Response to referee comments and updates for manuscript:
"The Ensemble Consistency Test: From CESM to MPAS and Beyond "

Teo Price-Broncucia[1], Allison Baker[2], Dorit Hammerling[3], Michael Duda[2], and Rebecca Morrison[1]

[1]University of Colorado - Boulder
[2]NSF National Center for Atmospheric Research
[3]Colorado School of Mines

We thank the editors and the referees for their excellent feedback. We believe that addressing this feedback has notably improved the manuscript. We are pleased to have the opportunity to resubmit to GMD for your consideration.

Referee comments are in *italics*. Our responses are in regular font. Changes to the manuscript are indicated in blue.

Referee #1

1. *I am surprised by the way the authors use the PCs by averaging out the variables over the whole domain. It is true that PCs are good for isolating the main axes of variability in a dataset. However, it is also clear that there are some solid spatial structures in atmospheric fields that are rooted in basic physics. I am thinking myself in terms of the latitudinal structure of zonally averaged surface temperature. Or I can also consider the vertical temperature profile of the middle troposphere or the stratosphere. These features are missed when the variables are averaged over the full domain before calculating the PCs. I am also aware of the fact that the way these features are represented might be very dependent on the horizontal/vertical resolution of the model being tested. However, checking the zonally averaged surface temperature in midlatitudes ([30,60] degrees in every hemisphere) and tropical temperatures [20S,20N] before calculating the PCs might give three different series which would be representing (even in a crude way) the meridional temperature gradient. Thinking in the same way, getting the temperature at 400 hPa minus the temperature at 700 hPa seems a crude but simple way of evaluating the vertical profile in the troposphere. A similar technique might be used to evaluate the lapse rate around 100 hPa, for instance, adding two different columns to the PCs (without removing the averages that the authors use). Using these diagnostics might improve the sensitivity of the system, can be extended to models of different resolutions and does not involve significant new computations. What do the authors think about this? Can they elaborate on this?*

We thank the referee for this comment and interesting idea for future work. We fully agree that many spatial structures exist in the outputs of the models discussed here and that much of the information about those structures would likely be lost in the spatial averaging step. However, it appears that for these models, it is very difficult to create a non-contrived modification that is entirely not detectable using the global means of the output variables, i.e. the global means contain enough of the signal to capture the change. This was certainly surprising to the authors and initial developers and users of the ECT method (relevant discussion is found on lines 103-110). But across multiple works and a large variety of test scenarios, global means have proven effective.

It is possible that in other domains (as opposed to the atmospheric models considered here), this approach might not be effective. The paper includes the following passage on this point at line 105, ”If a configuration change affects only the spatial distribution of an output field, without modifying the average magnitude of that field, spatial averaging would prevent the test from being effective. This behavior was found in Baker et al. (2016), where spatial averaging ended up erasing the effect of configuration changes in a global ocean model, where there are relatively very few output variables to consider and very different spatial and temporal timescales from atmospheric models”.

While the test can effectively detect changes to these models via a failure, it cannot always help a user determine the cause of said failure. There may be valuable information within the spatial distribution of the outputs that would assist in that task. The referee’s suggestion of additional diagnostic variables could be a relatively easy way to build in some level of spatial information that could assist in the identification of root causes or improve the application of the test for models where spatial information cannot be discarded (like the ocean model above). This is an active area of exploration for our group, and we appreciate the suggestions.

The following sentence was added to section 2 at line 105 for reader clarity. While modifications likely will result in new spatial distributions for model outputs, it appears very difficult to create a non-contrived change that does not *also* impact the spatial means in a detectable way.

2. *Line 293. The authors state that a correlation coefficient of 0.75 is a “limit” that they use to diagnose whether a correlation matrix is/is not rank deficient. However, this would be strongly dependent on the number of degrees of freedom. Why do not they use the spectrum of singular values derived from the singular value decomposition (SVD) of the correlation matrix? If no singular value is negligible, they know for sure they don’t have a problem. Since they apply this a “small” matrix of averaged series, using the SVD shouldn’t be computationally very expensive.*

   We apologize that this section did not read clearly. Similar to the referee’s suggestion, the PyCECT code implementation of the ECT identifies rank deficiency using a QR-factorization approach with a tolerance based on machine epsilon and the degrees of freedom of the system. The 0.75 cutoff is only used for analysis in the paper to see if correlated variables that do NOT result in rank deficiency (as determined by the above method) contribute disproportionately to false positives (it appears they do not). For reader clarity we have modified the language at line 288 as follows. Variables that are almost exactly linearly correlated, and thus result in a rank deficient covariance matrix, are already excluded as part of the PyCECT software (using a QR-factorization approach with a tolerance based on machine epsilon and degrees of freedom) because they have the potential to introduce numerical issues. Further, due to being almost exactly linearly correlated, they will not further aid in the characterization of the model. However, other variables still have a range of correlation intensity. A plot of correlation coefficients for MPAS-A can be seen in Fig. 10. From the 43 MPAS-A variables considered there were 13 variable pairs considered highly correlated. For this analysis we investigate variables having greater than a 0.75 correlation coefficient but not resulting in a rank deficient correlation matrix as identified using the QR-factorization described above.

3. *Line 332. The authors use the (often used) target of 95% of explained variance to identify the number of PCs that must be kept. I don’t think this is critical, but I suggest the authors (for future versions of the software) that determining whether the corresponding EOFs are or are not well determined by the sample might be safer from the point of view of the stability of the next steps. There are alternative methods in the literature for this,*

either based on the errors of the eigenvalues or the congruence coefficients. (Cheng, X., G. Nitsche, and J. M. Wallace, 1995: Robustness of low-frequency circulation patterns derived from EOF and rotated EOF analyses. J. Climate, 8, 1709–1713., North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. Mon. Wea. Rev., 110, 699–706.)

> We thank the reviewer for their thoughtful suggestion and citations. We agree that the use of a 95% cutoff is likely sufficient for now, but alternative methods may be useful in the future, especially if the test methodology is applied to outputs with much larger dimensionality. We have added the following language at line 332. (While not used in this approach, it is worth noting numerous methods of determining the appropriate N_{PC} exist as described in Cheng et al. (1995), Richman and Lamb (1985), and North et al. (1982). While variance explained has proven sufficient for our current method, an alternative method may be explored in future work.)

4. *I think the sentence in lines 174 and 175 must be better worded, I don't find it clear enough.*

> We agree with the referee that this sentence could be more clear. We have adjusted the language at line 174 as follows. The overall ECT results in a failure if N_{pcFails} PC components fail N_{runFails}, or more, of the new test runs.

5. *Line 174. "the the", Line 314. Is a "d" missing after "average"?, Line 454. "with with"*

> We thank the referee for their careful reading. All of these typos have been edited.

Referee #2

1. *In the process of spatially averaging, you lose spatial autocorrelation that may yield important patterns. For example, you lose the ability to characterize if your model is identifying features like the NAO appropriately. It helps with computational tractability of the problem, but is the ability to characterize the underlying differences spatial variability within the two sets of model runs an issue you are able to overlook? I would argue that for meteorological time scales this may not be as important, but in a climate time-scale application, this is certainly a tradeoff and potential limitation of the validation approach.*

We thank the referee for this comment, which is similar to one by Referee #1 and speaks to a key aspect of the ECT. We have copied our response to that question below but would add the following, specific to Referee #2. The ECT approach is not intended nor able to determine whether a model correctly captures some behavior, like the NAO for instance. It can only provide information on whether a new configuration is statistically distinguishable from a previous accepted configuration. Previous works (Baker et al. (2015)) tested spatial averaging using 1-year long runs, but it is an interesting question for future work whether that is the case for even longer runs. When the UF-ECT was compared to the ECT with year long runs results were generally aligned, with the only meaningful differences being changes that were only detectable using short runs (all using spatial averaging.)

Adapted from above: we fully agree that many spatial structures exist in the outputs of the models discussed here and that much of the information about those structures would likely be lost in the spatial averaging step. However, it appears that for these models, it is very difficult to create a non-contrived modification that is entirely not detectable using the global means of the output variables. This was certainly surprising to the authors and initial developers and users of the ECT method (relevant discussion is found on lines 103-110). But across multiple works and a large variety of test scenarios, global means have proven effective.

It is possible that in other domains (as opposed to the atmospheric models considered here), this approach might not be effective. The paper includes the following passage on this point at line 105, ”If a configuration change affects only the spatial distribution of an output field, without modifying the average magnitude of that field, spatial averaging would prevent the test from being effective. This behavior was found in Baker et al. (2016), where spatial averaging ended up erasing the effect of configuration changes in a global ocean model, where there are relatively very few output variables to consider and very different spatial and temporal timescales from atmospheric models”.

While the test can effectively detect changes to these models via a failure, it cannot always help a user determine the cause of said failure. There may be valuable information within the spatial distribution of the outputs that would assist in that task. This is an active area of exploration for our group, and we appreciate the suggestions.

The following sentence was added to section 2 at line 105 for reader clarity. While modifications likely will result in new spatial distributions for model outputs, it appears very difficult to create a non-contrived change that does not *also* impact the spatial means in a detectable way.

2. *When selecting variables, the variables selected are natively generated by the model. If the end user of the model calculates additional derived fields from the model, would you suggest modifying your parameters specific to that problem? It may be good to add a few sentences discussing how your method may change if the user is deriving fields from the native model output, which is often the case with these models.*

   This is a good suggestion to clarify in the paper as, like the referee mentions, users may often derive additional variables. Further, in practice, a model’s default output variables may not align cleanly with native and derived variable sets (and we have found it difficult to distinguish which is which in the investigated models unless one has a truly comprehensive knowledge of the model code.) This is handled by the ECT in two ways. First, we don’t expect cases where an error occurs in a derived field that is not detected in the natively generated output, especially since the ECT procedure captures relationships between variables via the PCA steps. If a user included derived variables that are linearly dependent prior to the ECT approach they will be removed by the PyCECT procedure to avoid creating a rank deficient matrix. In section 3.3 we investigated whether variables that were highly correlated (but not linearly dependent), contributed more to false positives and found they did not.

   For clarification we have added the following text at line 323. It is worth mentioning, that we expect the existing procedure to adequately handle the case were a user has added additional derived variables, or where it is difficult to distinguish which output variables are natively calculated versus derived.

3. *The authors state they retained 43 variables from the MPAS, yet they suggest there are 55 vertical levels in the data. DO the spatial averages include vertical averaging as well? Nesting down 40000+ values to a single spatial mean and then repeating that 55 times to get a single number from well over 2 million points seems like a lot. Are vertical levels of 3-D variables treated separately or together?*

   This is a good clarification. Yes, the spatial averaging included averaging in the vertical dimensions. As the Referee notes, this is a substantial reduction in the data and was surprising to the authors of the original ECT papers as well (as discussed above). The reasoning as to why such a high level of averaging is still effective likely comes down to a few reasons. First, since the goal of the ECT is simply to identify unwanted changes to the model (as opposed to characterizing the model's temporal or spatial behavior) much less data is required. The sensitivity of the approach is discussed in section 5.2. Second, by accurately characterizing the distribution of an ensemble of model outputs the ECT is able to distinguish when very small (in magnitude) changes to global means are still statistically distinct. Third, only a few principal component (PC) dimensions need to fall outside of the accepted distributions to identify a change. Since those PC's also capture relationships between variables this amplifies the sensitivity of the test even when using spatial means.

   For clarification we have amended the following text at line 98. For UF-ECT, spatial variables defined at each grid cell are spatially averaged to one global mean value at each time-slice (this includes averaging across the vertical component of any 3-dimensional variables).

4. *When using the Shapiro-Wilks test for normality purposes with 43 variables, the probability of committing a type 1 error will be quite high (roughly 89%) for each time slice considered. This is more problematic if you are considering vertical levels separately. Did you do any type of correction to the Shapiro-Wilks tests to account for the multiplicity problems?*

   We appreciate the referee's comment on the The Shapiro-Wilk test employed in section 3.2. In our work the Shapiro-Wilk test is only used to determine when the initial perturbations in one field (temperature in our case) have propagated through the model and resulted in normal distributions for other variables. Therefore variable distributions are considered independently and the total number of Shapiro-Wilk failures is examined over time to identify when it has stabilized (as seen in Figure 8). In this way it is not susceptible to multiplicity problems. However, if the number of variables failing the Shapiro-Wilk test did not decrease then stabilize, a user should be concerned that their model is not behaving like those we investigated and perhaps the initial perturbations are not propagating across fields. We also investigated in section 3.3 whether variable distributions need to be strictly normal for the test to be effective.

   For clarification we added the following text at line 258. If the number of variables failing the Shapiro-Wilk test does not decrease then stabilize, a user should be concerned that their model is not behaving like those we investigated and perhaps the initial perturbations are not propagating across fields.

5. *The selection of 95% variance explained for the N_{PC} cutoff seems to increase the risk that you are comparing PCs that are noise instead of signal. Did you experiment with more traditional methods for selecting PCs, such as a scree test, the method of congruence (Richman and Lamb 1985) or a North test (North et al. 1984)? Regardless, how do you account for the risk of noise versus signal when keeping so many PCs? This is even more egregious with the CESM where you're keeping 130 PCs. Almost certainly that amount of PCs includes some noise that may not be useful.*

We appreciate this comment and relevant citations from Referee #2. We did not experiment with alternative methods such as the scree test of method of congruence. In response to a similar comment from Referee #1 above we have added the following language at line 332. (While not used in this approach, it is worth noting numerous methods of determining the appropriate N_{PC} exist as described in Cheng et al. (1995), Richman and Lamb (1985), and North et al. (1982). While variance explained has proven sufficient for our current method, an alternative method may be explored in future work.)

Specific to Referee #2's question of how to account for noise versus signal in the presence many principal components it is also worth noting that the fifth part of the setup framework we propose includes the experimental testing of the false positive rate (see Figures 16 and 20). This provides confidence that the use of a 95% variance explained cutoff did not result in the inclusion of dimensions that were mostly noise. If it did, then it would not be possible to achieve a false positive rate of approximately 0.5%. We agree with Referee #2 that some amount noise is certainly present in later PC's, but it does appear that the 95% variance cutoff is sufficient to avoid unwanted impacts in the false positive rate in the models studied here.

6. *I appreciated the model resolution experiments as this was something I was certainly interested in. I also noted the authors' selection of the same # of PCs for both resolutions of CESM they tested. In their example they suggested this was okay since the differences were minimal in variance explained, but the authors had the luxury of knowing the # of PCs for the coarser resolution run already. In practice, is this a fair comparison, since in a real experiment the # of PCs you retained would be related to something else?*

We thank the referee for this question and agree it is worth clarifying further. In section 4.4 our goal was to explore whether a different model resolution resulted in substantially different UF-ECT parameters. As the referee mentioned, we had access to the variance explained at different numbers of PC's and for different resolutions. In the case of an arbitrary user of this procedure they would still have access to the estimate of variance explained for whatever resolution they used for steps 1 through 5 described on line 209.

As we found that UF-ECT parameters changed only slightly across resolutions we suggested it is likely acceptable to use the same parameters based on our experience with CESM. Users may find this sufficient justification to only use the setup framework for a single resolution and employ the determined parameters across other resolutions. However, if a user is unsure whether the impact of resolution in their model is likely to be similar to CESM, they could also follow the setup framework with a second resolution and compare as we did.

For clarity the following text has been added at line 478. Users of other models may find this sufficient justification to only use the setup framework for a single resolution and employ the determined parameters across other resolutions. However, if a user is unsure whether the impact of resolution in their model is likely to be similar to CESM, they could also follow the setup framework with a second resolution and compare as done here.

7. *When choosing variables to exclude, what criteria are used to determine if variables are "linearly correlated"? Is there a correlation threshold? If so, what threshold and why? Upon further reading I found this definition on line 293. I recommend moving it earlier so the reader has context when the idea of "linearly correlated" is first introduced in the text.*

We appreciate Referee #2 highlighting this. In response to a comment by Referee #1 above we have modified the language at line 288 as follows. Variables that are almost exactly linearly correlated, and thus result in a rank deficient covariance matrix, are already excluded as part of the PyCECT software (using a QR-factorization approach with a tolerance based on machine epsilon and degrees of freedom) because they have the potential to introduce numerical issues. Further, due to being almost exactly linearly correlated, they will not better help us characterize the model. However, other variables still have a range of correlation intensity. A plot of correlation coefficients for MPAS-A can be seen in Fig. 10. From the 43 MPAS-A variables considered there were 13 variable pairs considered highly correlated. For this analysis we investigate variables having greater than a 0.75 correlation coefficient but not resulting in a rank deficient correlation matrix as identified using the QR-factorization described above.

In addition, in response to Referee #2's suggestion to introduce this approach earlier we have amended the following text at line 118. Using constant or exactly linearly correlated variables can introduce numerical issues to the PCA step due to the resulting low rank matrices (these are identified using a QR-factorization approach described in Section 3.3)

8. *Something strange is happening with the text on line 197 with the Molinari citation. I think a comma, parentheses, or something else may be missing. Line 219, The word "don't" should be changed to "do not" to avoid use of contractions in scientific writing. I see the same issue on line 328. It may appear elsewhere, so please check and clean those up. In the figure caption for Fig. 17 I assume you mean p <0.05, not p <0.5.*

   We thank Referee #2 for catching these typos. All have been edited.

9. *While I realize the point is to show the application of the method to any objective model configuration, the reader may benefit from a table listing what the 43 variables are that you consider from the MPAS, at least to reveal the types of things you are considering in the PCA.*

   Thank you, this is a helpful suggestion. We have added an appendix of MPAS-A variables and descriptions as an appendix.

Additional Edits

1. In addition to edits based on referee comments we discovered one error we would like to correct before submission. In section 5.2 and Table 1 we described testing a single-precision configuration of MPAS-A against a double-precision accepted configuration. In the course of follow-up work we discovered that the test runs we had generated were not independent. Because the perturbations generated were on the order of machine epsilon for double precision, when applied to a temperature field represented by single precision their effect was erased. This resulted in the creation of 30 identical runs. Viewed independently, that single-precision run did lie outside the accepted distribution (which is why duplicating it resulted in a failure rate of 100%) but without independent runs we are unable to report a true ECT failure rate. We are also unable to perturb the test scenario using single-precision perturbations without also modifying the accepted ensemble.

   This test represents 1 out of 15 test scenarios explored in Section 5 (actually 39 scenarios if the different scientific parameter values are considered) and therefore we do not believe its removal represents a substantial modification to the manuscript or its conclusions.

   For transparency and clarity we have removed the relevant row of Table 1 and updated the language in Section 5.2 as follows.

   We initially hoped to test the impact of a change from double to single precision. Model developers of both MPAS-A and CESM believed changing precision in such a way should fail. In earlier work (Milroy et al., 2018) a precision-based test scenario passed the UF-ECT for CESM. However, in that case only one subroutine was modified, keeping field representations in double precision, whereas our change for MPAS-A would impact the entire model code.

Unfortunately, due to initial perturbations being on the order of machine epsilon for double precision, they are essentially erased when applied to a temperature field represented by single precision. This prevents the creation of independent test runs. How to effectively test configurations when the accepted ensemble used a higher level of precision than the test configuration remains an open question.

Discussion of the precision test in Section 6.3 was also modified as follows. These results prompt a variety of questions. The exact way in which FMA affects the distribution of chaotic model's outputs is still unknown. Past work (Milroy et al., 2018), has shown good agreement between changes detected using the ultra-fast runs and year-long runs. Further exploration of the effect of non-scientific changes on long runs is warranted.