Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-142-RC1, 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.

Anonymous Referee #1

Received and published: 28 July 2016

———————— General comments:

Overall the paper was well-written and clear. The TSC test idea is a clever application of the time step convergence work from Wan et al. (JAMES, 2015) and appears useful. Certainly this approach is promising and inexpensive, and the manuscript is a good start. More details on the manuscript are provided below, but my main concerns to address are as follows:

(1) The paper would have been stronger if the test parameters had been fleshed out more thoroughly, particularly the ensemble size, the false positive rate, and number of variables. For example, because this test returns a "fail" if a single variable fails, then a larger subset of variables will increase the possibility of failure by chance, so making

the reader aware of this relationship would be useful.

(2) More details on the scope of the test would be helpful. There are bits in section 2.1 and later in 4, but it would help to better quantify the scope beyond the "equation-solving" part. In particular, the selection of variables would seem to impact the scope. Because of the limited (?) scope, an example of a bug/issue that is not caught would be helpful. And ideally this counter-example would be discussed within a larger discussion of scope as relating to the choice of variables.

(3) The experimental results section would be stronger if the experiments more closely represented the stated scope (see previous comment). Then the reader would gain a better understanding of the tool's utility. The chosen experiments are essentially the same as those in Baker, et al. (GMD, 2015). While it is important to include those, it is not clear that the 10 variables chosen would be sufficient to catch errors in all parts of the code (as stated in section 2.1), so it would be helpful to have an example of an error that is not caught. Also, several times (e.g., section 2.1) "code modifications" are mentioned as an application for this test, but there is not an example supporting this statement (and including such an example seems important).

(4) Regarding the TSC's use of the t-test, please clarify the reason for the directional t-test. In particular, why does the test only check if the mean is larger than zero? (i.e., mu_j > 0) - as opposed to the non-directed alternative hypothesis: m_j != 0. Certainly mu_j can be negative, so is this scenario just not of concern? For example, in figure 3, if delta_RMSD for variable was negative for all 12 members, then TSC would issue a pass. I am not necessarily questioning the efficacy of the test procedure, but I have to wonder if systematically negative results can be problematic as well or even indicative of an issue with the simulation being tested.

———————— Specific comments:

(1) Section 1: line 20: Check the use of "reproducibility" in this context.

(2) Section 1: The first couple paragraphs are quite similar in parts to the text in Baker, et al. (GMD, 2015), including the same references and some of the same phrases, which is a bit awkward.

(3) page 3, line 24: The tool's application to "code modifications" is mentioned here and in section 5, but I don't believe this is being tested in the experiments. It may be of interest to look at CESM code modification experiments in the followup to Baker, et al. (GMD, 2015), which is:

Daniel J. Milroy, Allison H. Baker, Dorit M. Hammerling, John M. Dennis, Sheri A. Mickelson, and Elizabeth R. Jessup, "Towards characterizing the variability of statistically consistent Community Earth System Model simulations." Procedia Computer Science (ICCS 2016), Vol. 80, 2016, pp. 1589-1600. (http://www.sciencedirect.com/science/article/pii/S1877050916309759)

(4) Section 2.1: I would really like to better understand how the selection of the 10 variables affects (or does not affect) the scope.

(5) page 2, line 25: This is not exactly true as CAM-ECT has been used to pinpoint errors in specific code modules (e.g. FMA error on Mira detailed in Milroy et al. 2016).

(6) page 3, lines 27-28: Regarding "...when the accuracy limits related to the algorithmic implementation are taken into account." This doesn't appear to be considered in the rest of the paper.

(7) page 4, line 14: I agree with #5 as a desirable feature, but I don't believe that evidence was given in this manuscript that TSC fulfills #5. Certainly no evidence was given in Baker, et al. (GMD, 2015) that CAM-ECT satisfies #5, though one can imagine the framework could possibly apply. So if the claim is that TSC fulfills this while CAM-ECT does not, it would be stronger to provide specific evidence of such a case for TSC (i.e., an experiment to validate the claim).

(8) Section 2.3: Since the starting conditions for the TSC ensemble are samples from

"a previously conducted long-term simulation", does one need to update this simulation with answer-changing CESM tags, for example? Also does "long-term" mean 1-year or ????? Please give more details on how this part of the process works.

(9) Would the TSC test results be affected if the ensemble was created instead by perturbing initial conditions (since this does not require a previous simulation)?

(10) page 5, last paragraph: Should point out that this test (RH-MIN-HIGH from .8 to .9) is from Baker, et al. (GMD, 2015) for comparison.

(11) page 5, line 34-page 6, line 1: "[...] concept of self-convergence since no structural changes [...] have been introduced into the model." More generally (and relevant to the discussion in Sect 3.2), what if the modified model's 2s timestep behavior is closer to the 1s timestep reference model than to itself for 1s timestep? In other words, what if its convergence behavior to the reference model is different than its self-convergence?

(12) page 6, line 13: The "more substantially" comment is a bit vague. The change is already labeled "climate-changing", which itself seems substantial. Certainly this change is more substantial than, for example, changing the order of operations in the code or something similarly "minor". Clarify?

(13) page 7, first paragraph: Did the authors use the SIEVE method to verify all of the results presented? It is not clear. Also wondering if the example (NU) in Baker, et al. (GMD, 2015) that passed (but that Baker et al. claim should have failed) was independently verified by the authors with SIEVE?

(14) Followup to (12): Recommend that the authors come up with another example of a small scale change that CAM-ECT would not catch because of its use of the global and annual mean (but that TSC would) - other than the NU test from Baker et al. (GMD, 2015). This would probably have to be more subtle than the experiments in Baker, et al. (GMD, 2015). I think this recommendation is particularly pertinent given the list of desired features on page 2 (and that TSC should achieve #6 while CAM-ECT will not).

(15) page 7, line 16: A false negative example would be a great addition and improve the reader's understanding of the tool's scope.

(16) Section 3.2: The splitting into the two domains could be explained more (it is discussed a bit again later in 4.1). It seems a bit arbitrary and suggests that the DUST and CONV-LND failures cannot be detected otherwise. One issue is that by splitting into domains (effectively doubling the number of variables), the false positive rate is being increased. Would be helpful to have more guidance on variable selection and limitations.

(17) page 8, lines 16-28: I'm still struggling a bit with understanding the scope, which is discussed again here in terms of what will and won't be caught (e.g., aerosol concentrations). Please clarify earlier and consider including supporting experimental results.

(18) page 9, line 10: If the implicit assumption that the random variables (mu-sub-j) are Gaussian distributed is violated, will the TSC test results be affected? (And has this been explored? An example could be something like truncation...)

(19) page 9, Step 3 (line 30-> page 6): More clarification is needed here. For the t-test, the choice of .05% is conservative (as acknowledged in text), and it is clear that the specified t-statistic (4.437) is dependent on both the .05% cutoff *and* the sample (ensemble) size (M=12). However, there is a less intuitive dependence on the number of variables that should be pointed out (and discussed). Because the t-test is performed on each variable *individually*, then the number of variables examined certainly affects the overall test failure rates. The conservative choice of .05% may make sense for the 20 variable subset (meaning that a single variable has to fail quite badly to cause a failure of the overall test). However, if one were to use 2 variables (or 100 variables), the .05% may no longer be the best choice. I think this should be addressed given the discussion on page 8 (line 25) that one could choose to include more (and presumably fewer) fields.

(20) page 9, line 9): Please clarify the reason for the directional t-test and consider

updating/clarifying the accompanying discussion on page 9,line 25 -> page 10, line 4.

(21) page 9, line 10: A minor point, but technically one cannot "accept" the null hypothesis. (One can fail to reject the null hypothesis or reject it.)

(22) page 11, line 1: Was the .89 vs. .897 detectable by SIEVE? Also how long of a simulation was run for SIEVE in this case?

(23) page 10: Given the FMA issues found for Mira in Milroy et al. 2016 (and also for BlueWaters), I am questioning the Cori results a bit - also because the results in Table 1 for Cori are not as definitive as for the other machines. Cori uses FMA by default, or was it disabled for these experiments? How long were the simulations examined by SIEVE for Cori?

(24) page 13, line 25: " ...failing the test will very likely mean the climate will be different. Passing the convergence test should hence be considered a necessary condition..." I don't quite agree with this. Many of the parameterizations could be quite different in the short term (because of sensitivity), but the longer term behavior is basically the same. In other words, the weather after 150s may look different (e.g., raining or not), but the annual climate is the same. (This assertion is also made on page 7, lines 15-16)

———————— Technical Corrections

(1) page 2, line 3: Remove the final word "did" from the sentence.

(2) page 2, line 29: The second occurence of "simulation" should be plural.

(3) page 3, line 9: Spell out the number 3 (three).

(4) page 3, lines 23-25: Consider breaking this sentence into smaller parts.

(5) page 5, line 8: "thus saves" should be "thus saving".

(6) page 5, line 18: "dependences" should be "dependencies".

———————— Final thoughts

I like the idea of this work, and I hope that the comments and suggestions provided will be helpful for the revision of the paper. I believe that more flushed out algorithm details, a clarification of scope, and better alignment of the experimental results with the stated features of the test will strengthen the paper and its impact and utility.

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