This manuscript introduces a method (ClimSIPS) which select subsets of CMIP models based on model independence, model performance and spread. In the second part the authors describe a case study for European summer and winter.

The manuscript fits the scope of GMD and is very helpful by dealing with the large ensemble of CMIP5 and especially CMIP6 models. Additionally, the change of the ECS distribution when considering only one model family member is very interesting. Nevertheless, the manuscript has reached an extreme length and is written very detailed. I am wondering if there are places where the text could be shortened.

Thank you for your review! We really appreciate your takeaways from what we agree is a bit of an interminable read. Following your feedback, we've made some substantial cuts to the article, listed as follows:

- Moved CMIP5 subselection (Figure 11) to the supplement and removed discussion of EURO-CORDEX as a benchmark from the main text
- Shortened the paragraphs on robustness and model agreement in Section 1.1
- Removed lists of initial condition ensembles used in the construction of the intermember distance metric from the main text
- Shortened discussion of within-model vs. between-model spread masking
- Reworded sentences to be more concise throughout
- Improved mathematical notation and added equations to be more precise with the cost function terms.

In total, we have reduced the length of the paper by several pages, even with additions requested during the review period.

Some small comments:

Line 53: Capital letter in the beginning of the sentence: "Modeling centers..."

Thank you; fixed.

Line 86/87: A verb is missing in the first part of the sentence.

Thank you for the catch. We've revised as:

L72–73: "Initial versions of ClimWIP based performance and independence definitions on the same set of predictors, which lead to concerns about convergence to reality."

Figure 1: The quality of the figure is quite bad and difficult to read.

That's very good to know, thank you! We've increased the dpi of the png image from 300 to 800 to rectify this.

Line 568: Is there a special reason for choosing this reanalyse dataset?

There was not a special reason for using the observational datasets we used to demonstrate the method. The reanalysis datasets were chosen based on availability at the time of method development under the assumption that we would perform further sensitivity testing to account for observational uncertainty. As we continued to develop the method, though, it became clear that this exact definition of performance was becoming a tangential avenue of inquiry. In

ClimSIPS, the performance metric is the simplest to swap out; models simply need a scalar rank, which can be obtained in many different and interesting ways. We envision most users will want to define their own performance metrics and therefore decided to focus method evaluation energy elsewhere (e.g., on automating selection of the spread-maximizing ensemble member from each SME).

To address this in the paper, we've added the following:

L560–562: "We found using a single observational estimate for each predictor to be sufficient for demonstrating ClimSIPS; the method's sensitivity to representations of observational uncertainty, different predictor combinations, and alternative performance definitions all warrant further exploration."

Line 945 and 947: Is this grade of precision of the numbers really needed here?

This is a good point. It is definitely not needed for the combinatoric explosion argument. We've amended the sentences to read:

L920–923: "34 choose 5 subselection iterated over more than 16 billion cost function values, which, run in parallel, took approximately two hours to run on 24 cores. Not evaluated here, 34 choose 10 subselection, with the cost function computed 6.6×10^{12} *times, would take considerably longer, an estimated three weeks to run, even in parallel on 48 cores."*