

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.

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

L135: Table1?

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., Graff, 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.

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?

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

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.

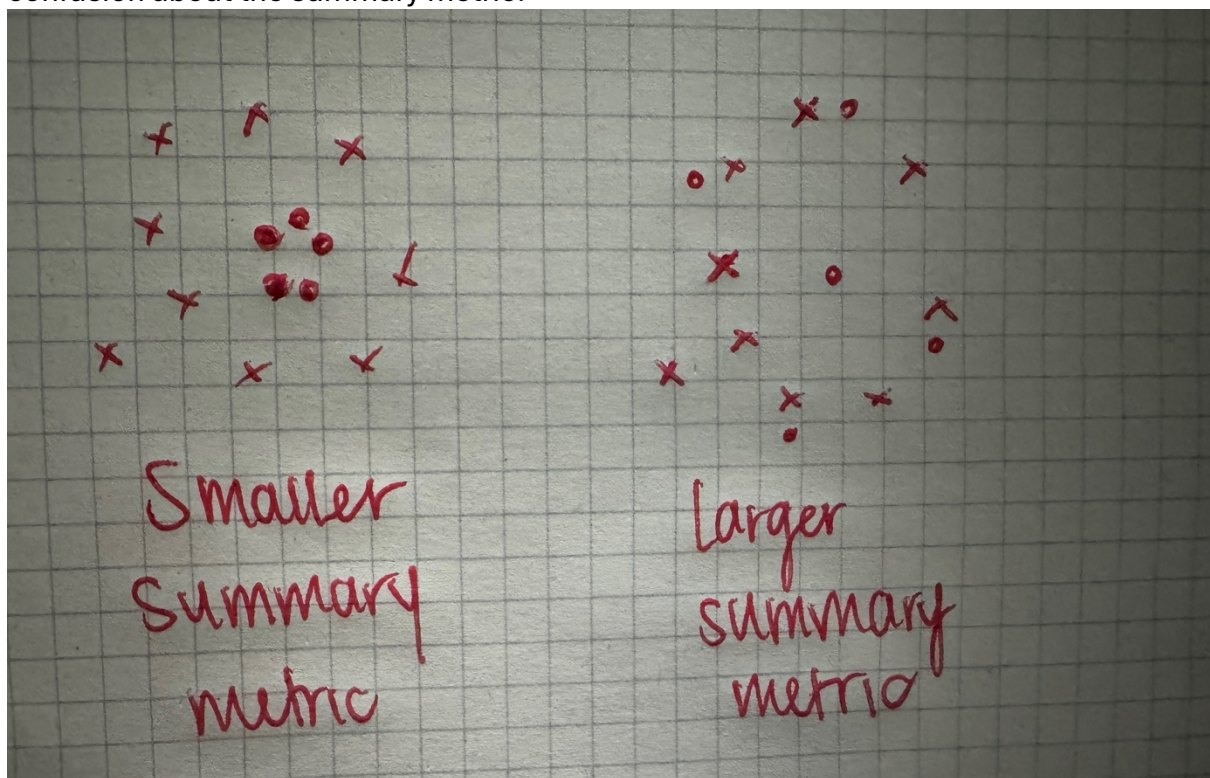
L215: Can you elaborate on “strikingly similar”?

L222: What features?

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?

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.



L305: stray box in the equation.

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

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 $RMSE < 0.1$?

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?