Reviewer 1:
This is my second review of the manuscript “Uncertainty-informed selection of CMIP6 Earth System Model subsets for use in multisectoral and impact models” by A. Snyder and colleagues.

Thank you to the authors for extending the paper at several places and including more context. I think I understand the authors aim with this method better now.

Overall, I only have small technical comments left and think the manuscript can be published in its current form.

Thank you for your feedback. Our responses are in blue and **all line numbers referenced refer to line numbers in the new track changes manuscript.**

Minor comments

Table 2: The authors could have a look at ECS values from Mark Zelinka, which I think includes the missing models: https://github.com/mzelinka/cmip56_forcing_feedback_ecs
Thank you! We have sourced these values, added this citation, and re-run our analyses to include these models as candidates for selection (they were always included in the set of models making up the full data in each experiment). The final subsets for each experiment were unchanged.

Figure 1: I assume the figure title of the right panel should not read ‘all ESMs’?
Also, in the rest of the manuscript, the authors have switched to referring to their model collection as ESM/GCM.
Thank you for catching this, we have corrected this typo.

Figure 2: IASD is never introduced as far as I can tell.
Thank you for catching this, we have added the acronym to the text before Figure 2 where we first note that interannual standard deviation is one of our indices (~L178).

Figure 6: IAV is never introduced.
Thank you for catching this, we have added the acronym to the text where we introduce the Hawkins and Sutton calculations (~L347)

Reviewer 2:

General Remarks

This study has benefitted from the first round of revisions, and I find the methodology much easier to follow as part of the main text. It is an interesting approach to model selection; the uncertainty partitioning aspect clearly took a lot of effort and will be useful for many end-users of CMIP6. However, I am searching for more assistance in interpreting the results from Figures 2 through 7 in the text. I'm not sure what aspects of the figures I should be looking at, there is very little discussion of the similarities and differences between Experiment 1 and Experiment 2, and I am missing the justification for why the selected subset is superior to other possible subsets. It may be too much to compute Figure 5 for all 72 subsets you are considering, but it would support the selection you've done if there was some comparison between the subset you selected and the ones you did not. I have detailed places where results can be elaborated on further, as well as a few figure style suggestions, in the specific comments portion of this review. Thank you for taking the time reviewing the revised manuscript. Our responses are in blue and **all line numbers referenced refer to line numbers in the new track changes manuscript.**

With regard to the above, we have clarified the goal of our selection criteria in the manuscript (L112-122), and the selected subset of models meets that criteria the best compared to other possible subsets: to first represent (via PCA) a space that maximizes total variance of the full set of data in each experiment and to second select models that cover the range of that space while also preserving the IPCC distribution of ECS values. We are not evangelizing about a particular subset of models being the best but we are proposing a method that can be applied to new sets of simulations/models/regional discretizations and showing how it works for the choices outlined in Table 1. We have added text to our conclusions to re-emphasize this at the close of the paper (L473-474). We added the "post-facto" description of how our selection compares to the full ensemble when the Hawkins and Sutton partition is performed as a sanity check, looking for overall agreement across regions. We submit that the overall result is satisfactory (across regions and the two variables). Of course there will always be some specific region where the subset of models does not closely represent the full ensemble, given the relative strengths and weaknesses of individual ESMs, and if the selection was done with specific regions in mind it would be likely different.

Specific Comments

L34: (CMIP; Eyring et al 2016) > in LaTeX, (CMIP; \citealp{Eyring})

L35-36: Same as L34

L38-39: \citep[e.g.][]{X,Y,Z}

L63-64: Same as L38-39

L73-74: Same as L38-39

Thank you for highlighting the above copyediting. We have corrected them.

L135: Table1?

Thank you, we've corrected this.

Table 2: ECS values are available for your missing models:

- CMCC-CM2-SR5: Values reported in the IPCC's Assessment Report 6 Working Group I Chapter 7 Supplementary Material (The Earth's energy budget, climate feedbacks, and climate sensitivity) Table 7.SM.5.
- EC-Earth3-Veg-LR and FGOALS-g3: https://github.com/mzelinka/cmip56_forcing_feedback_ecs
- NorESM2-MM: Seland, Ø., Bentsen, M., Gra?, L., Olivié, D., Toniazzo, T., Gjermundsen, A., Debernard, J., Gupta, A., He, Y., Kirkevåg, A., Schwinger, J., Tjiputra, J., Aas, K., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., and Michael, S.: The Norwegian Earth System Model, NorESM2 – Evaluation of the CMIP6 DECK and historical simulations, https://doi.org/10.5194/gmd-2019- 378, 2020a.

Thank you! We have sourced these values, added the citation to Zelinka (as this covers all 4 models) and re-run our analyses to include these models as candidates for selection (they were always included in the set of models making up the full data in each experiment). The final subsets for each experiment were unchanged.

L159-161: Is it fair to compare individual realizations to an ensemble average for things like interannual standard deviation? Additionally, how do you handle the fact ensemble spread in precip. is much larger than ensemble spread in temperature for many regions?

We believe it is fair, since standard deviation is first computed on each individual ensemble member available and then these standard deviation values are averaged across the ensemble. We have further clarified this order of operations in the manuscript (L175-183). Doing the ensemble average first and then taking the standard deviation of the resulting time series would have issues along the lines you highlight, that an ensemble average from a size of 1 would potentially have much higher IASD than an ensemble average from a size of 50. As we do it in the manuscript, we feel it is fair (although of course larger ensembles give a more robust estimate of ensemble average IASD). As for different spreads, all PCA analyses in this work are conducted on centered and scaled variables, as noted in the legend of figure 2.

Figure 1: Can the side-by-side panels be on the same y axis scale? Additionally, the figure titles are identical, is this intentional?

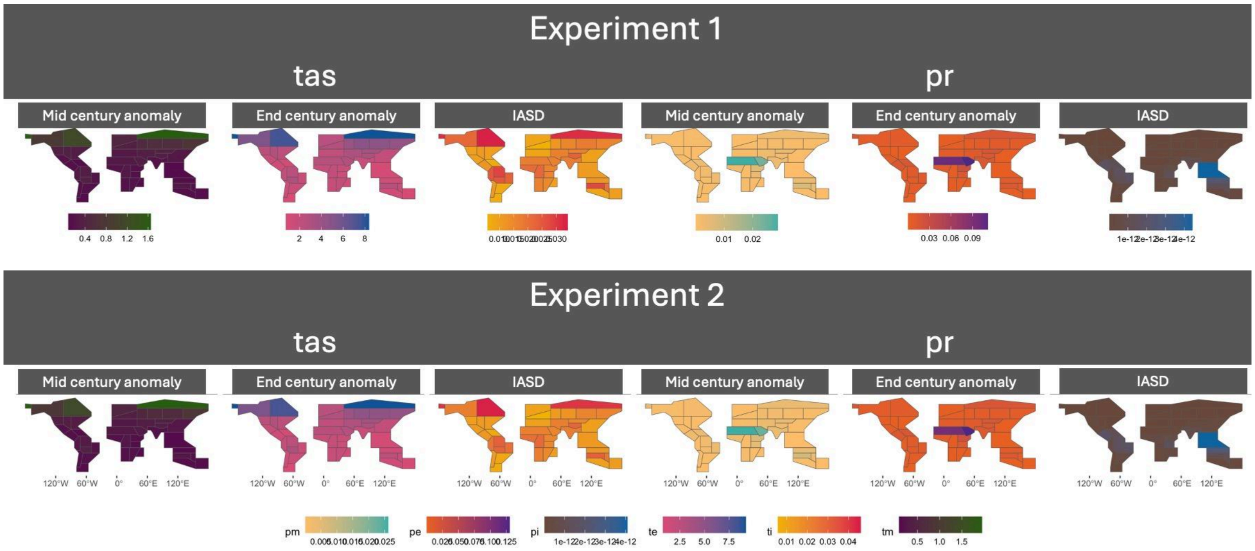
Thank you for this suggestion and for catching our typo in the titles. We have adjusted the y-axes to be consistent and corrected the title.

L206: Can you give a sense of what the sign of the PC represents? Could you show the full ensemble variance spatially? That might help with the interpretation of the EOF.

Thank you for raising these questions. We have attempted to add text clarifying this (L238-249).

But briefly: with eigenvectors, the sign of the total vector (as represented by all six maps together in each row) is irrelevant (if v is an eigenvector, so is -v with the same eigenvalue). For the components within each eigenvector, eg the temperature IASD map being quite red in PC2 compared to the lighter overall blue mid century anomaly map in PC2, the sign is also not directly meaningful because the variables are centered and scaled. The critical piece in understanding the EOFs is tracking dominance, for lack of a technical term, in each PC. One can interpret PC1 as showing temperature being the driving factor for that fraction of total variance, PC2 primarily as showing the importance of internal variability in explaining that portion of total variance, precipitation trends being mostly what is explaining the portion of total variance explained by PC3, etc. These are good sanity checks (e.g. temperature is the most important thing in explaining the biggest fraction of total variance in climate data, we probably expect that) but interpretation is not as relevant as the mathematical fact that PCA results in an orthogonal coordinate system that maximizes total variance.

We did check the plots of total variance across models and scenarios for each of the six indices in each region, plotted here, but we are not sure what value it would add to the paper to include. Note that because this is on the raw data rather than the centered and scaled data the PCA was performed on, each index has its own units and ranges and so of course does the variance of each index.



L215: Can you elaborate on “strikingly similar”?

We have added additional text (L254-257).

L222: What features?
Thank you, we have clarified (L265). We simply mean that having dependent models with, for example, shared cloud physics in the full data means that the full data has a bias towards that physical representation of clouds compared to other representations present in fewer models.

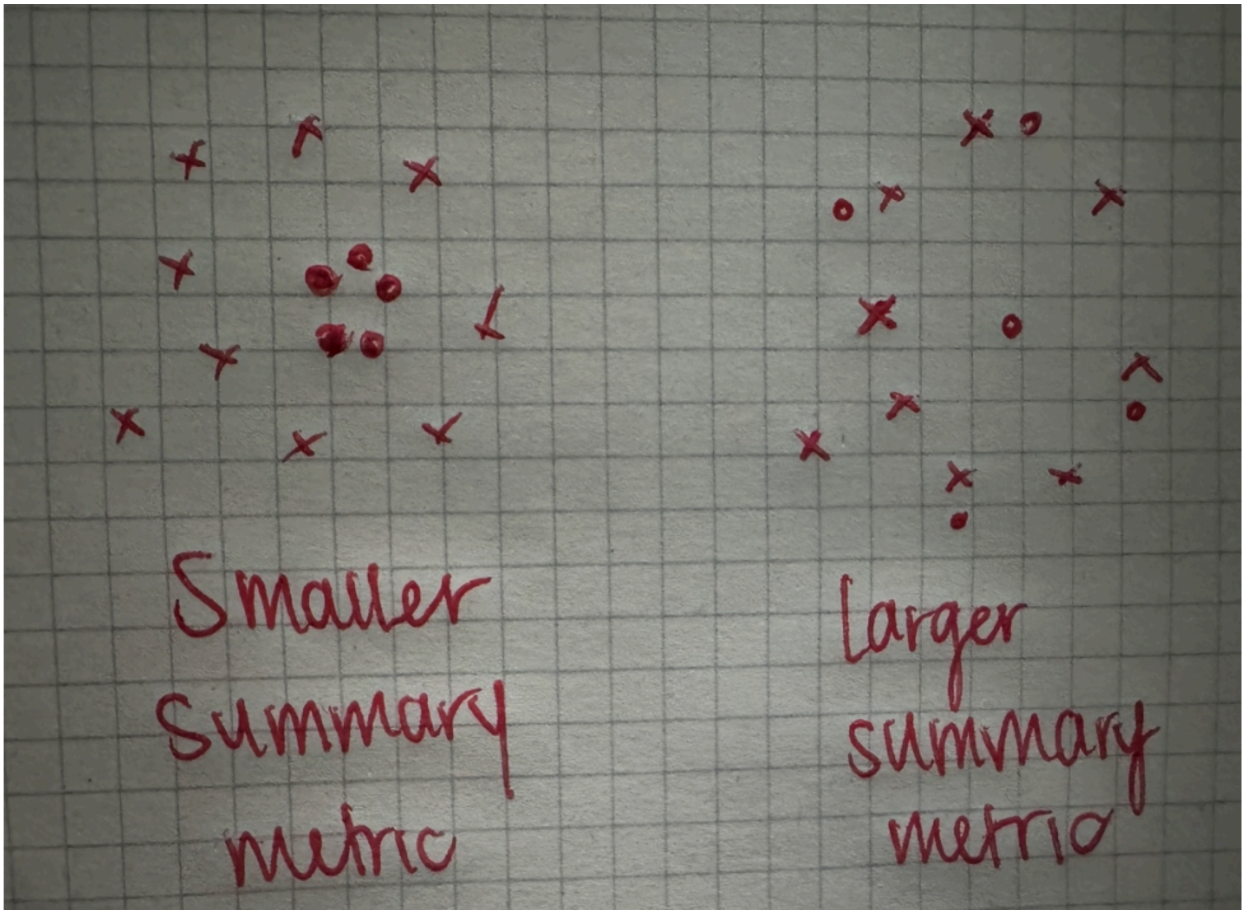
L235: Can you comment on what is happening with MRI is PC2?
Figure 3 and Figure 4: These scatters are not so legible, and I'm not sure what is to be gained from scattering each PC against the others? Would scatters of each PC against total variance illustrate the message?

We agree that the figures are dense and the message we hope to convey with them is not well-illustrated currently. We have clarified in the text that the point of these figures is less the particular values plotted, and more that the cloud of points together represents a five-dimensional surface we are intelligently selecting models to span from those currently available (L278-291).

These are not quite scatters against the PCs, they are projections of each model's index data into the vector space defined by the PCs, i.e. plotting the cij values from L203. We have clarified this in the text as well (L281-286). This is the 5 dimensional cloud of points that our metric selects spanning representatives of. We aren't sure what you mean by scattering a principal component against total variance. Figure 1 does plot the fraction of total variance that each principal component explains.

L258-260: I think I am misunderstanding. Is the idea to have the subset sit at the center of the distribution? Or to cover the spread? I drew a scenario that I think illustrates my confusion about the summary metric.

We have clarified the goal of our selection criteria in the manuscript (L112-122). It is closer to the latter.



L305: stray box in the equation.

Thank you, we have corrected this.

Figure 5: Why do you think you lose temperature agreement in so many regions in Experiment 2?

It is likely reflecting that the full data is different in Experiment 1 than it is in Experiment 2, and therefore so are the respective distributions of ECS for each Experiment's full data. So effectively, the baseline data against which our subsets get compared are different. Therefore the effect of a constrained ECS distribution is just different relative to the ECS distribution of Experiment 1's full data than to Experiment 2's full data.

Figure 6: My read here is that in all the regions you flag, model uncertainty is always under-represented by the subset (due to the constraint on ECS you impose) and the partition

between scenario uncertainty and interannual variability in the subset approaches the full ensembles over time, but scenario uncertainty of the subset is always under that of the full ensemble in 2040. Is this to be expected? Do we see some cases where the subset has more scenario uncertainty than the full ensemble early in the record? Though model uncertainty is shifted with respect to the full ensemble it seems to evolve through time in a similar way in most cases. Isn’t that more important than just a RMSE < 0.1?

Thank you for highlighting this. We have clarified in the text that the RMSE threshold is primarily a way to manage inspection of so many time series (L406-412 and L446-448), and we have further emphasized the point you raise - that matching the overall evolution in time is more critical than exact agreement. We actually do not think that scenario uncertainty is less at 2040 than in the full ensemble. Scenario uncertainty is given by the difference between the solid and dashed curves which behave in most cases indistinguishably from the green wedge at 2040, and always consistently with the expectation that scenario uncertainty would be the lesser source of uncertainty, often close to negligible, at that time. We did compare the scenario uncertainty fractions for the subset and full data and in most regions from 2035-2045, they differ by only 1-2% of total variance attributed to scenario uncertainty. I.e. the subset might say 7.5% of total variance in a region is due to scenario uncertainty in 2040 where the full data indicates 6%. As you point out in your comment, from a wholistic perspective those values are saying the same thing: in 2040, a very small fraction of total variance in every region is due to scenario uncertainty, which we expect because the 2030s-2040s is when the ScenarioMIP experiments considered in this work (SSP126, 245, 370, 585) begin to diverge meaningfully.

Figure 7: Again, I see a difference between cases where the subset is shifted down w.r.t. the full ensemble partition (e.g., Experiment 2 ESB) and cases where the partition is fundamentally different in time (e.g., Experiment 2 EAS). Have you investigated why this might be in more detail?

We feel this is outside the scope of this work as the focus of our work is proposing a methodology. We are proposing a method that has some good qualities in a global sense, so we are not expecting perfect performance everywhere. We emphasize in the introduction and conclusions that the specific choices around spatial regions and discretization considered in this work are driven by our experience with models that require global coverage. We often accept issues in specific regions in these applications in exchange for the coverage needed. We also have added text regarding this to the results (L450-455).

Reviewer 3:

Review of esd-2023-41

Title: Uncertainty-informed selection of CMIP6 Earth System Model subsets for use in multisectoral and impact models

Authors: Abigail Snyder, Noah Prime, Claudia Tebaldi, Kalyn Dorheim

Overall Recommendation: Accept

This study, submitted to Earth System Dynamics, highlights a new approach to select subsets of CMIP6 GCMs for use in multi-sectoral and impacts modeling. The approach is novel and attempts to capture multiple sources of uncertainty in the subset and does have potential utility for multiple applications beyond what was mentioned in the manuscript. This is the second time I have reviewed this manuscript and I'm pleased that the authors addressed my previous comments. As my remaining comments are minor issues, I recommend moving forward with publication. I look forward to seeing the revised manuscript in print.

Thank you for your feedback. Our responses are in blue and **all line numbers referenced refer to line numbers in the new track changes manuscript.**

Minor Comments:

Line 135 – There's a typo where Tables 1, 2, and 3 are all mentioned in this line. I believe Table 2 was what the author's meant to refer too.

Thank you, we've corrected this. A victim of moving things around with track changes across platforms.

Lines 138-142: "Models for which we... (more details in Section 2.2)." – How do you know the distribution of ECS is preserved when those without an ECS value in literature are removed? Presumably, these models have an ECS value, it is simply that the ECS values for thos GCMs have not been assessed by other literature.

This is a very good point.

Each subset considered is a set of 5 models with available ECS distributions that satisfy the IPCC distribution. You are correct that other subsets including the models we did not previously have ECS values for do exist that satisfy this. We have attempted to further emphasize in the conclusions that this is a method-focused work that scientists can adapt for information they have access to and prioritize, rather than a claim that our final subsets are in anyway 'the best'. They are simply the most numerically optimal at this particular task with the particular choices outlined in table 1.

Incidentally, the other two reviewers did provide citable ECS values for the 4 models we did not have previously. We re-ran our analysis to include subsets with these as candidates (considering a total of 150 subsets compared to the previous 72 for experiment 1), and the resulting selected subset in both Experiment 1 and Experiment 2 did not change. Two of the four newly added models (CMCC-CM2-SR5 and FGOALS-g3) are dependent with models we previously considered for selection in Experiment 1 and the projections between these dependent pairs of models are quite similar in the eigenvector space (which always was based on all available models, independent of ECS values), so it is in retrospect not surprising that new subsets with those two models were not selected to span the space efficiently. For subsets containing the other two of the four models (EC-Earth3-Veg-LR and NorESM2-MM), it is more down to chance that subsets containing them were not selected.

Line 189 – Like the comment for Line 135, there’s a typo where Tables 1, 2, and 3 are mentioned in the same line.

Thank you, we’ve corrected this.

Lines 199-201: “Based on this figure...flexibility of this method.” – The authors did respond to my question with respect to the number of chosen eigenvectors. I suggest including the response here rather than only in the response to the reviewers.

Thank you, we have done so (L228-232).