Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-91-RC3, 2019

Geoscientific
Model Development
Discussions
Open Access

Interactive comment on "Replicability of the EC-Earth3 Earth System Model under a change in computing environment" *by* François Massonnet et al.

Anonymous Referee #3

Received and published: 25 July 2019

article

1 Overall review

Climate models have a particular problem related to debugging and testing. The underlying dynamics is chaotic, so sensitive to small (floating point roundoff-level) changes. The community has had to extend classical software engineering practice to cases where testing cannot rely on exact bit-for-bit reproducibility. This paper adds to a growing literature on this issue, and this paper cites most of the key papers from that literature.

The particular example chosen here examines two different releases of a widely used climate model EC-Earth. The two versions were tested on two different supercomputers with different hardware, compiler versions, and optimization levels. Comparisons of output were done across a commonly used set of model metrics from Reichler and Kim. The results showed the existence of a possible "uninitialized variable" bug in one version of the model. In the newer version of the model, there is no conclusive evidence of hardware and software causing a changed climate.

The paper is a minor addition to an existing literature, but is useful for forcefully making the case that hardware and software induced answer changes should be very carefully examined as source of differences between two model runs, and the community should systematically adopt rigorous testing processes.

2 Specific Comments

- The discussion should mention how the results compare to those in prior perturbation studies mentioned in section 2.2 [e.g., Baker et al. (2015)]
- In section 3.1.4, provide a few more details about the H-TESSEL model (e.g., resolution, grid type, number of vertical levels)
- Section 3.2 should include an explanation of why a 20-year period is necessary to detect code errors that may arise later [e.g., more than 1 year as in Baker et al. (2015)] in the coupled climate model simulations, and whether 20 years is also sufficient for testing different configurations (e.g., active biogeochemistry vs none) or grid resolutions. In particular, how is the 20 years reconciled with the Servonnat et al finding cited at the bottom of page 4, that about 70 years are needed to account for low frequency ocean variability?

- The details about the Monte Carlo simulation in the 2nd sentence of the Figure 1 caption should be moved to paragraph 2 in section 3.2.3.
- The authors do not explicitly compare Figs. 2 and 5 in the text. The authors should likewise explicitly compare Figs. 3 and 6 in the text.
- Figure 3 and Figure 6: The text after the first sentence in the titles should be placed in the caption. The color bars need units, and should have the same numerical range to clarify side-by-side comparison.
- page 12, lines 3-6: Lay readers (e.g., model end-users) may wonder why this bug wasn't fixed. Add a sentence similar to the 3rd sentence in section 5 that emphasizes that the testing framework is a diagnostic tool that alerts the user to potential issues in model code, but does not identify or fix specific problems.
- The null hypothesis that is stated on page 13 line 1 should be introduced in section 3 (methods)
- Page 16: The authors should highlight the importance of adopting software development best practices generally, such as compiling and running climate models without optimizations and debugging flags (e.g., -fpe0) in the second bullet point. In the reviewer's experience, the typical goal of climate model end users is to simply get the model to build and run due to time constraints and/or lack of knowledge about model software and build systems. Users will run simulations with full optimizations only, and may not be aware of issues with code until they examine the output.
- In the conclusions, reiterate that this paper only demonstrates that EC-Earth 3.1 is non-reproducible, not that EC-Earth 3.2 is.
- The data and code availability section should state clearly that the model codes themselves are not publicly available, and therefore a reviewer or reader cannot

independently verify these results. At best they could independently test in a different model to which they may have access.

3 Grammar / Style

- Change “hereinafter” to the more common “herein” or “hereafter”
- page 2, line 8: “accuracy” is probably the wrong word, consider changing to “precision or stability”
- Move parenthetical explanation of floating-point math from page 4, lines 8-9 to page 3, line 15, where floating-point math is first mentioned.
- page 7, line 5: “unique and identical” is confusing here, the sentence needs to be rephrased.
- page 10, line 5: “nail down” is used incorrectly; consider changing to “narrow down.”
- page 13, line 9: Change “exist” to “be”