Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-49-AC2, 2017

Geoscientific Model Development Discussions

Interactive comment on "Nine time steps: ultra-fast statistical consistency testing of the Community Earth System Model (pyCECT v3.0)" *by* Daniel J. Milroy et al.

Daniel J. Milroy et al.

daniel.milroy@colorado.edu

Received and published: 20 July 2017

Thank you for the thorough review and helpful suggestions; we address each item below.

1. [. . .] *I suggest authors use more test cases of porting validation in this paper, such as most of the corresponding test cases used for CAM-ECT (especially the cases with change of processor architectures).*

   We appreciate the suggestion that porting examples may be of particular interest, and have included more porting experiments in the revision.

2. *I agree that Figure 1 is a good motivation for the number of "9", however, I am afraid that the authors' statement that "While nine time steps may not be strictly optimal, we have no reason to believe that more time steps results in a more accurate ensemble consistency determination" (P5L24) is not convincing enough, because Figure 1 cannot represent the testing results of UF-CAM-ECT, which means that this paper does not show that "9" is a "right" number of time steps according to the testing results of UF-CAM-ECT. To further prove "9" is a "right" number or to find a "right" number, I suggest authors evaluate the consistency of the testing results between different numbers of time steps, for example, gradually increasing the number of time steps from a number smaller than 9 (6?) to a number larger than 9 (45?).*

   The major reason we present Figure 1 is to demonstrate sensitive dependence on initial conditions in CAM and to show that choosing a small number of time steps may provide sufficient variability to detect statistical difference resulting from significant changes. In fact, Figure 2 provides the stronger argument for choosing 9 time steps: we do not have any evidence to suggest that an ensemble formed from, e.g., time step 45 will contain variability that improves the UF-CAM-ECT sensitivity or classification accuracy. Since the behavior of most CAM variables (especially those resembling the top two in Figure 2) is consistent through time step 45, the plot indicates that there may be no advantage to using more steps. While some variables do manifest a similar trend to that of the bottom row (increasing variability with time step), integrating that greater variability into an ensemble does not necessarily translate to a more accurate test. In fact, choosing a smaller number of time steps is advantageous from the standpoint of capturing the state of test cases before CESM feedback mechanisms can take effect. We acknowledge that we did not sufficiently explain our choice of time step and have updated the text in section 2.2 in light of this

comment. We appreciate the reviewer highlighting this point.

3. *The cost of UF-CAM-ECT depends on the ensemble size. Given the number of "9" of time steps, the ensemble size is much larger than the ensemble size involved in CAM-ECT. It is unclear about the relationship between the number of time steps and the ensemble size. However, it may be guessed that a larger number of time steps may require a smaller ensemble size. If that is true, it is possible that the cost of UF-CAM-ECT may become smaller with a bigger number of time steps, especially when the cost of model run always is not linear with the number of time steps because the initialization cost of a model run is significant and generally is unscalable with the increment of processor cores. More evidences about this point are welcome.*

   The relationship between model time step number and ensemble size is unlikely to be strictly inverse. For example, in Milroy et al. 2016 we concluded that ensembles of size 300 or 453 are necessary for accurate CAM-ECT (12 month) test results.

   We agree that model initialization and I/O overhead contribute to run time nonlinearity in the number of time steps. In fact, the majority of the cost of a nine time step run is due to initialization and I/O, as evidenced by the average run time of t_1 (96 seconds), t_9 (110 seconds), and t_{45} (118 seconds). (The average at each time step was computed from 10 runs of 900 MPI processes and two OpenMP threads on Yellowstone.) As we noted above, while nine time steps may not be strictly optimal, we do not have evidence that more time steps would result in greater accuracy in ensemble consistency determination, or that forming an ensemble from a later time step would reduce computational cost by reducing the ensemble size.

   Moreover, our primary consideration in this study is to find the smallest CESM time step that permits UF-CAM-ECT to evaluate experimental output in maximal agreement with CAM-ECT, and to detect small-scale changes via instantaneous global mean values before model feedback. An ensemble created at the ninth time step has these desirable properties. Optimizing the cost of ensemble generation and test evaluation is not a main consideration of this study, as UF-CAM-ECT is already an improvement of a factor of 70 over CAM-ECT in this regard. Any incremental improvement in UF-CAM-ECT would be negligible in this context.

4. *It is understandable that UF-CAM-ECT may have different test results from CAM-ECT, for example, the test results shown in Table 3. However, it is still a challengeable situation that UF-CAM-ECT may issue a pass when CAM-ECT issues a fail, for example, the manufactured examples CPL_BUG and CLM_HYDRO_BASEFLOW in Table 3. I do not fully agree the authors' statement that "in practice the two examples we gave in Sect. 5.3.2 were quite contrived as we could not identify more realistic ones" (P13L5), because similar modifications may really happen in model development or research. For example, scientists may only change the land surface data of several grid points when simulating the atmosphere for some scientific researches, or changing a few ocean grid cells into land grid cells in coupled climate model simulations.*

   We did not mean to imply that similar modifications would not be performed for research and development purposes. We reworked the sentence as follows: "While discovering examples where UF-CAM-ECT issues a pass and CAM-ECT issues a fail is conceptually straightforward (e.g. a seasonal or slow-propagating effect), in practice none of the realistic changes suggested by climate scientists and software engineers resulted in a discrepancy between CAM-ECT and

UF-CAM-ECT. We constructed the two examples presented in Sect. 5.3.2 accordingly, which went beyond changes described as realistic by climate scientists and software engineers."

[...] *it may be risky to state that "Therefore in practice, applying CAM-ECT as a second step should only be considered when UF-CAM-ECT issues a fail" (P13L9) and that "The ultra-fast test is cheap and quick, and further testing is not required when a passing result indicating statistical consistency is issued" (P14L1). More evidences or discussions about how to make users safely trust the passes issued by UF-CAM-ECT without the loss of chances for bug detection based on CAM-ECT are welcome.*

We performed numerous experiments attempting to find examples of cases where UF-CAM-ECT issues a Pass, but CAM-ECT issues a Fail. We enlisted the help of climate scientists and elicited the input of CESM software engineers to conceive of examples of such a split decision. The only cases we found were those reported in our paper. We will add emphasis in the manuscript that for important applications a researcher could consider using both tests, but otherwise UF-CAM-ECT appears sufficient.

5. *Considering UF-CAM-ECT is prospective to be general to CAM in various simulations and even general to various models, it will be interesting to know whether UF-CAM-ECT keeps the same results or even the "same" failure rates for the same set of test cases under different simulations with different input data, different parameterization schemes, different time steps, or different resolutions.*

   We agree that studying the failure rates across different scenarios is of interest. We have already successfully used CAM-ECT with both the Finite Volume (FV)

   as well as the default Spectral Element (SE) dynamical cores. We have also used it with fully coupled (active ocean) models. We have not explored multiple resolutions, but plan to do so and have added these suggestions in the section on future work.

6. *It will be welcome if authors list out the failure rates and the failure results at the same time in each table.*

   We appreciate this suggestion and have added failure rates for UF-CAM-ECT to tables 1 and 2.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-49, 2017.