Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2020-86-RC3, 2020

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Projecting ozone hole recovery using an ensemble of chemistry-climate models weighted by model performance and independence" *by* Matt Amos et al.

Anonymous Referee #3

Received and published: 19 March 2020

General Comments

The manuscript by Amos et al. introduces a method aimed at creating an alternative to the multi-model mean when examining an ensemble of model simulations. The weighting technique presented here accounts for both model performance and model similarity, with higher weightings given to models that compare well to observations and that are more independent of other models in the ensemble. The method is demonstrated using predictions of Antarctic ozone hole recovery from the future simulation of the Chemistry-Climate Model Initiative. The demonstration notably incorporates several observational metrics of species/physical parameters relevant for Antarctic strato-

spheric ozone.

A need for this type of analysis exists; it has been conveyed from many in the global modeling community that a simple multi-model mean, where all models are equally weighted, is not sufficient. Therefore, I consider the subject of the manuscript important, and the scientific findings presented through the demonstration of the method are relevant for ACP. I have some broader questions regarding the general applicability of the technique, but view the paper overall to be well-written, appropriate for the journal, and ready for publication after the authors have addressed the following minor comments. It is also a fitting submission for the CCMI special issue.

Specific Comments

L. 148 and throughout: What about when poor model performance manifests as poor simulation of the dynamics? If a model has an accurate chemical mechanism, maybe it looks good in the SD simulations, but it's poor when simulating the Antarctic vortex in free-running mode – doesn't that mean it should not be trusted to get the future ozone hole recovery right?

L. 194: How are the temperature metrics influenced by the SD versus free-running simulations? What would happen if you calculated the individual and total weights (i.e., Fig. 3) using Ref-C2 instead of Ref-C1SD? And, similar to the previous comment, what if the nudging in the SD run is what causes a realistic decrease in temperature, not the coupling between decreased ozone and temperature (i.e., if ozone is poorly simulated, but the nudging imposes realistic temperature changes, will a high weighting be awarded to this model, for this metric, despite getting temperature "right for the wrong reasons")?

L. 205: At this point, it's not clear if each of these metrics will be tested individually, or if all of them will be combined as in Eq. 2, or various combinations will be tested? Later in the manuscript, a dropout test is described; perhaps that should be moved earlier (under "3.3 Evaluating the weighting framework") to place it in the larger context

sooner?

L. 290: Are the results of this out-of-sample test shown anywhere? It's difficult for me to grasp what these RMSE values mean, in context, though I'd be curious to see the results of the test.

L. 336: On the sensitivity of the final weightings to which performance metrics are included: You performed a dropout test, leaving one metric out at a time. But, what about leaving out two? For instance, since CNRM-CM5-3 stands out so significantly in its Total Weight, what if you leave out the two metrics where that model apparently excels above the other models in performance, i.e., Polar vortex breakdown trend and Ozone-temperature gradient? I am concerned that the Total Weight for each model is highly sensitive to the combination of observational metrics included, beyond what is tested by the single-metric dropout test.

Technical Corrections

L. 50: "...and climate" should be "and climate forcers" or similar?

L. 200: "high" should be "highly" and "except of" should be "except for"

L. 252: Should "Figure 2" instead be "Figure 3"?

L. 267: "The total weighting, formed from the summation of individual metric weights..." Is this right? This conveys to me that all of the individual metric weights are simply totaled. Since CNRM has a couple weights >0.4, then its resulting total weight of 0.27 suggests this is not right... Found by Eq. 2 instead?

L. 273: "The lowest model weight is 55 % the value of a uniform weighting." I understand this to mean that the lowest red bar in Fig. 3 is 55% the magnitude of the dashed black line, but the smallest bars (CCSRNIES-MIROC3.2 and EMAC-L47MA) look to be less than half, judging by eye. Is this statement correct?

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2020-86,

2020.