Geosci. Model Dev. Discuss., 8, C2158–C2160, 2015
www.geosci-model-dev-discuss.net/8/C2158/2015/

Geoscientific Model Development Discussions

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

Anonymous Referee #1

Received and published: 23 September 2015

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.

Additionally, the results themselves are not as surprising as the authors assert. The substantive research reported in this article consists of

1. experiments exploring the internal variability of two CMIP5 models
2. heroic attempts to reproduce the results of papers published by other authors

I will address each in turn.

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 multi-decadal climatologies on the two machines.

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.

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.

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).

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

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.

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