Geosci. Model Dev. Discuss., 8, C2377–C2381, 2015
www.geosci-model-dev-discuss.net/8/C2377/2015/

Geoscientific Model Development Discussions

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: 2 October 2015

We thank Referee #1 a lot again for the further comments on our responses. The comments from the referee will not only help us further improve the manuscript, but also help correct our opinions and give us new insights. We are very glad to discuss more with the referee.

1. The authors repeat the following truism: "Reproducibility is a fundamental principle of scientific research", which means little without defining what is meant by reproducibility. In most scientific fields (e.g. physics, biology, psychology) reproducibility does not mean that the exact data of a study can be replicated to 8 decimal points, but rather that the results paint a similar conclusion. I would expect that the same will be true in any computational science involving floating point calculations (e.g. cosmology,

fluid mechanics). Is climate modelling unusual among the computation sciences in this respect?

Response: For our opinion in this manuscript, reproducibility means that reproduced results paint a similar or even exactly the same conclusion. "Scientifically reproducible" referred in this manuscript means that "results paint a similar conclusion". I think for other computational science, reproducible can also be classified as "scientifically reproducible" and "exactly reproducible". Exact reproduction is necessary for climate modelling because conclusions from climate simulation results can be sensitive to round-off errors. Currently we are not sure of that whether scientific conclusions in another computational science can be sensitive to round-off errors.

2. In point 2 on page C2164, the authors make an argument as to why bit-level reproducibility is essential in climate modelling. The argument follows a chain of reasoning, which rests crucially upon the following assertion: "For the simulation results that are sensitive to round-off errors, it is almost impossible to reproduce the results scientifically but not exactly." It is far from impossible to do this. There are many papers using coupled atmosphere-ocean for which the runs could look very different due to differences in forcing or states of variability, but for which this would make little differences to the conclusion of the paper (e.g. studies of the impact of a model change, or in the mechanisms behind a particular process).

Response: Thanks a lot for this comment. We strongly agree with the referee on this point. Given a number of runs of the same simulation under different computing environments, even when the simulation results are statistically sensitive to round-off errors, it is highly possible that multiple runs of the simulation produce the same conclusion. Therefore we should correct the statement "For the simulation results that are sensitive to round-off errors, it is almost impossible to reproduce the results scientifically but not exactly". The statement "To reproduce the simulation results that are sensitive to round-off errors when the original simulation setting is not completely known, fellow scientists always have to try a number of simulation runs under different

simulation settings, or even have to conclude that the original results are irreproducible after a lot of failed tries." may be much better.

3. On the other hand, if the *conclusions of a study* (not just the simulation data itself) are sensitive to the precise initial state and computational platform on which the experiments are performed, e.g. because the conclusions are only valid for particular phases of a mode of internal variability, then the study should never have been published on the basis of a single experiment. If such a paper is published, and an attempt to repeat the experiment produces very different conclusions, then the reproduction attempt has been a useful exercise in that it has demonstrated that the conclusions of the original paper are unsafe and that more ensemble members or a longer integration is required, or that the result is only valid for an individual model. This is a perfectly reasonable way for our science to progress, and does not require bit-level reproducibility across multiple platforms.

Response: It is possible that a lot of papers with unsafe conclusions have been published. We strongly agree with the referee on the point that "reproduction attempt has been a useful exercise in that it has demonstrated that the conclusions of the original paper are unsafe and that more ensemble members or a longer integration is required, or that the result is only valid for an individual model". Here the biggest challenge is that how fellow scientists to "safely" confirm that the original results in a published paper are unsafe when attempts to repeat the experiment produce very different conclusions. For example, some students in our group tried to reproduce the conclusions in some published papers for their research, following in the experimental setups introduced in the paper. They got significantly different conclusions after a number of tries and then called the authors for the original simulation settings but no reply was received. I told them it was not "safe" to conclude that the results in the paper were unreliable, because they did not know the differences between the original simulation settings and the simulation settings recreated by them, and whether the differences were reasonable to result in significantly different conclusions.

So we think that to "safely" confirm unsafe conclusions of the original paper, the whole original simulation setting should be known. The survey in the manuscript shows that fellow scientists rarely can independently obtain the whole simulation setting for the results published in papers and only the authors of a small proportion of papers are convenient to provide the whole original simulation setting. The worldwide bitwise identical reproducibility in the manuscript aims to improve the independent repetition and independent reproduction of original results in published papers. According to the definition in the manuscript (Section 2.2), it only requires original scientists of published results to ensure the whole simulation setting publicly available, but does not enforce every reproduction by fellow scientists at the bitwise identical level.

In summary, we believe that the public availability of the whole simulation settings in published papers will improve the progress of science and agree with the referee on that the reproduction of a conclusion "does not require bit-level reproducibility across multiple platforms".

4. As a final note, within my own institute, bit-level reproducibility *on a single platform* is an important requirement for model development. We take a great deal of care to ensure that repeated model runs produce bit-identical results, even when changing the number of cores on which the model is run. We find this a useful tool for identifying bugs in new code, and in filling gaps in recent experiments following an archive failure. However, even with this capability and with the "whole simulation setting" still available, we are unable to produce identical results when we upgrade to a new HPC platform. We would if we could because it would save significant effort in port validation, but it is simply not possible. So the chances of another centre bit-reproducing the results of a simulation are small, even with access to the "whole simulation setting". This is also demonstrated by the authors' own results, which show only a âĹij30% success rate for the simulations for which they had the full information required.

Response: We are very glad to discuss about this issue with the referee. We still work on how to make bitwise identical reproduction across different computing platforms. Al-

though it currently may be impossible to achieve bitwise identical results across different processor families and across different compiler families, in the companion GMDD paper ("Enhancement for bitwise identical reproducibility of Earth system modeling on the C-Coupler platform") of this manuscript, we show a preliminary conclusion that bitwise identical results can be achieved across different versions of the same processor family and across different versions of the same compiler family (Section 4.1.2 in the companion paper). Our recently finished work about bitwise identical compiling setups (submitted to GMD several days ago) further shows that a set of bitwise identical compiling setups can be across different compiler versions and different compiler flags. So we believe that model centers have chances to keep bitwise identical results of original simulations on a new HPC platform.

Port validation is necessary when bitwise identical results are not achieved on a new HPC platform. We note that there is a recent paper from NCAR about a new ensemble-based approach for port validation. A hypothesis of port validation is that the original results on the old HPC are correct and a new HPC is not valid if it is failed in the port validation. The challenge to port validation is that this hypothesis may not be true in some cases because there may be bugs or risks in the old HPC. For example, in the work about bitwise identical compiling setups, we propose a new approach based on bitwise identical testing that can detect the compiler bugs triggered in model simulations. There is an example that a compiler bug of the Intel compiler version 13 is triggered when running the ocean model POP2, which may have not been detected and reported before.

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