

Interactive comment on “Importance of bitwise identical reproducibility in earth system modeling and status report” by L. Liu et al.

L. Liu et al.

liuli-cess@tsinghua.edu.cn

Received and published: 24 September 2015

We thank Reviewer #1 for the comments. This response include two parts. We'd like to clarify our opinions in part 1 and reply the comments one by one in part 2.

Part 1. The opinions in this manuscript include:

1. “Reproducibility is a fundamental principle of scientific research.” Results of a published paper should be able to be independently reproduced by any fellow scientists in the world.

It should be a well-accepted concept that “reproducibility is a fundamental principle of scientific research”. Many journals pay efforts to aim reproducibility of results in a published paper (for example, <http://www.nature.com/news/journals-unite-for-C2163>

C2163

reproducibility-1.16259). Specifically to Earth system modeling, journal Geoscientific Model Development try to improve reproducibility of results in a published paper. It not only encourages authors to publicly share their code and asks them to state the availability of the code their papers, but also lists “technical aspects of running models such as the reproducibility of results” as its scope for publications.

2. Bitwise identical reproducibility is necessary to Earth system modeling.

We think conclusions of the reproduction of original results can be classified into three levels, i.e., irreproducible, scientifically reproducible and exactly reproducible. It has been found that some simulation results of Earth simulation results can be sensitive to round-off errors. For the simulation results that are sensitive to round-off errors, it is almost impossible to reproduce the results scientifically but not exactly. In other words, the original results are either irreproducible or exactly reproducible. Bitwise identical reproducibility, which guarantees the reproduction of exactly the same results, therefore is necessary to Earth system modeling. If future technical advances make bitwise identical reproduction easily achieved for scientists, they will also make scientific reproduction as well as repeatability easily achieved.

3. The bitwise identical reproducibility of Earth system modeling is currently at a very low level.

Our survey with more than 300 high-impact papers shows that the results in only a small proportion of papers (less than 20 papers) can be repeated or bitwise identically reproduced under the help from the authors. If we talk about independent repetition or reproduction, there will be almost no papers. We guess that such a low proportion is much lower than the goals of journals to enhance the reproducibility of published papers.

4. Suggestions about bitwise identical reproducibility of Earth system modeling.

Our goal in this manuscript is to give some suggestions about bitwise identical repro-

C2164

ducibility of Earth system modeling. We believe that if the original simulation environment is well recorded and can be independently and easily obtained by fellow scientists, bitwise identical reproduction or at least repeatability can be achieved. Some opinions or suggestions in the manuscript may be not right, we hope such a manuscript can stimulate a lot of discussions for correct opinions, suggestions and solutions, and attract a lot of efforts from journals, scientists and engineers to improve reproducibility of future published papers. There are increasing research or engineering works in other fields such as life sciences to improve the reproducibility of published results.

Part 2. Response to each comment. We will modify the manuscript accordingly when revising the manuscript.

1. This seems to me more of an opinion piece than a research article, and I do not feel that the opinions expressed are supported by the results presented. I therefore recommend that it is not accepted for publication in GMD in its current form.

Response: Yes, this manuscript is more of an opinion piece than a research article.

2. The results of (1) demonstrate that bit-level differences can grow into differences in multi-decadal internal. It is well known that bit-level perturbations grow within a few days to synoptic-scale differences (e.g. Rosinski and Williamson 1997 <http://dx.doi.org/10.1137/S1064827594275534>, Goel and Dash 2007 <http://dx.doi.org/10.1016/j.envsoft.2006.06.011>). One would expect that this could then lead to differences in the states of slow climate modes. Indeed, when porting a climate model from one machine to another, one sees similar differences between multidecadal climatologies on the two machines.

Response: Yes, "it has been shown that climate simulation results can be sensitive to round-off errors (Song et al., 2012; Hong et al., 2013)". In this manuscript we just further show some results about this conclusion. Thanks lot for these references. We will cite them.

C2165

3. It is incorrect to assert, as the authors do on p4378, that the control experiment is "correct" and the others are attempts at reproduction. Rather, due to the growth of perturbations, the multiple simulations are all equally valid statistical samples of the model's climate for this experiment.

Response: We feel that the word "correct" introduces some misunderstandings. The sentence "Any simulation of a model can be considered as "correct", while the other simulations can be viewed as an attempt of reproduction." should be modified to "The results of any simulation of a model can be treated as the original results to be reproduced, while each other simulation can be viewed as an attempt of the reproduction." for clarification. Even the original result are not correct or wrong scientifically, it should be considered as "correct" in a reproduction. For example, fellow scientists may find that some results in a published paper are possibly wrong and want to reproduce the corresponding simulation and results to study how the wrong results are produced. In this manuscript, we use standard deviations, a statistical way, to analyze some results.

5. I would also argue that it is incorrect to state, as in section 2.1, that "More and more evidences, including this study, have shown that round-off errors can introduce significant uncertainty to climate simulation results." No references are given for the "more evidences", and I don't agree that this study demonstrates this conclusion. There is already uncertainty in climate simulations arising from e.g. forcing uncertainty or unconstrained modes of internal variability which will differ depending on (among other influences) the initial state of the simulation. I don't think it has been demonstrated here that the growth of round-off errors add additional uncertainty to that. As a thought experiment, if one attempted to quantify the internal variability uncertainty using an ensemble with a spread of initial conditions, or quantify the parametric uncertainty using a perturbed parameter ensemble, would adding additional ensemble members with bit-level perturbations in the initial state increase the ensemble spread? I would expect not.

Response: Yes, it has not been demonstrated that the growth of round-off errors adds

C2166

additional uncertainty. “More and more evidences, including this study, have shown that round-off errors can introduce significant uncertainty to climate simulation results.” should be modified to “More and more evidences, including this study, have shown that climate simulation results can be sensitive to round-off errors.”

6. Making a more general point on this topic, I would argue that reproducibility in climate science would be better served by other centres repeating the same experiments with *different* models (or the same model with a slightly different setup) and determining whether the same *conclusions* can be drawn, so determining the extent to which the results are subject to structural model uncertainty or internal variability. Conclusions from a single-model study which have not been replicated in other models (or have been contradicted by other models) generally carry little weight, so I think the authors are trying to solve a problem which does not exist (or at least of which I have seen little evidence).

Response: We strongly agree with the referee’s point. It should be a fundamental and successful approach to improve Earth system modeling and scientific research. Given the original results from the original simulation environment, we can conclude that an attempt of reproduction is successful if the reproduced results are (scientifically) the same with the original results, no matter whether the simulation environment for the reproduction is the same as or different from the original simulation environment. So we can classify the reproducibility into two types: the reproducibility with the same simulation environment and the reproducibility with a different simulation environment. This manuscript focuses on the first type. We believe that any published results should satisfy the first type of the reproducibility. If the original simulation results cannot be reproduced by fellow scientists under the same simulation environment, the original simulation results may not be trustable.

The first type of the reproducibility can improve the scientific research based on the second type of the reproducibility. For example, given the original findings with model A in a published paper, fellow scientists may want to confirm the original findings with

C2167

another model B. When the new findings with model B is different from the original findings (we think this situation always happens), fellow scientists may want to reproduce the original simulation and original results using model A, for more details and results in the original simulation, to carefully study why there are differences. When the new findings are the same with the original findings, scientists can also carefully study why they are same with the detailed results after the successful reproduction of the original simulation. It is possible that two simulations with different models produce the same findings but different details or different mechanisms.

The problem in this manuscript is that most of published results of Earth system modeling do not achieve the first type of the reproducibility, because fellow scientists cannot know the original simulation environment that generally are not well recorded by original scientists. We believe it is an important problem to solve because not only reproducibility is a fundamental principle of scientific research, but also the worldwide bitwise identical reproducibility will improve the sharing of simulations and simulation results and will provide more chances for scientific research and improve the reliability of scientific results.

7. Regarding the work in (2), I am astonished at the scale of the task which was attempted, and congratulate the authors on the fact that they successfully reproduced any of the experiments. This is the only aspect of the paper which could be described as presenting “novel concepts, ideas, tools, or data”, but I do not feel that the results are “sufficient to support the interpretations and conclusions” (both quotes taken from the GMD review criteria). Out of 14 papers for which sufficient information and data was provided, only 5 were bitwise reproduced, which suggests that a researcher aiming to bitwise reproduce an experiment would have a low probability of success even if the authors’ proposed standard was adopted by all. Personally I am surprised that the success rate was this high and would expect that an average researcher (who in general has access to only a single HPC platform) would have a lower success rate than 35

C2168

Response: This manuscript discusses about worldwide bitwise identical reproducibility that aims independent reproduction of published results by fellow scientists. We believe that if the original simulation environment is well recorded and can be independently and easily obtained by fellow scientists, bitwise identical reproduction or at least repeatability can be achieved. The implications of from the survey can be concluded as: 1) almost no results in published paper can be independently reproduced by fellow scientists; 2) if fellow scientists call for help from the original scientists, only a small proportion of fellow scientists will get effective help from original scientists; 3) if fellow scientists obtain the original simulation, they can reproduce the original results at the bitwise identical level with a high rate. So the survey well supports our claims in the manuscript.

8. The methodological description in this section is missing any information on the range of computing platforms which the authors had at their disposal? Was it a single HPC system, or were they able to choose a system which resembled that used in the original research? Of the 5 successfully reproduced simulations, were any on hardware different from the original experiment? I would be very surprised and intrigued to learn more if this was the case.

Response: Thanks a lot for this comment. We will add the missing information about the computing platforms. Of the 5 successfully reproduced simulations, we use the same CPU family (Intel CPU) but different versions from the original experiment. We failed to bitwise identically reproduce the original results with a processor in a different CPU family.

Interactive comment on Geosci. Model Dev. Discuss., 8, 4375, 2015.