GMD Review 1

The authors describe an approach for analysis of inter-dependence of CMIP climate model simulations, their weighting and sub-selection for different purposes based on desired range of spread, performance and dependency.

The manuscript is well written and fits the scope of GMD. I find the developed methodology innovative and useful.

One of the important findings is that the model performance is rather a “model” characteristic, whereas the spread is more diverse for individual members of an ensemble of the same model. Further, the reduction of the spread of ECS after the family-democracy is taken into account is also a very important conclusion.

Please find below comments that should be addressed before the paper is accepted for publication:

We’d like to thank you for taking the time to help us improve our manuscript; we are thrilled to hear you find the methodology useful! All comments have been addressed in the text and as indicated below for quick reference.

line 234 – 237: I suggest explaining better that the “multi-model ensembles” correspond to “families”, e.g. replacing the word “ensembles” with “families”.

Thank you for the catch, we’ve changed ensembles to families to maintain consistency.

L209-212 : “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).”

line 260-265: the results of the sensitivity testing are shown somewhere? it should be stated explicitly (e.g. “see below”)

This is a good point. We’ve decided to make the sensitivity testing more transparent by including an additional figure in the Supplementary Material. In the main text, we’ve updated the following in reference to the new figure:

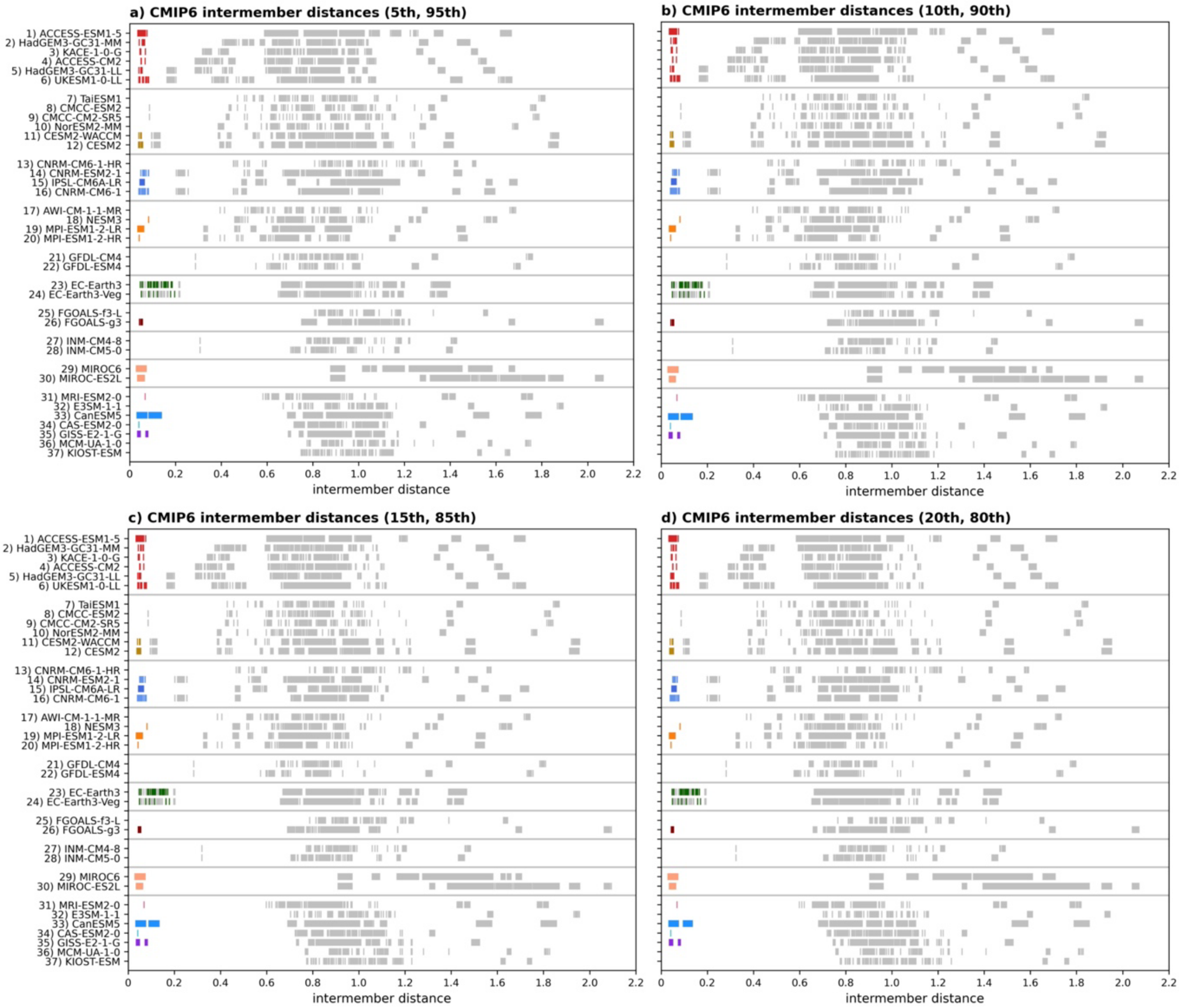
L246-248 : “To first order, I_{ij} is robust to methodological choices; the sensitivity testing did not reveal major shifts in whether a model was considered relatively dependent or independent with respect to the other models in the ensemble (See Figure 1, Supplementary Figure S1).”

L308–310 : “Results are not highly sensitive to precise percentile thresholds used to exclude regions of low between-model spread and high within-model spread; intermember distances are largely consistent for thresholds between the 5th and 20th percentile for between-model spread and the 80th and 95th percentile for within-model spread (Sup.Fig. S1).”

To the Supplementary Material, we’ve added:

L15-23: “While developing the fingerprint mask, we explored sensitivities to the percentile thresholds that define "low" between-model spread and "high" within-model spread. Shown in Supplementary Figure S1, we varied the threshold to mask between-model spread at or below the 5th, 10th, 15th, and 20th percentile. In concert, within-model spread was masked at or above the 95th, 90th, 85th, and 80th *percentiles. Intermember distances were similar in the four cases. They primarily differed by how closely members of initial condition ensembles group together. Ultimately, we chose the 15th and 85th percentile thresholds to define independence but would have obtained similar results with the 20th and 80th percentile thresholds. However, we felt that masking 40$\%$ of the domain began to challenge the notion of global similarity in the independence predictor fields and thus moved forward with the 15th and 85th percentile thresholds.”*

The new figure is captioned as follows:



Supplementary Figure S1: *“A comparison of CMIP6 intermember distance sensitivity to the definition of "low" between-model spread and "high" within-model spread. Regions at or below/above the following percentile thresholds are masked: below the 5th and above the 95th (panel a), below the 10th and above the 90th (panel b), below the 15th and above the 85th (panel c, used in the study), and below the 20th and above the 80th (panel d). For each model, distances between initial condition or perturbed physics ensemble members are marked in color, and distances to members of the remaining models are marked in light gray.”*

line 270 – 274: I suggest shortly mentioning that the benefit of longer time period is not visible for all models, denoting the contradictory result of EC-EARTH3 models, for which there is still the overlap even for the longer period.

We agree that this is the right place to identify what is happening with EC-Earth3. We’ve amended the text to read:

L254-263: “The grouping effect of the longer predictor averaging period helps to further distinguish initial condition / perturbed physics ensemble members from members of other models (Fig.1, light gray) in most cases. This differentiation is particularly clear in the case of CESM2-WACCM. The longer climatological averaging period distinguishes its three ensemble members from those of CESM2; with the shorter period, the two CESM2 model variants overlap (Fig.1, models 11 and 12). In contrast, though, the longer averaging period fails to subdue internal variability enough to differentiate EC-Earth3-Veg from its base model, Earth3 (Fig.1, models 23 and 24). The remaining internal variability in EC-Earth's global SAT and SLP fields is traceable to oscillations in the EC-Earth3 preindustrial control run from which both model variants are branched (Döscher et al. 2022). Functionally, this means that despite differing by coupled dynamic global vegetation, EC-Earth3 and EC-Earth3-Veg would be identified as one model by our independence metric.”

line 294: I suggest adding a note that the concept of fingerprints will be explained further below.

This is a good point. To clarify the terminology “fingerprint”, we’ve reworded the passage to:

L282-283: “Further, spatial masks can be explicitly designed to leave behind "fingerprints" tailored to meet dependence objectives. Here we design a spatial fingerprint...”

line 546 – 547: why the evolution of SAT over Europe should be representative of the GCM’s ability to simulate correctly the response to aerosol forcing? There are also other factors to be taken into account, so why specifically only the aerosol emissions are mentioned here?

Thank you for bringing this to our attention, the way that we’ve worded things really overstates our rationale for including two SAT climatological periods. The idea more was to evaluate to what extent a model resembled observed European SAT during a period prior to and post- the adoption of air quality directives in Europe to ensure biases during either period were accounted for in the performance metric. We did not explicitly evaluate how models respond to aerosols and should not imply that. Therefore, we’ve changed the sentence to read:

L536–538 : “We employ two periods of annual-average European SAT climatology, 1950-1969 and 1995-2014, to establish (1) if notable European SAT biases exist in the period prior European air quality directives (Haug et al. 2004) and (2) if a model's "present day" European SAT is significantly warmer or cooler than observed.”

line 608: I recommend to explain a bit the term “pool” – it can be the whole multi-model ensemble or somehow pre-selected subset. The term pops-up suddenly and makes the reader a bit confused.

This is a good point. We’ve decided to adapt the notation over all as follows:

L601–603 : “The first step of ClimSIPS is for the user to decide the number n of selections (s_i) they would like to make from a selection pool of N available models (s_1,..s_N). In this study, we demonstrate the method by selecting subsets of varying sizes from selection pools of varying sizes, henceforth referred to as a "N choose n subselection".”

In general, we feel that the $s_i,…,s_N$ notation helps to highlight instances where we normalize based on the whole selection pool. This improves interpretability in equations 7-9.

lines 612-620: the notation "s_i" should be explained properly, that it denotes individual simulations.

We also hope the adaptation of the notation from the selection pool being represented by N to it being represented by $s_i,..s_N$ helps here some. In addition, we've added the following to indicate that s_i refers to individual simulations:

L603-605 : "To illustrate the method, we select two model simulations, s_1 and s_2 from a purposefully reduced five model selection pool, $s_1,..s_5$, in a 5 choose 2 subselection."

line 875: please add a reference to the proof of the statement "intermember distances within both CMIP ensembles did satisfy metric criteria". (is it shown somewhere or not shown?)

This statement is rather abrupt and definitely comes too late. To address this, we've added the following to the method description:

L332-338 : "In Figure 3, we show how intermember distances based on the sum of normalized RMSEs calculated from SAT and SLP fingerprints help to uncover model relationships within CMIP. Intermember distances are presented for each model in one dimension (Fig.3a,c) and, as recommended by Abramowitz et al. (2019), for the ensemble as a whole in a low dimensional projected space (Fig.3b,d). The second display strategy is appropriate because we find our intermember distance matrix meets the formal mathematical definition of a metric space. To be mathematically a metric, the distance from a model to itself must be zero, and distances between models must be positive, symmetric, and adherent to the triangle inequality, which states that the distance from A to B is less than or equal to the distance through an intermediary point C (Abramowitz et al. 2019)."

We then call back to the explanation with:

L856-860: "Additionally, climatological SAT and SLP fingerprints allayed a concern that computing RMSE distance between models does not require the overall collection of intermember distances to meet the formal mathematical definition of metric space (Abramowitz et al. 2019). We found that intermember distances within both CMIP ensembles did satisfy metric criteria, with all sets of three models upholding the triangle inequality of dist(A,B) <= dist(A,C) + dist(C,B). Intermember distances could therefore be both understood as distances and visualized in low-dimensional space."

Part II – a comment on ternary plots and recommended subsets: The ternary plots are definitely useful for the analysis of different selection criteria. An issue, that is not commented on, is that some of the subsets "reside" a large part of the triangle, whereas some other subset have only a small fraction of the triangle. In some cases, the subset minimizes the cost function for only very narrow intervals of the coefficient values.

Complexity is an interesting feature of the subselection triangles, and we agree that it is worth discussing further. The size of the selection region is determined by the distributions of the selection criteria. In our primary JJA CEU case study, all three selection criteria have more or less a "core" with a few outliers (e.g., MIROC6 and MIROC-ES2L for independence or E3SM-1-1 and CanESM5 for spread). Thus, to first order, "small" regions of the selection triangle tend to occur when balancing performance and independence (with 10% or less priority given to spread) or when performance priority begins to give way to the other two (~70% performance priority). In the first instance, many models have similar performance values and can be selected alongside the independent MIROC models to minimize the cost function as performance priority gives way to independence priority. In the second instance, performance priority no longer requires the highest performing model to be included and subsets are comprised of other constellations of models until spread maximizing and more independent models eclipse them in the cost function. “Large” regions of the triangle tend to occur once performance is not a key player anymore and nearly always involve the independent MIROC models or the spread-maximizing models like E3SM-1-1 or CanESM5. Because those models stand away from the core to such a degree, they minimize the cost function for large regions of the triangle. In short, the distribution of selection criteria matters and outliers create larger regions. Because of CMIP6’s “core”, the triangle is more complex.

As a discussion, we’ve added the following. To discuss region size in general:

L687–691: “The size, shape, and number of regions within the subselection triangle are determined by performance, independence, and spread distributions; the larger the selection pool, the more difficult it becomes to predict the combination of models that will minimize the cost function. A subset can minimize the cost function for a small region in α- β space or even a single value of α and β. Small subset regions are as valid as larger ones; they simply reflect that independence, performance, and spread are distributed such that there are several model combinations in contention to minimize the cost function in that region of the subselection triangle. Conversely, when a subset minimizes the cost function for a large region of the subselection triangle, it suggests that it is comprised of outliers given priority in the cost function to such an extent that other model combinations cannot reach the minimum.”

To highlight region size relative to the whole domain:

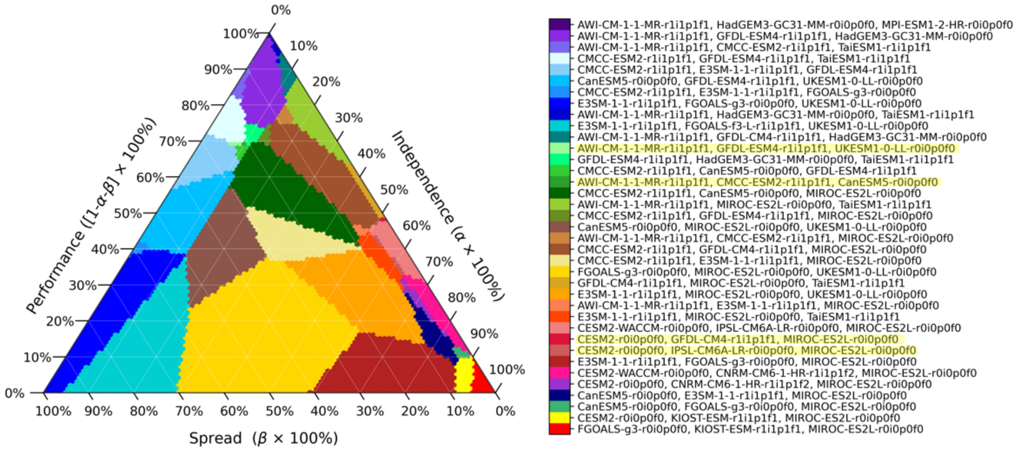
L756: “In total, recommended subsets cover 15% percent of the subselection triangle.”

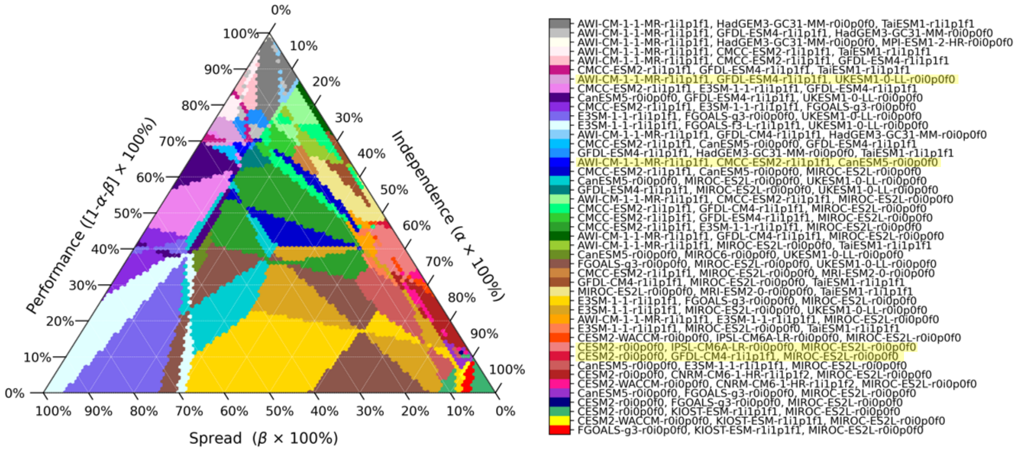
To discuss small recommended regions specifically:

L764-769: “Small recommended subset regions (<10 pixels in α-β space) occur at approximately 70% performance, 10% independence, and 20% spread, likely because performance priority has reduced enough to allow spread outliers like UKESM1-0-LL and CanESM5 to be in contention alongside various models within the core of the performance distribution. Similarly, small recommended subset regions near 50% performance and 50% independence result from the selection of various models in the performance core with the independent MIROC-ES2L.”

I suggest that it should be discussed, that in the case of the subsets that correspond to a very small fraction, there might be other subsets that have cost function values close to minimum and would maybe satisfy the criteria for a wider interval of the coefficients?

We looked into this by looking at the cost-function’s secondary minimum, i.e., the subset of models that is next in line to minimize the cost-function. For the JJA CEU 35 choose 3 by ensemble means case shown in Figure 8, the primary (top) and secondary (bottom) minimum subselection triangles are shown below:





The several pixel small regions that were recommended are highlighted. In all cases, they are expanded (in pixel / % of domain) by the secondary minimum as:

- AWI-CM-1-1-MR, CMCC-ESM2, and CanESM5
  - (1 / 0.01%) > (8 / 0.08%)
  - This subset has a similar cost function to:
    - CMCC-ESM2, GFDL-ESM4, and CanESM5
- AWI-CM-1-1-MR, GFDL-ESM4, and UKESM1-0-LL
  - (3 / 0.03%) > (40 / 0.39%)
  - This subset has a similar cost function value to:
    - CMCC-ESM2, GFDL-ESM4, and TaiESM1
    - CMCC-ESM2, E3SM-1-1, and GFDL-ESM4
- CESM2, GFDL-CM4, and MIROC-ES2L
  - (2 / 0.02%) > (5 / 0.05%)
  - This subset has a similar cost function to:
    - CMCC-ESM2, GFDL-ESM4, and MIRCOC-ES2L
    - CESM2-WACCM, IPSL-CM6A-LR, and MIRCOC-ES2L
- CESM2, IPSL-CM6A-LR, and MIRCOC-ES2L
  - (8 / 0.08%) > (81 / 0.79%)
  - This subset has a similar cost function value to
    - CESM2-WACCM, IPSL-CM6A-LR, and MIRCOC-ES2L

Several larger regions are also expanded by the secondary minimum, but it is not that the cost function minimum is unstable and the secondary minimum is stable, unfortunately.

Could there be some additional selection criteria that the recommended subsets should minimize the cost function for a larger fraction of the ternary plot?

This is an interesting idea and we are considering adding some penalties to, for example, remove subsets that include more than one family member from contention. An option to pre-filter by performance is already implemented. We are not sure how these sorts of penalties will affect complexity of the subselection, but with ClimSIPS, it is straightforward to explore!

It would make the selection more robust. In some cases, the recommended subsets are represented by only several "points" in the ternary plot (e.g. Fig 8). I believe that it is desirable to recommend subsets that would be useful for as wide range of applications as possible, to make projections used for similar applications physically consistent.

As far as robustness of the selection, we do feel that the beauty of CMIP is there is not a one size fits all answer for every study. There are just a few "good" models to use in all cases. As priorities shift, so do the combinations of models. But we do agree, physical consistency for similar applications is important as well. Alternative definitions of performance or spread in other variables could help coalesce subset regions.

Part II + Discussion and conclusion: Regarding the recommended subsets derived from CMIP5, the ClimSIPS method suggests similar subsets as used in Euro-CORDEX for driving regional climate model simulations over Europe. The authors claim, that this agreement implies, that their method is suitable for choosing subsets for driving RCMs. This implication is questionable, as it is not clear, what exactly was the basis for the choice of Euro-CORDEX driving GCMs from CMIP5. I do not doubt that proposed method is suitable for choosing appropriate subsets from CMIP6, I just do not agree with the comparison to CMIP5 subsets implying the suitability of ClimSIPS. Please, consider modifying the statements appropriately. The argumentation should be based on the nature of ClimSIPS, which is well described in the paper.

Thank you for bringing this up, and we are very happy to hear you find the method suitable on its face. Following this feedback, we've decided to move away from the CMIP5 subselection and only focus on CMIP6 subselection in the main text. CMIP5 subselection will remain for those interested in the supplement. We feel this move has made a very long paper a bit more manageable to read. The statements you mention have been removed from the main text and text descriptions in the supplement. Thank you again for your careful read of the study!

Language, copy-edits:

line 85 – Sentence beginning „Initial versions ..." – the verb is missing in the sentence.

Thank you, fixed.

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."

line 149 – „The study, an extension **of** the work ..." – the "of" is missing

Thank you, fixed.

line 508 – "in" is missing in "For use **in** cases..."

Thank you, fixed.