Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2020-23-RC1, 2020

Earth System Dynamics
Discussions

Interactive comment on "Reduced global warming from CMIP6 projections when weighting models by performance and independence" *by* Lukas Brunner et al.

Anonymous Referee #1

Received and published: 3 June 2020

Summary The authors present a methodology for weighting CMIP6 models based on several performance metrics as well as on their independence from each other. This provides narrower bounds on future global mean temperature changes than in the unweighted ensemble, primarily by down-weighting the highly sensitive models that happen to have poor performance with respect to two reanalysis products and/or are closely related to other models. I found the paper to be nicely motivated, well organized and supported, and a useful contribution to the literature. There are a few areas that I think need to be clarified, and so I recommend minor revisions.

Major Comments

* Figure 1 and the discussion around lines 241-242: the terminology of 0% to 100% trend-based seems too ambiguous to me and should just be written out explicitly. Couldn't the terms that are included just be stated explicitly in the figure? The figure doesn't really stand on its own, since one has to refer to these lines to know what exactly is meant by these. Additionally, it is not clear what the intermediate values (33%, 50%, 66%) correspond to exactly. Upon multiple readings, I still cannot understand what is meant by these percentages at all, and I'm not completely sure what is actually meant by "50% tasTREND and 50% anomaly- and variance-based diagnostics" that forms the basis of the remaining analysis. Please clarify.

* Discussion of Figure 2 around line 270: Should one have intuitively expected this from the math? I cannot seem to rationalize why using a model that is close to the CMIP6 MME to weigh CMIP6 would pull the CMIP6 MME mean away from the pseudo-observational "truth". This seems like a deficiency in the weighting. Shouldn't the weighting be resilient to this and do very little "harm" in this case?

* Figure 4: The combined and performance-only weights are shown, but not the independence weights. Is there a reason for this? Is it worth also showing the ECS or TCR from these models on this plot, so that one could see that higher ECS/TCR models tend to be down-weighted? I assume this is correct, to the extent that models that warm the most over the 21st Century have high ECS/TCR, but I don't recall the authors coming out and saying it. Modifying this figure in this way could be a compact way of making that point.

* Figure 4: I'm surprised to see several well-regarded models having relatively low performance weights (UKESM, HadGEM, CanESM, CESM), whereas some models that are typically poor performers seem to do well here (GISS, FGOALS, INM-CM). Any comment? Is it possible that your performance metrics are too restrictive (just involving tas and psl, two fields that may not adequately discriminate models with good vs bad moist physics that governs feedback and ECS), allowing poor performing models to get high weights?

Minor Comments

*line 61: should be “model’s”

*line 78: should be “method’s”

*Line 250: I don’t see where the 10-20% statement comes from. By my eye, the medians range from near 0% to slightly larger than 25%.

*Figure 1: titles should be “leave-one-out”

*Figure 2 caption: should be “which”

*Figure 2: To clarify, the similarity between pseudo-obs and MME is only assessed over the “Diagnostic period” right? (Side-note: “diagnostic period” only appears in the figure and is not discussed in the text.) By my eye, MPI looks closer to the MME than does CanESM, so I’m a bit confused here. Is the reason because similarity in the evolution of GMST only one of the several metrics employed, and MPI does worse in the ones that cannot be gleaned from this figure?

*Line 309: “allows us” or “allows one”; also, it seems like some reference to all the performance metrics work done by Gleckler et al seems appropriate here. I believe they also advocate for comparing against multiple observational datasets.

*Line 314: I don’t see the motivation for these 3 groupings. Is it in any way objective?

*Figure 6: too small to read, suggest stacking the two panels vertically rather than placing them next to each other horizontally

*Line 334: should be “model’s”

*I don’t think the average reader should be expected to know how to interpret a figure like Figure 5. Only the meaning of the colors are explained in the caption. What does the rest signify?

*Line 391 “The weighting also largely reconciles CMIP6 with 5”: what is this referring

to specifically, and is there a figure in particular being referenced?

*Figure 4: Are all weights less than or equal to 1 in absolute units, and only exceed 1 when expressed relative to equal weighting as is done in the figure? Otherwise I’m a little confused about why a model would have a weight in excess of 1. How exactly is wi used? weighted avg of X = sum(wi*Xi)/sum(wi)?

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2020-23, 2020.