Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-49-RC2, 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.

Anonymous Referee #2

Received and published: 12 June 2017

General Comments:

Software testing, which targets to detect and fix the bugs in the software as early as possible, is an essential step in software development, while the efforts for the software testing of Earth system models (ESMs) seems far from enough, especially when the code lines of ESMs expand rapidly and the climate simulation results still contain obvious uncertainties, which make model developers more difficult to successfully detect software bugs and then fix them. In another aspect, the computer environments (including compilers, processor architectures, etc.) always have to be upgraded with the fast advancement of technologies. Due to the chaotic nature of the climate system

as well as ESMs, bitwise identical outputs may not be obtained after porting an ESM to a new computer environment. Therefore, non-bit-for-bit (BFB) consistency testing are highly desirable for model development.

This paper proposes an ultra-fast consistency testing approach for CESM (pyCECTv3.0), based on the authors' previous works CAM-ECT (Baker et al., 2015; Milroy et al.). This new approach UF-CAM-ECT is easy to understand and is potential to significantly accelerate the software testing, and this paper is well motivated. It of course is prospective to improve the model development. Considering there are some points in the papers are not clear enough currently, I recommend a major revision before the further consideration for publication.

Specific comments:

1. As it is difficult to theoretically prove the effectiveness of UF-CAM-ECT, representative test cases are required for evaluation. I note that this paper follows some test cases used for the CAM-ECT (Baker et al., 2015; Milroy et al.), while many tests cases of porting validation that have been used for evaluating CAM-ECT are not used in this paper. Considering porting validation is an important usage case for UF-CAM-ECT, 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).

2. UF-CAM-ECT in this paper specifically uses the output after 9 time steps of model simulation for testing. The number of "9" is generally motivated from Figure 1 that shows the sensitivity of output fields to the number of time steps. 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?).

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.

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. So 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 discus-

sions 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.

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.

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

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