Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-142-AC2, 2016

Geoscientific Model Development Discussions

Interactive comment on "A new and inexpensive non-bit-for-bit solution reproducibility test based on time step convergence (TSC1.0)" *by* Hui Wan et al.

Hui Wan et al.

hui.wan@pnnl.gov

Received and published: 21 October 2016

We thank Dr. Sacks for his insightful questions. Our responses are detailed below.

Comment: *This is a clever idea, and the paper is very well written. I'd like to be convinced that this technique truly has more power than seemingly simpler techniques. For example, can some of the same experiments be redone with this set of runs?:*
(1) control: unmodified model with 1s time step
(2) baseline for comparison: unmodified model with 1s time step, with a roundoff-level perturbation in the temperature field
(3) test code: some change in the code with 1s time step
Basically, I'd like to be convinced that the "time step convergence" is truly needed here,

and that it truly provides more power than just comparing two versions of the model with a short time step. Does the above, conceptually simpler test give false positives or false negatives in cases where the TSC test gives the correct answer?

Response: This is an excellent question that touches upon some aspects of the old and new test methods that we did not elaborate on in the discussion paper. Essentially, Dr. Sacks asked whether the old PerGro test would become useful again if the model time step was set to 1 s instead of 1800 s. Our answer is "yes, but that revised test could still give false negatives in some circumstances where the TSC method gives the correct answer".

The original PerGro test is no longer useful for the default CAM5 model because even in a trusted computing environment, initial perturbations of $\mathcal{O}(10^{-14})$ K grow so rapidly that the resulting solution differences are often undistinguishable from solution differences caused by unintended code changes or incorrect porting. In our response to referee #2's comments, three reasons are listed as reasons for the rapid growth: (a) long time step, (b) state-dependent randomness in the radiation code, and (c) poorly conditioned code pieces. Reducing model time step addresses issue (a) (see Fig. 1b in our response to referee #2), thus helps to alleviate the perturbation growth; but problems (b) and (c) still exist, and lead to divergence of trusted solutions that can masks subtle but systematic solution changes. Below is an example.

We conducted PerGro test runs using 1 s time step and with radiation called every other time step (so that the time step ratio between radiation and the other parameterizations stay the same as in the default model). We then conducted simulations with the dust emission parameter changed from 0.55 to 0.45 as in the DUST case presented in the discussion paper, also with 1 s time step and with radiation called every other time step. The exercise was repeated using 11 additional sets of initial conditions. As can be seen in the figure below, the temperature RMS differences induced by the parameter change (solid orange lines) stayed substantially below the reference curves (dashed black lines) in the first $\sim$10 time steps, then quickly approached the reference curves

but did not exceed them in any of the ensemble members. We extended the simulations to 300 steps and the results remained the same. Based on the description of the PerGro test at http://www.cesm.ucar.edu/models/cesm1.0/cam/docs/port/pergro-test.html, one would consider the DUST case as a clear “pass", while both our TSC method and the CAM-ECT assigned the case a “fail”.

It is worth noting that the PerGro method perturbs and monitors only the temperature field. Since the impact of dust emission is limited to a rather small number of grid points in very short simulations, it is not surprising that the emission change cannot be detected by PerGro even with 1 s time step. The TSC method makes use of the fact that a change in model time step directly affects all prognostic equations. We monitor multiple state variables, and also calculate the test diagnostics on land and ocean separately, thus achieved higher sensitivity with the TSC method.

Comment: *I'd also like clarification on the following point: On a continuum from non-answer-changing to answer-changing, I see mention of the following types of changes: (1) bit-for-bit identical, (2) answer-changing only at the round-off level, (3) answer-changing only within the limits of numerical accuracy due to the discrete time step size, and (4) climate changing, according to criteria like SIEVE or CAM-ECT. The TSC test distinguishes changes at level 3 or lower from those at level 4. But is there actually a level in between (3) and (4): changes that affect the model evolution in an appreciable way, but are not large enough to cause statistically detectable changes in climate? It seems that many bugs might fall into this intermediate regime – e.g., accidentally flipping the sign on a minor term in an equation. Do the authors feel that there is a set of changes that falls between (3) and (4), and if so, how do they expect these changes to be categorized by the TSC test?*

Response: This additional level between (3) and (4) might exist in principle, in which case the TSC test would assign a “fail” to the results and would not be able to distinguish them from level-(4) differences.

We also would like to point out that level-(3) and level-(4) changes are not strictly defined in a quantitative sense. For example, two simulations representing indistinguishable climate according to SIEVE based on the AMWG diagnostics package might be distinguishable using additional metrics or using CAM-ECT. Similarly, two simulations determined to be consistent using CAM-ECT based on the global and annual averages might turn out distinguishable using grid-point-wise model output and monthly time series. As for level (3), the relatively strong time step sensitivity in CAM5 implies that the numerical accuracies are substantially different when time step is changed, so level (3) is not a fixed criterion either. As can be seen in Fig. 2 of the discussion paper, if we had chosen to conduct a TSC test using a 1800 s time step instead of 2 s, the results from the RH-MIN-HIGH case (which were determined by CAM-ECT as climate-changing) would have been assigned a “pass” by TSC. In the revised manuscript, we point out these ambiguities, and clarify that while answering the “climate-changing or non-climate-changing” question using a specific set of metrics provides *one* assessment of the solution similarity/difference, the TSC method provides a different assessment of the magnitude of solution changes. From a theoretical point of view, the relationship between those two kinds of tests is not entirely clear; practically, because there are flexibilities in the design of the TSC test (e.g., time step size and pass/fail criterion), it should be possible to set up the test so that the outcome closely matches the results from a predefined climate reproducibility test. Evidences are provided in the current manuscript, and future work is planned to further evaluate the strengths and limitations of the TSC method.

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-142, 2016.

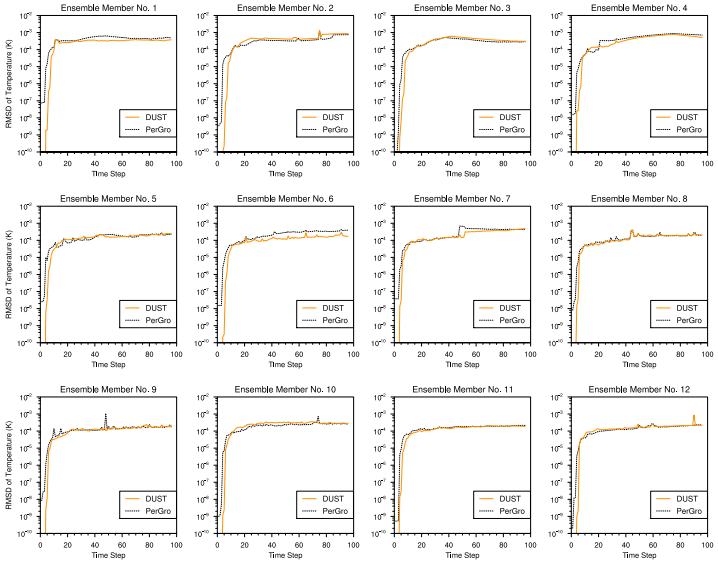

Figure 1: Comparison between the temperature RMS differences caused by initial perturbation of $\mathcal{O}(10^{-14})$ K (dashed black lines, "PerGro") and the differences induced by changing the dust emission parameter from 0.55 to 0.45 (solid colored lines). All physics parameterizations used 1 s time step except for radiation which was calculated every other step. The simulations were conducted on Titan at the Oak Ridge Leadership Computing Facility using the default compiler setups. The 12 ensemble members used initial conditions sampled from different months of a previously conducted multi-year climate simulation with the default CAM5.3 model and the FC5 component set.

Fig. 1. Temperature RMS differences.