Comment to Topical Editor

Dear Prof. Dr. Christoph Knote,

Thank you very much for being topical editor for this manuscript and for managing the review process in such a straightforward manner. We have published the source code on Zenodo with a specific DOI and refer to it in the manuscript accordingly.

Thank you very much and kind regards,
Christian Zeman

Answers to Reviewer 1

Overall:

The authors have responded thoughtfully and adequately to all of my concerns and questions. The new manuscript is a significant improvement. The authors clearly put considerable effort into this revision. The additional figures and subsections are quite helpful, and the overall writing quality has improved.

We thank the reviewer for the thorough review of the changes and the good comments. We would also like to thank the reviewer again for the very useful initial review comments, which have helped a lot to significantly improve the manuscript.

A few notes:
---
(1) Figure 4: I quite like this addition that allows the easy comparison across variables
(also Figure 11).

Thank you.

(2) I like that there are 16 "key variables" now being investigated, as in Figure 4. The variable choice was explained in the reviewer response (e.g., "We now propose a set of key variables we think reflects the most important processes of an atmospheric model ...") and also is commented on in the discussion section of the paper (e.g., line 637). My suggestion is to add a sentence or two when Figure 4 is first introduced (~line 457) that comments on the choice of 16 variables.

Thank you. We have now added the following sentence around line 457:
“We believe that such a set of variables offers a good representation of the most important processes in an atmospheric model (i.e., dynamics, radiation, microphysics, surface fluxes) and, considering the often high correlation between different variables, is therefore likely sufficient to detect all but the tiniest changes in a model.”

(3) Section 4.9 (and Fig.10) : I appreciate the inclusion of the results on spatial averaging. I think the bars in Fig. 10 are showing the average of 100 randomly drawn samples, and I am wondering if there was more variability in the global rejection rate for the smaller ensemble and subsample sizes (on the right of the x-axis) than for the larger ones on the left (or is it the other way around). Just curious...
Also I am wondering about the reason why the 2 bottom plot scenerios in Fig 10 (the CPU and false positive ensembles) have less sensitivity to the spatial averaging than the diffusion modifications. The paper states that this result "is interesting" (line 565), and it may be that the D = .005 case essentially representing the largest perturbation, followed by the smaller D=.001, then the (likely smaller) CPU perturbation, and then the "smallest" (i.e., "no change"). It does seem to make sense that in the presence of little or no perturbation (so little variability), then the effects of different tile sizes would matter very little... I'm not requesting any paper modifications - just thinking about this.

While we have not really tested this, we think that the variability of the global rejection rate is likely going to be higher for the smaller ensemble and subsample sizes. Taking Fig. 9 as an example, the rejections are often not that clear for smaller ensemble and subsample sizes. Furthermore, we also think that the ratio between subsample and ensemble size will have an effect on this, as the quantiles are closer to the mean with a bigger ratio (when, for example, comparing 20/50 vs 100/200 vs 150/200 in Fig. 9).

The absence of any sensitivity of the CPU ensemble to spatial averaging is indeed interesting. We think that, in this case, the test is quite clearly not able to detect differences and, therefore, the number of rejections is very similar to the no-change ensemble, where the number of false positives are not affected by spatial averaging. We also think that not all changes will show such a high sensitivity to spatial averaging as diffusion. Some changes that can be detected might show no sensitivity to spatial averaging at all. However, this would have to be investigated further to give a clear answer.

(4) Appendix A: This is also a nice addition.

Thank you for the suggestion to add something like that.

Minor:
---
(1) line 372: "visibly or easily" => " visibly or be easily"
(2) line 451: "quite much" => "quite a bit"

Thank you, we have change it in the manuscript accordingly.