Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-143-RC2, 2019

Geoscientific Model Development Discussions

Interactive comment on "Uncertainties in climate change projections covered by the ISIMIP and CORDEX model subsets from CMIP5" *by* Rui Ito et al.

Anonymous Referee #2

Received and published: 5 September 2019

General comments

This manuscript aims to quantify the spread of CMIP5 projections and biases covered by the subsets of models used in the ISIMIP and CORDEX experiments. The first section of the results examines the spread of model performance in reproducing the temperature and precipitation over the historical period (1986-2005), relative to a range of observational and reanalysis products. The rest of the results examines the spread in projected end-of-21st-century changes in annual mean temperature and precipitation, and how it compares to the spread covered by randomly selected subsets. The main findings are that (i) the small ensembles used in ISIMIP and CORDEX generally

perform well over the historical period but are not optimal in minimizing historical biases, and (ii) the ISIMIP ensemble outperforms the CORDEX and randomly selected ensembles in covering the full CMIP5 range of projected temperature changes, but both ISIMIP and CORDEX cover a smaller spread of precipitation changes than randomly selected subsets.

This manuscript presents a valuable study to put the CORDEX and ISIMIP subsets in the context of the full CMIP5 ensemble. At this stage, it is mostly descriptive and would greatly benefit from a more comprehensive discussion, including the benefits/limitations of the metrics used, and the implications of its findings. Please clarify how this specific study sets itself apart from existing studies such as McSweeney & Jones (2016), and how your results fit into the context of the existing literature. Minor adjustments to language and sentence structure are needed to improve the readability of the manuscript.

Specific comments

Section 1 Introduction P2 L. 19: Please specify what these previous studies have found, and why we have yet to reach a consensus on the method to select small ensembles. Also, revisiting these papers in a discussion section would provide the necessary context to interpret the results and whether the methodology used in this study is distinct from, or improves upon, the ones used in these studies.

P3 L. 23-36 Please clarify what is specific to this study: is some aspect of the methodology new? Is this performing an existing analysis to a new set of data? Is the added value of the manuscript to specifically address whether region-specific subsets (CORDEX) outperform a globally consistent sub-set (ISIMIP)? Stating this explicitly would improve the value and readability of the manuscript.

Section 2.1 P4 L. 18 Why focus on land area only? Regional precipitation, including over the seas/ocean, is relevant for impact studies. Please clearly state the scope (and the application) of this study.

P4 L. 22 Can you justify why you excluded low-precipitation models from the precipitation analysis? I understand these models significantly bias your ensemble but this undermines the stated aim of the study (i.e. to quantify the spread of the full CMIP5 ensemble covered by the ISIMIP and CORDEX subsets). You arbitrarily reduced the model spread covered in this study. Please at least provide more information as to how many models were excluded (and thus the size of your remaining ensemble), why 0.1mm/day was chosen as a threshold, and the reasoning to exclude these models from the precipitation study but to keep them for temperature (if the argument is that the climate they produce is too unrealistic to be a plausible representation of today's climate).

P4 L. 32 Please include a definition of the skill score used here, or a reference to a published paper using the exact same skill score. In Taylor (2001), two examples of skill scores are used, to illustrate that the skill score can be adjusted depending on whether you value high correlation or matching the variability most. In addition, it is explicitly stated that the value of Ro should be reported every time a skill score is used.

P5 L5 Is R the min-max range of a given subset? Please clarify. Using 'uncertainty range' is misleading; it sounds like you are sampling your ensemble. If I understand correctly, you generate 10,000 values of RSub, then look at ensemble spread (in Fig 4).

Section 3 This section is entitled 'Results and Discussion' but mostly contains the description of the results. Regardless of whether it is included in this section or a separate section, the manuscript needs a more comprehensive discussion (see other comments below).

Section 3.1 P.5, L. 28-29 Please include the top 50% ensembles in the supplementary material, so that future model selection can rely on your analysis to select less biased models.

P5 L29-30 Can you suggest why the spread is different between the two ensembles

over the Northern Hemisphere? In Fig.1, the spread for ISIMIP is sometimes significantly larger, sometimes significantly smaller than CORDEX. Could this be due to the number of models in CORDEX? Do some of the regions have models that overlap across ISIMIP and CORDEX? Simply stating that they are different in some regions is not very informative.

P6 L1-3 This paragraph would benefit from an earlier explanation about how the skill and bias metrics are different and the insight gained by using both. Include some interpretation of why the model ensembles that perform relatively well in a bias metric perform less well in the skill metric. (same comment for P6 L11)

P6, L. 15-16 Please include the top 50% ensembles in the supplementary material. In addition, please include in the discussion whether the 'best performing models' perform well both in temperature and precipitation, and whether selecting according to high skill or low bias makes a difference. As you state that a better ensemble can be selected, please give the evidence from your results that this can be done robustly.

Section 3.2 P6, L.25 Please place this into context by mentioning other studies that have looked at emergent constraints, even if it's only in specific regions (e.g. Bracegirdle et al, 2018 for Southern Ocean winds; Bracegirdle and Stephenson 2013 for Arctic warming).

Section 3.3

In general, this section is confusing. It would benefit from clearly stating what is being compared, and referring to specific aspects of Fig 4 to support your statements. Specifically: P8 L1-2: Please clarify which metric you use to make that statement (i.e. the total coverage on the y-axis). I got confused because the performance of FRACORDEX remains low compared to FRARandom_C , even as the number of models increases. Please state explicitly where you are comparing it to the full range, or to Random_C (3 sentences later). P8 L16: Please specify which FRA you are talking about: the median of FRARandom_C,? Or FRACORDEX? Or both? P8 L19-22: This

is an interesting point, but if you make the point that increasing the number of models produces a higher FRA, please show the evidence for it. The latter part of the description is unclear so adding the technical details and a figure would make a stronger point.

Summary and Conclusions

This section provides a general summary of the findings, but would benefit from providing context as to how these results compared to other studies (e.g. those cited in the introduction), and how the findings advance the general understanding of the field. In addition, statements in P9 L4-5 and P9 L 15-16 seems to indicate CORDEX performance to be bad relative to randomly selected ensembles, while P9 L8-9 states 'relatively wide coverage of both uncertainties'. Please clarify so that the message is not ambiguous. For example, it is ok to state that CORDEX is not performing well compared to the randomly selected ensembles, but is marginally better than ISIMIP at sampling uncertainties in projected change in precipitation.

Please include a more comprehensive discussion of your methods and results, including:

Two metrics for "good performance" are used in parallel throughout the study (low bias and high skill score). Please comment as to how similar/distinct these two metrics are, and on the insights gained by using both (qualitatively or quantitatively). Similarly, how different are the 'top 50%' ensembles? i.e. does using skill or bias for selection of the best performing models significantly affect the ensemble?

The results section 3.1 focuses mostly on whether the ISIMIP and CORDEX fall within the observational spread (e.g. L. 2-3 on page 6). It would be helpful to distinguish whether this is mainly due to a large spread in the model ensembles, or whether a systematic bias is seen in certain regions (e.g. Fig 1 shows model ensembles overestimate precipitation in most regions). Also, please include a discussion of the expected variance of model ensembles. In coupled models, the timing of climate variability modes

is unlikely to match that of observations, so the variance over a 20-year period is likely to be higher in model ensembles than observations.

In this study, the performance of CORDEX and ISIMIP are considered independently for the temperature and precipitation changes (with precipitation being scaled by the temperature change). Please discuss whether there is any evidence that a selection on one variable (e.g. precipitation) is sufficient to select good performing models, or whether a combined approach is necessary to select models. In climate impacts, people care about the plausibility and diversity of climate sampled, not a single variable.

Technical corrections

P1 L18 (and after) High performed models -> high performance models or high-fidelity models

P1 L20-25 Please rework this section to clarify the meaning. As you have not previously introduced the 10,000 sampling strategy, these two sentences are confusing.

P2 L33 Please rephrase this sentence for better readability. For example: "In addition, paper X and Y showed that combining region-specific subsets covers more uncertainty than a single, globally consistent, subset of models."

P5, L11 Please rephrase that last sentence for readability.

P7, L9-12 This sentence needs to be reworked for readability.

P7, L15 "with those of the 10,000" -> remove "the"

P7 L16 "randomly sampled subsets" of what?

P9 L17 Be more specific: 'it depends on the number of models used' is too vague to be informative -> "FRA increases with the number of models used", or "regions covered by bigger ensembles generally have higher FRA"...

P13 L8 "areal mean of the reference data" -> normalized by the regional average of

GPCC data.

P14 Figure 2 Why does Antarctica have no top 50% in temperature? Explain that somewhere (main text or figure caption).

P16 Figure 4 “uncertainty range” -> range Also, why are red dots missing in some regions in Fig 4a?

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-143, 2019.