Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-142-AC3, 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: 22 October 2016

We thank the referee for the careful review. Our responses are detailed below.

Comment: *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.

Response: The intention of this manuscript is to describe a first implementation of the TSC test procedure in the CAM5 model and to demonstrate that it is a practical and useful method for model testing. We acknowledge in the revised manuscript that the test setup can be hardened, and future work is planned to further evaluate the specific choices (e.g., ensemble size and the pass/fail criterion), and to evaluate the strengths and limitations of TSC by comparing it with other methods.

We agree that if the pass/fail criterion stays the same as described in the discussion paper, monitoring a larger set of variables will increase the possibility of failure by chance. We point out in the revised manuscript that $\mathcal{P}_0$ (i.e., the threshold $\mathcal{P}_{\min}$ for failing the test) was empirically chosen by comparing the behavior of the simulations that were expected to pass the test with those expected to fail. If the number of variables is changed, one should re-do the simulations in several trusted computing environments, calculate and plot the time series of $\mathcal{P}_{\min}$, determine the typical values, then adjust the threshold accordingly. There might be other pass/fail criteria that are less sensitive to the choice of variables, and this could be a topic for future work.

Comment: *(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.*

Response: We clarify in the revised manuscript that TSC was designed from the point of view that CAM is a general circulation model that solves a large set of differential, integral, and algebraic equations. The model variables (i.e., arrays in the code) can be

categorized into the following types:

I. Prognostic and diagnostic variables whose equations are coupled to one another, so that any change in variable A will, within one time step or after multiple time steps, affect variable B in the same category. Examples in this category include basic model state variables like temperature, winds, and humidity, as well as quantities calculated as intermediate products in a parameterization, for instance the aerosol water content (which affects radiation and eventually temperature), and the convective available potential energy (which affects the strength of convection hence temperature and humidity).

II. Prognostic variables that are influenced by type-I variables but do not feedback to type I. An example could be passive tracers carried by the model to investigate atmospheric transport characteristics (e.g., Kristiansen et al., 2016).

III. Diagnostic quantities that are calculated to facilitate the evaluation of a simulation, but do not feedback to type I. Examples include the daily maximum 2-m temperature, the ice-to-liquid conversion rate in the cloud microphysics parameterization (which is a quantity calculated merely for output in CAM5.3), and any variable specific to the COSP simulator package (Bodas-Salcedo et al., 2011).

Code pieces in the model can be categorized accordingly.

Our standpoint is that the essential characteristics of the simulated climate are determined and represented by type-I variables, and the TSC test is designed for code pieces in this category. Since all variables in this type are coupled, and since our test method monitors instantaneous and grid-point-wise values before chaos sets in, any significant bug or compiler error (that affects the solution of the *coupled* equation set) should be detectable through the monitoring of a single variable, as long as there is sufficient integration time for the impact to evolve to a discernable magnitude and propagate to that variable. When the simulations are short (for instance on the order

of minutes of model time as in TSC), tracking multiple variables can help increase the sensitivity of the test (decrease the chance of false negative) since discernable solution differences might show up earlier in some variables than in others.

The list of variables monitored by TSC can be extended to type-II variables defined above, if the user wishes to cover the related code pieces in the testing. Diagnostic variables of type I or type II should not be included in the list because the concept of time step convergence does not apply. Consequently, bugs in the implementation of any “diagnostic-only” calculations, e.g., a satellite simulator, would not be detected by TSC. Also, issues with code pieces that are not exercised, for instance the restart capability, would not be caught by the test either.

Comment: *(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).*

Response: Please see our response to the previous comment for a clarification on the scope of our test, and for examples of bugs/issues that would not be caught by TSC.

As for the “code modifications”, two examples from Milroy et al. (2016) that represent code optimization strategies are included in the revised manuscript: “division-to-multiplication" (DM) and “precision" (P).

Comment: *(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:*

mu_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.

Response: The test metric of the TSC method is the model's time stepping error in simulations conducted with 2 s time step compared to trusted reference solutions conducted with 1 s time step. If both the model equations and the discretization methods stay the same, the time stepping error is expected to stay the same. If bugs are introduced, or if the code is not compiled or executed correctly, the resulting numerical integration will not be solving the originally intended equations, thus not converging to the original reference solutions, resulting in larger apparent time stepping errors. In a non-answer-changing case, while $\Delta\mathrm{RMSD}$ can be negative by chance for an ensemble member, it is very unlikely that it will be negative for all members. The only situation we could imagine systematically negative $\Delta\mathrm{RMSD}$ to occur would be the implementation of a new and more accurate set of time stepping algorithms that featured smaller sensitivity to the step size change of 1 s to 2 s, but yet produced very similar solutions at 1 s time step when compared to the original code. Such a case of algorithm update would be considered a substantial code change, so methods like TSC and PerGro would not be the most natural tests to perform since they are designed to assure that the solutions are unchanged. Once the merits of new algorithms have been confirmed and a new default model is established, a new set of reference solutions (with 1 s time step) and trusted solutions (with 2 s time step) should be generated and used for future testing. This is explained in the revised manuscript.

Comment: *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.

Response: The first two paragraphs of the manuscript have been rewritten.

Comment: *(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)

Response: Thanks for the reference. Two examples of code modification from Milroy et al. (2016) that represent code optimization strategies are included in the revised manuscript: "division-to-multiplication" (DM) and "precision" (P).

Comment: *(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.*

Response: Please see our response to general comment #(2) for a categorization of the model variables. The TSC method described in the manuscript is designed to test all code pieces that affect type-I variables, and the 10 variables we chose all belong to that type. Monitoring more (fewer) variables of the same type would not affect the scope of the test but could affect the test's sensitivity for a chosen integration length, i.e., it could decrease (increase) the chance of false negative, since bugs or issues associated with a specific piece of code might take longer time to cause discernable solution differences in one variable than in another. Adding type-II variables, on the other hand, would extend the scope of the TSC test.

Comment: *(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).

Response: The respective sentence in the discussion paper read "Moreover, since each ensemble member is a one-year simulation, it is unlikely that the method can be used to test a small subset of the model components, or a code that is still in debugging stage thus numerically unstable for long simulations (criterion 5)." The statement is now revised. Based on the categorization of model variables discussed earlier in our response to general comment #(2), we expect CAM-ECT to be capable of pinpointing issues associated with variables of type II and type III. The FMA error on Mira as described in Milroy et al. (2016) is an interesting case worth further investigation. To keep the manuscript focused, we do not include any detailed discussions on that topic, but some of our thoughts are included here:

Based on our definition of the type-I variables, we do not expect CAM-ECT to be able to pinpoint issues in code pieces that affect the calculation of type-I variables. The argument is that since all variables in this type are inherently coupled, any substantial change in one equation should have affected all the type-I variables after a year of model integration. In the Milroy et al. (2016) paper, it was reported that six output variables from the CAM model were identified as suspects for further inspection. We contacted the authors and obtained the actual list of those variables. Five out of those were in fact type-III ("diagnostic-only") variables as we suspected, but it was curious that the sixth variable was CLDLIQ, the mass concentration of liquid-phase condensate in stratiform clouds. Given the important role of this prognostic variable in the model, it is counterintuitive to us that values of this variable obtained on Mira were inconsistent with the control ensemble while values of other closely related variables like temperature, humidity, and cloud properties were consistent. Could it be that the inconsistency in CLDLIQ was very minor thus the impacts on other variables were negligible? Would we see more substantial inconsistencies and in more variables if spatial patterns were included in CAME-ECT? The answers to these questions are unknown

at this point. We also learned from Mr. Milroy and Dr. Baker that a number of code lines and local variables in the cloud microphysics parameterization were identified as being affected by FMA. It was again counterintuitive to us that those local variables included the microphysical tendencies of cloud droplet and ice crystal number concentrations, but the corresponding state variables were deemed consistent between the Mira results and those from the trusted computers. To us, this again indicates that the case is worth further investigation in the future.

Comment: *(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.*

Response: The subsection on test scope has been rewritten. What we meant by the sentence cited above has been rephrased: From the point of view that CAM is a general circulation model that solves a large set of differential, integral, and algebraic equations, we consider the results as unchanged if the numerical solutions are found to have the same time stepping error when compared to a predefined set of reference solutions.

Comment: *(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).*

Response: In the earlier study of Wan et al. (JAMES, 2015), in addition to assessing time step convergence in the the full CAM5 model, convergence analysis was also done for configurations that exercised the dynamical core plus only one parameterization or parameterizations group at a time, e.g., deep convection, shallow convection, large-scale condensation, or the stratiform cloud microphysics. This was an attempt to

find out which of those parameterizations led to the convergence rate of 0.4 (instead of 1) in the full model. Simulations were also conducted using the dynamical core plus a very simple saturation adjustment scheme, or with the cloud microphysics parameterization of CAM5 but with the formation and sedimentation of rain and snow turned off (see Figure 3 in Wan et al., 2015, JAMES). Those simulations conducted with a small portion of the CAM5 code were likely to blow up if the integration had proceeded longer than a few hours or days, and certainly would not produce any realistic climate, but they clearly revealed different convergence rates and time step sensitivities associated with different components of the model code. We imagine the same strategy of breaking down the code into small exercisable units and evaluating convergence could be used to pinpoint bugs when, e.g., a code refactoring leads to unexpected failing results from the TSC test. This is why we believe the TSC method fulfills feature #5, and we clarify it in the revised manuscript.

Comment: *(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.*

Response: We clarify in the revised manuscript that the initial conditions for individual ensemble members were samples from the first year (after 6 months of spin-up) of a previously conducted 5-year simulation. In our experience, climate simulations of 1–5 years are frequently carried out during model development or evaluation, making such initial conditions easy to obtain. The basic requirements for the initial conditions for TSC are that (i) they contain reasonably spun-up values for the model state variables (e.g., not all zeros or spatially constant values for the hydrometeors or aerosol concentrations), and (ii) they represent synoptic weather patterns in different seasons. Those initial conditions do *not* need to represent well-balanced states in the quasi-equilibrium phase of a multi-year climate simulation. In fact, the default model time step of 1800 s was used when creating the initial conditions for this study, while the control and test

simulations in TSC used 1 s or 2 s time step, so the model state was certainly not well-balanced during those TSC simulations. We think the same set of initial conditions can be used after answer-changing code tags are established – until a point when the list of prognostic variables in the model becomes substantially different. Then it would be useful to regenerate the initial conditions, and rethink which variables should be monitored by the test.

Comment: *(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)?*

Response: We have not tried this idea yet, but expect that the answer would depend on the magnitude of the initial perturbations. Since our intended simulation length is on the order of minutes to an hour, small perturbations like those used in PerGro and CAM-ECT would not have time to trigger sufficient spread (variability) among the ensemble members. The need for ensemble was demonstrated by Figure 3 in the discussion paper.

Comment: *(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.*

Response: Done.

Comment: *(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?*

Response: Given the complexity of the model and its time stepping algorithms, we would argue it is very unlikely that a modified model's behavior at 2 s time step will be

closer to the reference solution at 1 s of an old model than to the reference solution at 1 s time step of the new model. As mentioned earlier in our response to general comment #(4), the only situation we could imagine to see that kind of results would be the implementation of a new and more accurate set of time stepping algorithms that featured smaller sensitivity to the step size change of 1 s to 2 s, but yet produced very similar solutions at 1 s time step when compared to the original code. Such a case of algorithm update would be considered a substantial code change, so methods like TSC and PerGro would not be the most natural tests to perform since they are designed to assure that the solutions are unchanged.

Comment: *(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?*

Response: Two simulations that are both "climate-changing" can differ from the control simulation by different magnitudes. We revised the wording of the respective sentences as follows:

"If we had introduced larger changes in the model, e.g., by changing cldfrc_rhminh to 0.999 instead of 0.9 from the default value of 0.8, or by replacing a certain parameterization by a different scheme, the impact might be more visible at the default step size. In contrast, if the parameter change were smaller, e.g., from 0.8 to 0.82 instead of 0.9, the red and blue convergence pathways in Fig. 2 might not diverge until a step size on the order of a few seconds."

Comment: *(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?*

Response: We point out in the revised manuscript that SIEVE is not necessarily a

satisfactory procedure or gold standard due to the ambiguity in the criteria for "climate-changing". For example, two simulations judged to be indistinguishable by SIEVE based on the AMWG diagnostics package might be distinguishable using additional metrics or using CAM-ECT. In the revised manuscript, we clarify that we did not independently verified any of the parameter perturbation examples from the Baker et al. (GMD, 2015) paper. Rather, based on the magnitudes of the parameter changes and our understanding of the mechanisms through which those parameters affect the simulated atmospheric motions, we expected all those changes to cause solution differences discernable by TSC.

Comment: *(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).*

Response: Since the test diagnostics of TSC are calculated from instantaneous grid-point-wise model output while CAM-ECT uses global and annual averages, we believe it is reasonable to expect that the former has a larger chance to catch regional differences in the solutions. The NU case has provided evidence to support this reasoning. It is worth noting we also stated in the discussion paper that

"On the other hand, since a large number (120) of model output variables are used in CAM-ECT and the simulations are relatively long (1 year), the chance of missing a climate-changing modification (i.e. getting a false 'pass') is relatively small."

We agree that further examples of small-scale solution changes would be informative, but they would not affect the key messages we are trying to deliver in this manuscript. In a more generally sense, it would be useful to compare TSC with CAM-ECT using

additional (more subtle and challenging) test cases so as to further understand the strengths and limitations of either method. We would be delighted to collaborate with the CAM-ECT developers on that.

Comment: *(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.*

Response: As mentioned earlier in the response to general comment #(2), any bug in "diagnostic-only" parts of the model code, e.g., the calculation of daily maximum 2-m temperature, or the implementation of a satellite simulator, would not be caught by TSC. Another type of false negative is discussed in our response to the next comment.

Comment: *(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.*

Response: Without splitting the domain into land and ocean we would indeed get false negative results in the both cases (DUST and CONV-LND). If we had not chosen to include the aerosol number concentrations in the list of monitored variables, the DUST case would also end up being a false negative. In these cases, the false negative results might be avoidable if the simulations were considerably longer. Limited sensitivity is a price one might have to pay in exchange for the rather low computational cost. We think the test setup described in the manuscript is a practically useful choice if the user is primarily interested in the basic atmospheric state and clouds and aerosols. If the model was going through an active development in, say, the representation of atmospheric chemistry, it would be beneficial to add concentrations of some chemical species to the list of monitored variables. The TSC method is flexible in this regard, although we would like to emphasize again that only prognostic variables of type I and

type II defined in the response to general comment #2 can be used for the test diagnostics. The concept of time step convergence does not apply to variables that are not calculated using an evolution equation.

As for the impact of the number of variables on the false positive rate, we clarify that the $\mathcal{P}_0$ in this manuscript (i.e., the threshold $\mathcal{P}_{\mathsf{min}}$ for failing the test) was empirically chosen by comparing the behavior of the simulations that were expected to pass the test with those expected to fail. If the number of variables is changed, one should redo the simulations in several trusted computing environments, calculate and plot the time series of $\mathcal{P}_{\mathsf{min}}$, determine the typical values, then adjust the threshold accordingly. There might be other pass/fail criteria that are less sensitive to the choice of variables, and this could be a topic for future work.

Comment: *(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.*

Response: Revision is made in line with our responses to general comment #(2) and specific comments #(15)-(16).

Comment: *(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...)*

Response: We have not explored this. The manuscript only describes the first implementation of TSC and provides evidences that this is a useful method. The test setup can be hardened in the future.

Comment: *(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.*

Response: We agree with the referee's comment. As mentioned above, we clarify that the $\mathcal{P}_0$ in this manuscript (i.e., the threshold $\mathcal{P}_{\mathsf{min}}$ for failing the test) was empirically chosen by comparing the behavior of the simulations that were expected to pass the test with those expected to fail. If the number of variables is changed, one should re-do the simulations in several trusted computing environments, calculate and plot the time series of $\mathcal{P}_{\mathsf{min}}$, determine the typical values, then adjust the threshold accordingly. There might be other pass/fail criteria that are less sensitive to the choice of variables, and this could be a topic for future work.

Comment: *(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.*

Response: Done. See also our response to general comment # (4) above.

Comment: *(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.)*

Response: Corrected. We now use "fail to reject".

Comment: *(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?*

Response: 10-year simulations were conducted in both cases and compared to a control simulation using the AMWG diagnostics. The case of 0.897 was indistinguishable from the control by SIEVE using the standard plots, but given the rather direct impact of this parameter on the cloud formation in the model, we thought the difference might be detectable by additional metrics. The expected "fail" was rather an educated guess that was later confirmed by TSC. This is clarified in the revised manuscript.

Comment: *(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?*

Response: As mentioned earlier in the response to specific comment #(5), we think the FMA issue is an interesting one worth further investigation. There is the possibility that the impact of FMA is far below the magnitude of the time stepping error in very short simulations thus not detectable by the TSC setup described in the manuscript. In that case, the use of multiple test methods might help better understand the impact of the FMA issue from different angles. Since the case is not yet well understood, and the Cori example is not essential for demonstrating the basic idea and utility of the TSC method, we do not show the Cori results in the revised manuscript.

A separate comment on the Cori result: in Table 1 the $\mathcal{P}_{\mathsf{min}}$ values were shown only at 5 minutes and 30 minutes after model initialization. While the two numbers from Cori were indeed less definitive than those from the other machines, from the complete time series shown in Figure 6 of the discussion paper, the Cori results seem less suspicious. This made us realize that "pass/fail" criteria based on results at a single time instance are more likely to lead to false positives and negatives. Since the model equations are evolution equations, and the TSC method looks at a time window well within the deterministic forecast limit, it might be beneficial to determine pass or fail using model results from more time steps within a certin time window.

Comment: *(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)*

Response: It sounds like the referee was thinking about chaos and predictability. Since the TSC test only looks at a time window of a few minutes to an hour, We believe the problem should be sufficiently deterministic.

Comment: *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".

Response: All corrected or revised.

Comment: *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.*

We thank the referee for the detailed and very helpful review. The questions and suggestions, together with the comments from the other referee and from Dr. Sacks, prompted us to think deeper about our method. We have made substantial revisions in the manuscript to clarify the purpose and scope of our method as well as our under-

standing of its relationship to other methods. We also added discussions on the details of the test design, including the ensemble size, test diagnostics, and pass/fail criterion, and acknowledge that these can be further evaluated and hardened. We intend to continue this work and obtain more comprehensive understanding of the strengths and limitations of the TSC method.

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