1 Overview:

Review of "A new ensemble-based consistency test for the Community Earth System Model" by Baker *et al.*

Baker *et al.* addressed many of the minor comments from the first round of reviews but has not sufficiently addressed two of the major comments.

2 Major comments:

Assumption of non-skewed distributions
Indeed, we are looking at 120 variables, and the distribution of all are not exactly symmetric, though most are close to normal. However, the RMSZ scores are not used to make the consistency determination (global means are used), but are provided to scientists as they provide insight in many cases (please also see response to Minor Comment #5).

This issue has not been addressed. It seems likely that many of the variables in their list ("http://www.cesm.ucar.edu/models/cesm1.2/cam/docs/ug5_3/hist_flds_fv_cam5.html") could have non-symmetric distributions. The authors could address this issue by appropriately stating the assumptions with this methodology in Section 3.5 (ie., tell the readers when this method is valid and when Z-scores are inappropriate). Otherwise, cut the discussion of RMSZ scores from the manuscript. Do they really need to include something in the manuscript just because it was used in the past (especially if the assumptions in the methodology may not be valid)?! Discussion of legacy code seems like it would be more appropriate for a user's manual than a journal article...

Is PCA necessary?
The PCA is necessary in order to determine straightforward and objective pass/fail criteria. For example, assume we have 26 variables labeled A, B, C,... Z. If variables A–F are linearly dependent, then an issue in the code could lead to 6 variables (A-F) failing. A different code issue could lead to only one variable (e.g., Z) failing. So to design a test cutoff based on how many variables fail would be arbitrary if the dependencies are not considered. Using PCA removes this issue by yielding a linearly independent set, and we can more easily compare the number of PCs that fail between different cases. More importantly, with PCs, we are able to determine (and satisfy) false positive rates, which is important to the CESM scientists. The cutoff of 3 PCs corresponds to the specified 0.5% false positive rate (section 3.4). One can change the parameters to achieve different false positive rates. We do not assume that the leading PCs are more important, as we have found subtle errors in the code that do not affect the large PCs, but are relevant (i.e., we want to detect them) nevertheless.

We agree that looking at 50 components (instead of 120) does not provide much computational savings, but it is not clear that looking at the remainder of the PCs would be beneficial as the first 50 represent nearly all of the variance. However, we note that 50 is only the default for the CESM-ECT python code, and a user can chose to

look at more or less if desired.

This argument strikes me as circular. The authors argue that the leading PCs are not necessarily more important ("We do not assume that the leading PCs are more important, as we have found subtle errors in the code that do not affect the large PCs, but are relevant") then in the next sentence go on to argue that the leading PCs are the important ones ("it is not clear that looking at the remainder of the PCs would be beneficial as the first 50 represent nearly all of the variance"). Are the small PCs important? If so, why are you only using 50 PCs?

I'm also not convinced that your choice of pass/fail is a good one. As it stands, your method would say that if the two largest PCs (which explain 40∼50% of the variability based on Fig. 1) were different in 2 (or 3!) out of 3 simulations and the other 48 PCs were the same then the simulation would pass. This seems like a major flaw. The leading PCs are almost certainly indicative of something real in the climate system.

As I said last time, this choice of 50 PCs strikes me as rather arbitrary. Why not use all 120 PCs with $N_{\text{new}} = 3$, $N_{\text{runFails}} = 2$, and simply define a failure based on $P_{\text{varianceExplained}}$. Where $P_{\text{varianceExplained}}$ is the total variance explained by the PCs that failed (e.g., $P_{\text{varianceExplained}} = 5\%$). So if one large PC fails multiple times OR a few small PCs fail multiple times then the simulation will fail. It should be easy to define a false positive rate in this manner.