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

Authors: Christian Zeman and Christoph Schär

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.

A few notes:
----

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

(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.

(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.

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

Minor:
------

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