Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2021-248-RC1, 2021

Comment on gmd-2021-248

Anonymous Referee #1

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-RC1, 2021

Title: "An Ensemble-Based Statistical Methodology to Detect Differences in Weather and Climate Model Executables"

Authors: Christian Zeman and Christoph Schär

The authors describe an ensemble verification method that uses statistical hypothesis testing to assess the effects of minor model changes. Note that other statistical ensembles tests have been previously proposed (which the authors reference) and are in use for evaluating climate models. These types of statistical evaluations are important as bit-reproducibility is not a practical approach for verification (or correctness checking) given the chaotic nature of climate/weather models as well as modern heterogeneous architectures.

The authors' statistical ensemble approach uses a total of three ensembles: the so-called "control" and "reference" ensembles are both from the old or accepted model, and an "evaluation" ensemble is generated from the new model or configuration that is being evaluated for differences. It is important to note that their statistical tests are applied locally -- meaning at each grid cell. The difficulty here is that the local statistical tests can't be assumed to be independent because the data is very likely spatially correlated. (And climate models have many different variables, many of which will have different spatial correlation properties). Therefore, how does one decide on the global reject criteria? The authors rightly admit that this would be non-trivial to determine for all the variables, and for this reason, they generate the two ensembles from the same (old) model (control and reference) to compare and empirically determine the rejection rate distribution. This rejection rate distribution computed from the two ensembles (which are not different) then becomes the baseline against which the rejection rate from the new (evaluation) ensemble as compared to the reference ensemble can be evaluated (nicely illustrated in Figure 1 in the paper). On the whole, this process seems rather involved, though the authors do a nice job explaining their approach. Results are given for a number of experiments for which the anticipated output is then confirmed by the

approach. Unfortunately, the authors do not compare their test to other available ensemble statistical tests. Of course a comparison would be useful as well as interesting (the authors acknowledge this), but perhaps the amount of additional effort required is too high...

One aspect of this work that I do find concerning is that I do not have a sense of the robustness of this approach. In particular, with ensemble methods, the size of the ensemble typically directly influences the robustness of the method. In this case, I would expect that a much larger ensemble size would be needed to ensure that the empirically calculated rejection rate distributions do not vary too much with different sizes or samples (i.e., the box plots in Figure 1). My feeling is that this aspect of the approach needs more vigorous treatment as, at present, I would be hesitant to apply this method in practice. The authors do state that the choice of ensemble member, subsamples, and subsample members was "rather arbitrary" (line 209), but claim that they are "quite confident" that different choices will not significantly affect the behavior. I confess that I am skeptical of this claim and think that a more thorough evaluation of these parameter choices is needed to demonstrate robustness.

A second concern that I have is in regards to which variables to evaluate in practice. While the authors state that evaluating all of the variables would be "overkill", it is unclear to me how one would best determine the "key" set of variables to use for a particular application. The authors do recognize that "for a fully-coupled GCM, some further considerations will be needed", but it is unclear to me how they chose the key variables even in their limited-area tests (and how they know whether the set of variables they choose was sufficient).

An interesting aspect of this approach is the focus on evaluating the detectability of changes over time (several hours to several months). Looking at the rejection rates over time (e.g., Figures 3-8) allows for comparing the relative importance of the change being tested to the model's internal variability. Perhaps this approach gives insights that other methods may miss (though this is not demonstrated). Many (or most?) of the presented results show that modifications change from "rejected" to "not rejected" as time increases. While I quite like these plots, it is not clear to me in general how this should be interpreted (in practice) in terms of deciding whether the modification should be rejected or not. This change from "reject" to "not reject" over time really highlights the question of "what is this test for?". For example, there is a discussion in Milroy et al. (2018) about how changing a random number generator is a detectable change after several time steps, but not after a year. In that case, Milroy says that this type of change is unimportant to them as the overall climate statistics are consistent at a year. In this paper, I am unclear as to whether the authors are purposely trying to find changes that are rejected at early time steps even if later on in the simulation the changes are not rejected. Is this important to their evaluation? They say that the approach is particularly sensitive to small changes, but is this a good quality? And if so, why? Perhaps the answer is subjective depending on the variable or change being tested, but I believe this point needs to be clarified for the approach to be useful in practice. Particularly because, in my opinion, a strong motivation for statistical ensemble approaches is to remove as much of the subjectivity (i.e.,need for a climate expert) as possible from the decision on correctness when evaluating model changes.

---
specific suggestions/questions:

-Abstract: line 15: I wouldn't categorize the switch from double-precision to single-precision as "tiny" . Maybe this is not what was meant - if not, then please re-phrase or clarify .

-line 76: For Rosinski and Williams work, consider mentioning what model this was for and why this approach is no longer appropriate on the current model(s).

-line 103: Note that the POP CESM consistency test concept is almost entirely different as it does not use PCA - but looks at individual variables and at individual grid points (as mentioned) accounting for their location-specific variability.

-line 213: define "floored variables"

-last paragraph of 3.1: Will this method catch something that the other methods wouldn't?

-lines 248-249: Do you perturb all of those variables at the same time?

-line 263: Given the (somewhat large) choice of 1e-4, it would be helpful to show the ensemble spread over the first few hours as done in the Milroy 2018 paper that is referred to or somehow give more rigorous justification for this choice.

-Figure 1: This is a nice graphic. My only suggestion would be to add something to indicate that the control and reference ensembles are from the same model. (This is mentioned in the caption, but might be a nice addition to the figure.)

-line 484: how many variables are there in total?

---
minor items:

-line 4 (abstract): "unsuspicious" sounds awkward - maybe "innocuous" ?

-line 26: "are there" => "are intended"

-line 94: "from which many show high correlations with each other" => " many of which are highly correlated"

-line 118: "sensitivity" => "sensitive"

- line 150: suggest putting ()s around equation numbers to match equation label : "equation 1" => "equation (1)" (other lines in the paper as well)

-line 173: replace "went for" with more formal language

-line 182: "used local statistics" => "chosen local statistics" or
"selected local statistics" ("used" sounds like it has been tried
out already - like as in "used car")

-line 208: same as above (replace "used")

-line 232: "rejection" => "rejections"

-line 364: "Section" is misspelled

-lots of places: "floating point" => "floating-point" when used as an adjective

-line 395: "much less" => "fewer"

-line 444: "the the"

-line 475: (twice): "casted" => "cast"

-line 512: "be effective at" => "affect"

-Throughout: consider an editing pass - especially for missing commas