Geosci. Model Dev. Discuss., 8, C1001–C1006, 2015
www.geosci-model-dev-discuss.net/8/C1001/2015/

Geoscientific Model Development Discussions

Interactive comment on "Enhancement for bitwise identical reproducibility of Earth system modeling on the C-Coupler platform" *by* L. Liu et al.

L. Liu et al.

liuli-cess@tsinghua.edu.cn

Received and published: 7 June 2015

We thank the reviewer very much for the comments and suggestions. We'd like to reply them as follows.

1. I believe the authors should add a few sentences about the character of the climate systems, that it's a fairly stable but chaotic time dependent system and how a last bit round-off error grows in the system

Response: Thanks for this suggestion. We will try to give a figure to show the growth of round-off error in a simulation when revising the manuscript.

2. I believe the authors should add a brief discussion of how bitwise reproducibility is tested, who makes that decision, and how that should be determined. In particular, a

comment in the final section that suggests bitwise reproducibility can be determined from visual comparison of a plot should be removed. Also, it probably should be noted somewhere that if a run is not bitwise reproducible, it's fundamentally difficult to determine whether the solution is the same without a long run or a test that is proven to be statistically significant.

Response: We will discuss about the above questions when revising the manuscript. Models always logs global sum of some fields such as global energy for diagnosing the conservation of the simulation. We find that such kind of global sum is very sensitive to round-off error and therefore generally use it to check bitwise reproducibility.

3. I think the authors assume that models are reproducible on the same machine, with the same executable, with the same processor counts, using the same software stack and same build, but that is not actually stated anywhere in the paper. It is an important assumption that NOT all models (or hardware) achieve, but it is almost certainly a requirement that should be noted in the paper somewhere. This is addressed partly in section 4.2.5 but recommend it be discussed earlier in the paper.

Response: We believe that if exactly the same environment can be rebuilt, the results should be reproducible. The proposal of bitwise identical compiler version set and processor version set can make results more easy be reproduced. We will try to discuss it earlier in the revised manuscript.

4. What if a simulation undergoes changes in input files, source code, or input values in the middle of the run. Are these changes recorded automatically with the C-Coupler tool?

Response: On the C-Coupler platform, the simulation environment is recorded when configuring the simulation. Therefore the changes of the simulation environment in the middle of the run that are not invoked by the current simulation will not be recorded.

5. Section 1: page 2408, line 28, point 3. infringement means that copyright is being

violated. I think the authors want to say that code and input data can be downloaded from separate servers to protect intellectual property. That's not entirely true as intellectual property is often contained in source code or input datasets that are used for simulations even if that part of the model is not invoked.

Response: We misunderstood the meaning of infringement. We will correct it when revising the manuscript.

6. Section 3.1: What if the results are not bitwise reproducible which will often be the case? Can a scientist perform new simulations and compare to the original simluations? How should the scientist proceed? Is there a test to verify the simulations are fundamentally "the same" without being bitwise reproducible?

Response: Thanks a lot for this question. I think it is a big scientific question that requires a lot of future efforts from scientists in the whole world. A recent GMDD manuscript entitled "A new ensemble-based consistency test for the Community Earth System Model" focuses on this question. We will also make some efforts in the future for this topic.

7. Section 3.2: No need to define the filename precisely. Just state that the files used to store the setting information should be uniquely tied to the simulation and contain information about when it was created.

Response: We will improve the revised manuscript accordingly.

8. Section 3.3: Do the diagnostics have to be full precision with respect to the model? Is the ascii (base10) representation of the diagnostics adequate to ensure bitwise reproducibility? Are global sums of fields enough to ensure that the entire field is bitwise identical everywhere? How many and which fields need to be diagnosed? Who makes those decisions? How long does a simulation need to be run to ensure the results truly are bitwise reproducible (ie. one timestep, one day, one year)?

Response: We generally use the ascii representation of global sums of one field from

a several-model-day (for example 3 model days) simulation to check bitwise identical reproducibility.

9. Section 3.5: I think point 7 should be removed, but I understand why it's there. It's just not practical to make it part of the defined process.

Response: We will improve the revised manuscript accordingly.

10. Section 4.1.2: How can a user know what the appropriate compiler settings should be so future hardware/software is bitwise reproducible? Is it adequate to test various compiler versions and hardware "at present"? If there is a performance degradation of the model at bitwise reproducible compiler settings, how does a scientist make the tradeoff between model performance and possible (not guaranteed) bitwise reproducibility in the future?

Response: Thanks a lot for these questions. In this manuscript, we define the bitwise identical compiler version set and processor version set to show that different computing environments can be used to reproduce the same simulation results. It is true that we cannot correctly predict the future hardware/software "at present". However, compatibility is always an important principle for the development of hardware/software. We therefore believe that the compiler settings based on existing compiler versions can make bitwise reproduction more easily achieved with future compiler versions. The performance degradation due to bitwise reproducible compiler settings may be not significant because the communication between processes is always a bottleneck for the scalability of models on high-performance computers.

11. Section 4.2.5 (2.): I don't agree with this claim. If different compiler versions and/or different hardware processors produce different results, it is not reasonable to suggest there are bugs in the model code. There are many reasons this could happen including (as stated elsewhere in the paper) compiler optimation/flags, compatibility of different hardware processors with each other, etc. Maybe I'm missing the point and this needs clarification.

Response: We will improve the revised manuscript accordingly. We think that the experience of bitwise identical compiler version set and processor set can be used for testing the source code for the development of a model. For example, given the two sets derived from a version of a model, these two sets can be used to test the further code versions of the model, to provide more chances to detect bugs in the code.

12. Section 4.2.5: I suggest this section be rearranged in the paper. Point 1 should be moved to a point earlier in the paper to introduce some basic starting requirements for bitwise reproducibility. Point 2 should be deleted unless it's further clarified. Point 3 is already made in the introduction and conclusions.

Response: We will improve the revised manuscript accordingly.

13. Section 5: what is piControl? page 2428, line 3, define "short-time" simulations more clearly (one timestep, one hour, one day, one year, one century?). What was your validation criteria in point 4 on page 2428? How many fields, what fields, how where they chosen? It would be nice to see what happens to the performance when the optimization is increased to scientifically acceptable levels but levels that are not bitwise reproducible in different processor versions or compiler versions. Can the authors demonstrate the claim made earlier that the performance difference is small. It would be nice to see a bit more results here, including efforts to reproduce simluations across different processor types and compilers and some suggestions about why this is difficult. For instance, is it possible at the lowest possible optimization with strict IEEE mathematical operations? But I don't expect the authors to address this at this time. This issue is briefly noted in section 4.2.3 point 2 and should be added to section 6.

Response: Thanks a lot for the above questions. We will try our best to answer them when revising the manuscript. Some questions are difficult to answer for us currently. However, we will try to discuss them in Section 6.

14. Section 6: The idea of testing bit reproducibility from figures is introduced in section

6 without any prior discussion, and I believe this idea makes no sense at all. For many fields, you might not be able to distinguish differences in the first or second digit from figures, let alone bit reproducibility. Figures can prove two simulations are not the same, but they will never prove that they are the same at the bitwise level.

Response: As shown in a manuscript (entitled "Importance of Bitwise Identical Reproducibility in Earth System Modeling and Status Report") in the supplement of this manuscript, round-off errors can lead to significant changes to climate simulation results. Given a figure and that we want to reproduce the figure, we may get another significantly different figure after a non-bitwise identical repetition. Here in Section 6 we want to state that simulation results such as data files and figures should be linked to the corresponding simulation setting package. In our model development, it happened that we felt the results shown in a Figure were good and then wanted to reproduce the corresponding results but failed after a lot of tries, because we forgot which simulation as well as the whole simulation environment produced the results.

15. Code Availability: I find it extremely ironic that the sample models used for this study are not available to the public to check bit reproducibility due to permission limitations.

Response: We are very sorry of that. The sample models used in this manuscript are not developed by us and were not publicly open before. We therefore cannot distribute them publicly. However, the code of C-Coupler1 as well as the C-Coupler platform is always publicly open and for this manuscript, we make the code of GAMIL, the atmospheric component of the CSM FGOALS-g2, publicly available.

16. Presentation Issues

Response: Thanks a lot for these corrections. We will improve the manuscript accordingly when revising it.

Interactive comment on Geosci. Model Dev. Discuss., 8, 2403, 2015.