model components or biogeochemical cycling, which may influence their simulated climate in ways that are difficult to anticipate.“

   Should start with an uppercase letter ("Modeling centers...").

   Thank you; fixed.

   Model versions from the same center often differ by the inclusion of entire numerical sub-models describing specific Earth system components in addition to the basic four components atmosphere, land-surface, ocean and sea-ice. In this context it is interesting to note that the names and versions of all sub-models representing up to 12 climate system components is listed in Brands et al. (2023) for 61 nominally distinct GCMs from CMIP5 and 6. In this extensive metadata archive, you can see how the distinct modelling groups have built their models in terms of included sub-models / Earth System components.

What a great resource! It will certainly be widely used. We've added reference to it here:

L49-52: "Modeling centers often contribute several versions of their base model under different names as well (Leduc et al. 2016); these variants differ by, for example, the spatial resolution of some model components or entire sub-models (see Brands et al. 2023), which may influence their simulated climate in ways that are difficult to anticipate."

3. Lines 56-57: „Adding further complexity, even uniquely named models from different modelling centers fall along a spectrum of uniqueness."

   I do not fully understand what you mean with this sentence. Could you provide an example for „uniquely named models from different modeling centers" ?

   That is a confusing formulation. We've revised it as:

   L52-53: "Adding further complexity, models actually fall over a spectrum that ranges from effective replicates to fully independent entities."

4. Lines 58-61: I think the Brands et al. (2023) GCM metadata archive is relevant in this sentence as well and the authors might wish to refer to it. The archive could be alternatively cited in lines 78-80 and is useful for determining „a priori" dependencies within in the CMIP ensemble, as defined by Boé (2018).

   We've added the reference as:

   L53-55: "Different models share historical predecessors (Masson and Knutti, 2011, Knutti et al. 2013), conceptual frameworks, and, in some cases, source code (Boé, 2018, Brands 2022b, Brands et al. 2023)"

5. Lines 119-121. Meanwhile, the EURO-CORDEX model selection team has come to a final recommendation for the driving GCMs from CMIP6. Please see Sobolowski et al. (2023) for more details.

   Thank you for pointing me (Anna Merrifield) to this white paper! I am very happy to see that model independence is a key part of the EURO-CORDEX model subselection. A bit of a back story, this paper was largely prepared for submission in September 2022, but had to be "shelfed" for a few months for my maternity leave when my daughter arrived a few weeks early. While we hoped to have the paper out for consideration in the EURO-CORDEX selection process, time wasn't on our side this time. I hope the community will still find the method valuable!

   As we've worked to shorten the paper, we've removed reference to EURO-CORDEX.

6. Lines 205-217: Please indicate the time aggregation of the GCM and reanalysis data you are using. Is the study based on monthly-mean data?

   Absolutely, we've added reference to the monthly mean output here.

   L183-188: For inclusion in Part I, the models also must provide (1) an estimate of ECS, calculated from a 4XCO_2 run using the Gregory method (Gregory et al., 2004) and (2) the following monthly-mean output fields (with their abbreviation and model output variable name given in brackets): near-surface 2-meter air temperature [SAT; tas], precipitation [PR; pr],

and sea level pressure [SLP; psl]. Further inclusion into Part II's European case studies require the additional monthly-mean output fields of sea surface temperature [SST; tos], and all sky and clear sky downwelling shortwave radiation at the surface [rsds and rsdscs, respectively].

7. Lines 218-219 and elsewhere: Would make sense to use the terms „a priori“ and „a posteriori“ model dependence (Boé 2017) in this study?

   We considered this and have used the „a priori“ and „a posteriori“ terminology in other papers (see Merrifield et al. 2020). Here due to there being several comparatives surrounding dependence for the reader to remember already (within-model vs. between-model, individual vs. family, etc.), we decided to formulate more as a model output-based independence definition with a metadata-based justification.

8. Lines 230-238: The definition of the INV and SME groups is clear but more information is needed on how you define the FAM group. For example, ACCESS-ESM1-5 is here considered an „SME“ model, meaning that it „[...] is represented by multiple members (e.g., initial condition ensembles, perturbed physics ensembles, combinations thereof) but is not determined to be part of a broader multi-model family.“ However, a closer look at the „source“ attributes of the corresponding netCDF files from ESGF and at the reference articles (doi: 10.1071/ES19035, 10.5194/gmd-12-4999-2019, 0.5194/gmd-4-723-2011,0.5194/gmd-4-1051-2011) reveals that the entire ACCESS GCM family is based on versions of the atmospheric sub-models (or AGCMs) developed at the MetOffice-Hadley Centre. Namely, ACCESS-ESM1-5 makes use of the „HadGAM2“ AGCM that is also used by the HadGEM2-ES and HadGEM2-CC coupled model configurations. HadGAM2 was further developed into „MetUMHadGEM3-GA7.1“, constituting the AGCM used in both Hadley Centre‘s and CSIRO‘s coupled model configurations used in CMIP6, e.g. HadGEM3-GC31-MM (doi: 10.1071/ES19040) and ACCESS-CM2 (doi: 10.1071/ES19040). Thus, it is reasonable to put the HadGEM and ACCESS coupled model configurations to the same family, as was done in Brands (2022), because they essentially share their atmospheric component. Following your nomenclature, this would mean assigning a „FAM“ to ACCESS-ESM1-5. Note that all the aforementioned model metadata is available at one glance from Brands et al. (2023).

   This is an important point. Before we discuss though, we do contend that it is not ideal to have the FAM designation introduced before we describe how we make the FAM distinction. Because it serves as a good “quick reference”, we plan to keep it in the table but highlight the preview aspect more in the text as:

   L205-213: “Finally, to familiarize the reader with the concept of model families we will subsequently define, we also list the family group status of each model. The designation, "INDV", indicates a model is considered to be an individual represented by a single member. "SME" signifies that a model is represented by multiple members (e.g., initial condition ensembles, perturbed physics ensembles, combinations thereof) but itself is considered an individual entity. This means it was not found to be part of a broader multi-model family or “FAM” by the criteria we subsequently define. In total, the 218 CMIP6 simulations from 37 uniquely named models considered in Part I fall into 19 Groups (7 multi-model families, 8 single model ensembles, and 4 individuals) and the 75 CMIP5 simulations from 29 uniquely named models fall into 20 Groups (8 multi-model families, 5 single model ensembles, and 7 individuals). In Part II, 197 CMIP6 simulations from 34 uniquely named models and 68 CMIP5 simulations from 26 uniquely named models remain for the subselection exercise (Sup. Tabs. S1-S2).”

   We initially assumed that ACCESS-ESM1-5 would be a part of the Met Office-Hadley Centre model family:

L382-385: "We also anticipated three "extended" families based on an analysis of model metadata, summarized in Sup. Tabs. S1 and S2, and the work of Brands (2022b), which grouped models in CMIP5 and CMIP6 via shared atmospheric circulation error patterns. The first, shown in dark red (models 1-6) in Fig.3, is comprised of models with UK Met Office Hadley Centre atmospheric components."

However, we found it did not meet our definition for family member:

L385-392: "In CMIP6, intermember distances show five of the six models highlighted in red on the y-axis of Fig.3a, satisfy both the self-contained group and median intermember distance threshold criteria to form a family. This grouping makes sense as all five models (HadGEM3-GC31-MM, KACE-1-0-G, ACCESS-CM2, HadGEM3-GC31-LL, and UKESM1-0-LL) use the same MetUM-HadGEM3-GA7.1 atmospheric component (Sup. Table S1). The sixth model, ACCESS-ESM1-5, does not satisfy the self-contained criteria and is closer to other models in CMIP6 than it is to its anticipated family members. This likely occurs because ACCESS-ESM1-5 uses a CMIP5-era HadGAM2 atmospheric component rather than the CMIP6-era MetUM-HadGEM3-GA7.1 atmospheric component and highlights the potential for models in the same development stream to differentiate themselves from their successors."

So the crux is: Can model development make a model functionally independent from a predecessor? This gets at what "independent" in this context means, which for us is a historically distinct enough simulation such that agreement in projection of future climate has meaning beyond "this model is agreeing with itself". An argument can be made that models developed by the bigger modelling centers like the Hadley Centre or NCAR conceivably could be functionally independent generation to generation. For example, from CAM4 to CAM5, NCAR updated many aspects of their aerosol and cloud parameterizations, alleviating several longstanding radiation biases (Kay et al. 2012).

Reference: Kay, J. E., and Coauthors, 2012: Exposing Global Cloud Biases in the Community Atmosphere Model (CAM) Using Satellite Observations and Their Corresponding Instrument Simulators. J. Climate, 25, 5190–5207, https://doi.org/10.1175/JCLI-D-11-00469.1.

(The qualifier "functionally" acknowledges that most models in CMIP are very similar to each other to start with regardless of origin, so an argument can be made that no current model is truly independent…)

To prime readers for the idea that model development could lead to independence, we've added to the opening paragraph of Section 3 Revisiting Model Dependence:

L218–221: "In prior studies, it has been shown that a climate model's origins and evolution can be traced via statistical properties of its outputs (e.g. Masson and Knutti, 2011; Bishop and Abramowitz, 2013; Knutti et al., 2013). Output-based modelidentification can uncover hidden dependencies within the ensemble, e.g. models that are similar because they share components or lineages, but not names. The approach also has the advantage that it does not presume model similarity based on name alone; output from models in active development can evolve substantially from version to version (e.g. Kay et al., 2012; Boucher et al.,2020; Danabasoglu et al.,2020) while output from the same version of a model run at different modeling centers is often quite similar (Maher et al., 2021b)."

9. Lines 240-247: Could you also shortly refer to the disadvantages of the a posteriori / output data–driven approach to measure GCM dependence ? Here, only the advantages are described so far.

This discussion was definitely missing. We've added reference to the primary disadvantage of the a posteriori / output data–driven approach to the opening paragraph of Section 3 Revisiting Model Dependence:

L221–225: "Risks arise, though, if model output used to determine similarity converges within a multi-model ensemble broadly, and thus becomes ineffective at differentiating between dependent and independent models (Brands, 2022b). To reduce the risk of similar output conflating dependent and independent models, we update the model dependence strategy from the ClimWIP independence weighting scheme (Brunner et al., 2020b) to revisit the concept of model families within CMIP."

And to the discussion:

L854–856: The potential for between-model convergence is cited as one of the primary drawbacks of using model output to determine dependence (Annan and Hargreaves, 2017; Brands, 2022b)."

10. Lines 259-261: Please add an equation to define inter-member GCM distance.

Absolutely. We've formally defined I_{ij} here:

L237–245: "Intermember distance (I_{ij}) is calculated through pairwise RMSE between ensemble members i and j for each predictor field $\hat{y}$ individually. Individual predictor RMSEs (ϕ_{ij}) are defined as:

$$\phi_{ij} = \sqrt{\frac{\sum_{k=1}^{p} w_k \left|\hat{y}_i - \hat{y}_j\right|^2}{\sum_{k=1}^{p} w_k}}$$

which reflects an RMSE weighted over the p gridpoints in a latitude / longitude domain, with w_k indicating the corresponding cosine latitude weights. Each ϕ_{ij} is normalized by its respective ensemble mean value ($\overline{\phi}$) and all ϕ_{ij} are averaged together to obtain a single I_{ij} for each member pair. As in Merrifield et al., 2020, I_{ij} is comprised of two individual predictor fields, global-scale annual average SAT and SLP climatologies.

$$I_{ij} = \frac{1}{2}\sum_{l=1}^{2}\left(\frac{\phi_{ij}}{\overline{\phi}}\right)_l."$$

11. Lines 279-280: The observational density underlying theses gridded dataset is also reduced during the first half of the 20th century, particularly in the Southern Hemisphere.

This is a good point. We've amended the statement as:

L265–268: "However, we find that increasing the period back into the 19th century does not appreciably change intermember distances (not shown). Additionally, the 1905 start date may allow for backward-compatibility of the metric with future generations of CMIP should organizers decide to begin the historical period in the 20th century rather than the 19th century."

12. Lines 285-287: „relative change with respect to a historical period" is not considered „model performance", as far as I know. Traditionally, the term „model performance" refers to model error with respect to observations.

Thanks for bringing it to our attention, this sentence is confusing. What we were trying to convey is that using the absolute value of fields like SAT has traditionally been less common for model evaluation than using relative anomalies. We've revised the sentence to read:

L274-275: "The absolute magnitude of a climatic field tends to be seen as secondary to its relative change with respect to a historical base period for most applications (Jones and Harpham, 2013)."

13. Line 307: „confined to subtropical regions" > „confined to the tropics and subtropics"

Thank you, fixed as:

L294–295: "This "low" between-model spread is largely confined to oceanic regions in the tropics and subtropics for both the SAT and SLP 1905-2005 climatologies."

14. Lines 323-324: The Brands et al. (2023) metadata archive comprising names and versions of the sub-models used in each GCM configuration helps to identify the „very similar but differently named models" you refer to in this sentence.

We have also removed this sentence in our effort to shorten the paper, but reference to Brands et al. 2023 is made in several other places including:

L425–426: "…or due to similar ocean component models (Brands et al. 2023)."

L51-52: "…for example, the spatial resolution of some model components or entire sub-models (see Brands et al. 2023), which may influence their simulated climate in ways that are difficult to anticipate."

15. Figure 3b) I can here see 3 independent clusters instead of the 2 indicated in the caption.

Interesting, is it CanESM5 that you see as separate from the CMIP6 core visually? CanESM5 is on the independent side of CMIP6 core, about 6 units from its approximate center in the MDS projection while the MIROC models are about twice as far (~13 units). The broken axis may exacerbate this visual issue, but we feel it was necessary to allow readers to see how the bulk of the projection of the CMIP6 ensemble compared to the CMIP5 ensemble. To clarify, we've amended the figure caption to read:

Figure 3: "Note that in panel b, a broken axis is used to emphasize the structure of the primary CMIP6 model core with respect to the independent constituents, MIROC6 and MIROC-ESL."

16. Between page 16 and 17 it seems that some running text is missing.

Thank you for pointing this out. This may be a function of Figure 3 taking up all of page 16 so the running text jumps from page 15 to page 17. In the revision, we hope that this resolves based on our changes.

17. Lines 434-436: Similarities might be caused by the use of similar ocean models. Distinct versions of the same OGCM (NEMO) are used in the CNRM and IPSL GCMs (see Brands et al. 2023 for further details).

    Thank you! We've added the reference as:

L423–426: “Similarity in these cases cannot be traced to a particular atmospheric component model, but for CNRM and IPSL, similarity could have arisen through an effort to foster collaboration between the two French modeling groups after CMIP5 (Mignot and Bony, 2013) or due to similarities in ocean component model (Brands et al. 2023).”

References

Boé, J. (2018). Interdependency in multimodel climate projections: Component replication and result similarity, Geophysical Research Letters, 45, 2771– 2779. https://doi.org/10.1002/2017GL076829

Brands, S. (2022). Common error patterns in the regional atmospheric circulation simulated by the CMIP multi-model ensemble. Geophysical Research Letters, 49, e2022GL101446. https://doi.org/10.1029/2022GL101446

Brands, S., Tatebe H., Danek, C., Fernández, J., Swart, N. C., Volodin, E., Kim, Y.H., Collier, M., Bi, D., Tongwen, W. (2023). SwenBrands/gcm-metadata-for-cmip: First standalone version of GCM metadata archive "get_historical_metadata.py" (v1.1). Zenodo. https://doi.org/10.5281/zenodo.7813495

Sobolowski, S. et al. (2023). EURO-CORDEX CMIP6 GCM Selection & Ensemble Design: Best Practices and Recommendations. Zenodo. https://doi.org/10.5281/zenodo.7673400