Review of the submitted article « Machine dependance and reproducibility for coupled climate simulations: The HadGEM3-GC3.1 CMIP Preindustrial simulation »

The article by Maria-Vittoria Guarino, Louise C. Sime, David Schroeder, Grenville M. S. Lister and Rosalyn Hatcher is about estimating the potential changes in the climate simulated by a climate model when run on different computers, and the influence of internal variability in this type of study.

Overall, the paper has largely improved compared with the first version. However, I still find the same tendency along the manuscript for over-interpretation of the results, too many suggestions and not enough clear use and interpretation of the actual results, not enough precision and details on the methodology, and even worse, a wrong initial statement as the starting point of section 4.2 (minimum length for a piControl DECK simulation in CMIP6 of is 100 years when it is actually 500 years), that is the base of one of the main conclusions of the paper.

Unfortunately, following those elements, I still not recommend the manuscript for publication in GMD.

You will find below some specific comments supporting my decision.

Page 4, lines 9-10: ‘all’ the CMIP6 experiments are not analyzed against the piControl. Replace ‘all’ with ‘many’.

Page 5, lines 11-13: the sentence “However this method cannot detect code bugs, which may cause a model to behave differently on different machines” is incorrect. First, the method you present is suited to check the resolution of the equations of the model on a period shorter than one day. It could thus detect what you call ‘bugs’. And second: what are you actually calling bugs? In your study you are trying to detect whether the resolution of the equations of the model on two different machines could end up with two different climates. But which one is the right one? In case you find that the simulations differ in some way, is there a way to say that one is correct and the other is not? I suggest to replace this sentence with: “However this method is restricted to time scales shorter than one day. The centennial simulations presented in this paper will help understanding whether or not differences can arise on longer time scales in the HadGEM3-GC3.1 model.”

Page 6, line 8: I still don’t get why you divide your SNR by sqrt(2). I’m not saying that it’s not relevant, I just don’t find any explicit reason, supported by a reference, or a demonstration, to explain this choice in your manuscript. I ask for clarification in the text (lines 14-15 are not enough, need a reference), especially because you say that MO-AR differences are outside the internal variability range for values greater than one (and not 0.9, or 1.1).

Page 6, lines 9-12: please reconsider this paragraph with this proposition: “When SNR<1, MO-AR differences can be interpreted as fluctuations within the estimated range of internal variability. When SNR>1, the MO-AR differences in mean are outside the expected range of internal variability. It means that we either evidenced a true difference in mean, or that the estimated range of variability is underestimated”.

Page 7, lines 13-19: the simple and direct way to explain the behavior of your results is:

- on decadal time scale, the period is too short to adequately sample the longer time scales of the interannual variability; therefore the estimated mean is not stable, and the estimated standard deviation of the simulation is likely underestimated compared with the true standard deviation of the internal variability of the model; it is thus not surprising to have values higher than one when analyzing decadal periods
- on longer time scales, the estimate of the mean and standard deviation converge toward their 'true' values. Accordingly, we see that the differences between and MO and AR become smaller.
- For the 200-year long period, we find no value greater than one. Following this diagnostic, and for the variables we assessed, the results show that there is no significant difference in mean simulated with HadGEM3-GC3.1 on MO and AR

And this is valid only for the mean, and for the variables considered. You can thus reconsider your last sentence (line 18-19) by saying that "Our results show that there is no difference in mean when considering a 200-year long period between AR and MO". Your suggestion is that "the overall physical behavior of the model has not been affected by the porting" is premature regarding your analyses.

Page 8, lines 16-17: I reject your conclusion that "simulations using the HadGEM3-GC3.1 model are reproducible […] long-enough simulation in length is used". You can only conclude from your analyses that the mean of the variables assessed is not different between MO and AR for your piControl simulations. Actually showing that the model is reproducible would require that your diagnostics provide an exhaustive description of the model physical behavior, not only mean, but also variability, teleconnection patterns, trends... You can say that your analyses do not show that the model is not reproducible, and that's already a valuable information, that has not been provided by all the modeling centers (and you should receive credit for this).

Page 8, line 22: I propose "The large differences observed on time-scales shorter than 200-years are a direct consequence of the (potentially underestimated) internal variability of the model, triggered (at least initially) by the machine-dependent processes (compiler […] 3.1 for details)."

Page 8, line 27: this is a major point: because the analyses presented in your manuscript concern piControl runs performed with a fully-coupled GCM (typically one of the DECK experiments, see https://www.geosci-model-dev.net/9/1937/2016/gmd-9-1937-2016.pdf), I assume that your are talking about the minimum length of the piControl run in the DECK (if not, then you really need to add more details to be more specific). The actual minimum length for a piControl run in the CMIP6 DECK is 500 years. Therefore your section starts with a wrong statement, which is a pretty serious mistake.

Indeed, your results suggest that 100 years may not be enough to fully sample the internal variability of HadGEM3-GC3.1. The good news regarding this statement is that the CMIP6 protocol asked for 500 years. The bad news is that several CMIP6 models have much more internal variability than their previous CMIP5 versions, and that 500 years might not be enough. But this is another story.

Also, your conclusions and suggestions have a priori no reason to be applicable to other experiments/MIPs, such as AMIP, historical runs, scenarios, etc… if you want to make on any other given experiment/MIP, be more specific.

Page 8, lines 4-14: I don’t understand where you want to go with the 2/3 power law, although the result is surprisingly consistent among the variables. And the “plateau” you describe is supported by three consecutive points on your plots on figure 10, the last one being slightly higher than expected by the line. I would agree that there is a plateau if it was described by more than one single point being higher than expected. And once more, you conclude that this results “suggests” something. I would recommend using your results to “show” things, and stick to what they actually show.