Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2021-248-RC2, 2021

Comment on gmd-2021-248

Anonymous Referee #2

Referee comment on "An ensemble-based statistical methodology to detect differences in weather and climate model executables" by Christian Zeman and Christoph Schär, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-248-RC2, 2021

Review: Zeman and Schar, GMDD

The manuscript presents a null hypothesis testing approach to identify if non-b4b changes in a full complexity atmosphere model produces the same climate statistics as the original model.

In the current form, this study reads like an exploratory study with a lack of clear framework for automated testing - which seems to be the goal - and thus seems incomplete. For example, the authors do not prescribe which variables to evaluate, how many variables to use or how long the tests should be run or how many ensembles. While the authors hint at these in text in places, these aspects still lack clear answers. I think it will be best to do additional wider case studies - like those done by others and finalize the framework based on all the results rather than leaving it out for the future.

Also, there is little novelty in the work. While the authors evaluate the null hypothesis at each grid point for the atmosphere - which is a little different from the Baker et al. (2015), Milroy et al. (2018), Mahajan et al. (2017, 2019) and Massonnet et al. (2020) tests for the atmosphere models - the need for doing that is not clear and has not been explored in this

study. Atmospheric mean flow fields are highly homogeneous with longer correlation length scales. Fig. 2 is a good example of this which shows high spatial correlation of the 500hPa geopotential height. It may thus be important to argue for the need for this grid point based test more strongly. The authors say that it is more fine-grained and thus would help with debugging. I am not sure how looking at some grid points failing the test would help with debugging. I think a clear case needs to be made, if possible with examples/case studies. If not, I think a comparison with tests that use the domain averages (Baker et al. 2015, etc.) would help justify the need for these fine-grained tests.

The main difference from previous testing methodologies is the use of mean rejection rates that are derived from sub-samples of control and evaluation ensembles - essentially conducting an ensemble of tests. Other studies only use one test to make a pass or fail decision. However, other tests, for example Mahajan et al. 2019, do use such an ensemble of tests to detect the false negative rates, which is kind of similar to this approach. This difference should be pointed out more clearly in the paper.

Also, while the authors conduct several case studies, showing that the tests can catch certain small differences, it is not clear how small these differences really are. I think the authors need to pay more attention to the detection capabilities of the test. It may be good to look at more parameters that are used in other studies to establish the robustness of the testing approach.

I think this may be a useful alternative test to the existing methods, but it needs to be more formalized in its prescription with supporting results and comparisons with other studies.

Other Specific Comments:

Lines 155-170: Discussion of FDR approach. There are several approaches to FDR. See for example, Ventura et al. (2004). It may be good to cite these different approaches here given the nature of the discussion. Also, Mahajan et al. (2021) recently used the FDR approach for testing statistical reproducibility in an ocean model and found it to be quite sensitive. It is interesting to note that the atmosphere model does not show sensitivity to this approach - although details are not presented here. Nonetheless, It may be good to cite this work here, which appears relevant to this discussion.

Lines 190-200: Mahajan et al. (2017 and 2019) also used the Monte Carlo approach that is being used here, i.e. they also use a large control ensemble (100 or so members) to establish the rejection rates. They indeed found that this approach yielded similar results to pooling the ensembles together. The approach of pooling ensembles together is called permutation testing and it may make sense to use the term here for clarity. Also, the line, 'Depending on the difference between the two models ….' seems hand wavy. Please clarify or omit.

Lines 220, that paragraph. The FDR approach does not suffer from this issue of the arbitrariness of the significance level of the local null hypothesis, where the p-value is corrected based on the significance level of the global null hypothesis. This should be discussed here since FDR was discussed earlier in the text.

Computational costs for Monte Carlo tests. In a few places, the computational cost of running Monte Carlo approaches is mentioned. Given the current computers with accelerators, I think it is generally a weak argument. For example, conducting Monte Carlo tests for all the 150 variables, say used by Baker et al., should not be much of a computational hindrance.

It might be important to discuss situations where this test may be more useful than others, particularly those that evaluate longer runs and situations where it may be useful to run these other longer tests.

References:

Mahajan, S, 2021: Ensuring statistical reproducibility of ocean model simulations in the age of hybrid computing. In Proceedings of the Platform for Advanced Scientific Computing Conference (PASC '21). Association for Computing Machinery, New York, NY, USA, Article 1, 1–9. DOI:https://doi.org/10.1145/3468267.3470572

ValÃ©rie Ventura, Christopher J. Paciorek, and James S. Risbey. 2004. Controlling the Proportion of Falsely Rejected Hypotheses when Conducting Multiple Tests with Climatological Data. Journal of Climate 17, 22 (2004), 4343–4356. https: //doi.org/10.1175/3199.1 arXiv:https://doi.org/10.1175/3199.1