Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-83-RC1, 2019

Geoscientific Model Development Discussions

Interactive comment on "Machine dependence as a source of uncertainty in climate models: The HadGEM3-GC3.1 CMIP Preindustrial simulation" *by* Maria-Vittoria Guarino et al.

Anonymous Referee #1

Received and published: 12 June 2019

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. The subject is particularly interesting and as mentioned by the authors, probably overlooked.

However, I fundamentally disagree with their approach and find their demonstration and arguments confusing, leading to wrong conclusions. I do not recommend the publication of the article in GMD. You will find below some general comments followed by specific comments explaining why.

General comments

The authors assume that computing the difference between two preindustrial simulations run from the same initial conditions and with the same boundary conditions is a measure of machine dependence (last paragraph of the introduction, page 3, lines 4-5; page 9, lines 14-15; page 10, lines 32-34; page 11, lines 10-11). This is fundamentally wrong: it is primarily a measure of differences due to internal variability. The goal of this study is to prove that the difference between PIMO and PIAR cannot be due to internal variability. In other words, the null hypothesis of the test is : Âń the differences between PIMO and PIAR are due to internal variability Âż. Rejecting this null hypothesis can be a proof of machine dependence.

As illustrated in section 2 (an excellent illustration of the influence of two different computing environments on the trajectory in a Lorenz attractor, that should be put to the credit of the authors), the same chaotic model on two different machines will not follow the same trajectory after a de-correlation time because the differences in the way the operations are treated, the differences in roundings (among others) will act as these were infinitesimal perturbations. The model trajectories will thus diverge, and this is expected. But we have no mean to say whether this is due to the computer, or to the behavior of the chaotic equations that would behave the same way on one single computer with added random perturbations.

One way to show that the model is different when run on two different machines is to compare the statistical properties of both attractors, for instance mean and variance, and demonstrate that the differences are not due to random processes (internal variability) or a lack of sampling. This last point is addressed in the paper but with the wrong approach. The sentence "We will focus on estimating how long constant-forcing climate simulations should be for machine dependence uncertainty to become negligible" (page 3, lines 4-5) illustrates that the authors didn't understand what estimating machine dependence is about. Increasing the length of the time series to estimate the mean and variance will allow stabilizing these estimated statistics (i.e. reducing the uncertainty on the estimates). The presence of low-frequency components (mainly the ocean), depending on the climate model considered (they don't all have the same internal variability) implies using multi-centuries time series, which is what the authors do; but the only thing that should decrease with the increasing length of the time series is the uncertainty on the estimates of the statistical properties of the climate of the model. The machine dependence should actually become more and more significant (come out of the noise due to internal variability) with the increasing length of the time series considered, if it truly had an influence. As well, the sentence page 11 lines 10-11 illustrates the confusion between internal variability and machine dependence.

The whole section on the physical implications is out of subject. You are speculating on differences due to internal variability.

Eventually, the authors make suggestions about CMIP6 that are well beyond the reach of their results (page 11, lines 18-23). The analyses of the paper only illustrate that internal variability can cause differences between two climatologies. CMIP6 encourages the modeling groups to provide ensembles with as many members as possible to study internal variability and intercompare the models, and we already see modeling groups providing tens of members of in the CMIP6 DECK historical experiment.

Machine dependence is an important subject in climate modeling and I agree that we should keep an eye on it routinely. However, from a CMIP multi-model comparison point of view, these potential differences are much smaller than the inter-model differences: to my knowledge, a true machine dependence with a CMIP climate model is still to be proven (potentially because they are very small). I invite you to read the publication by Milroy et al (2018) (https://www.geosci-model-dev.net/11/697/2018/) to see an interesting approach to compare model results on different computers. If you read French (I assume that the second co-author does, apologies if you don't), you can also have a look at this technical note from the Laboratoire de Météorologie Dynamique: http://cmc.ipsl.fr/images/publications/technical_notes/jerome_LMDZinfo9.pdf

Specific comments

Page 2, lines 25-27: machine dependance implies that the climate of a given model run on computer A is different from the climate of the model run on computer B. Therefore, machine dependance is a source of uncertainty for the climate models (if we can show that it is significant). Starting from here, reducing this uncertainty would imply showing that the climate simulated on computer A is more relevant than the climate simulated on computer B. Answering this question totally depends on the scientific question you ask. Selecting climate models, or weighting the projections according to selected criteria (like the so-called emergent constraints) is a research subject in itself, and until now the different attempts have not been able to drastically reduce the model uncertainty on future projections. Additionally, the differences between the climate simulated by the CMIP(5,6) models are easily shown by intercomparison results (see Chapter 9 IPCC-AR5) when it is pretty tricky to evidence a machine dependence (potentially much smaller than differences between the models). Therefore, I agree that the machine dependence uncertainty could be estimated someway and taken into account, but I think it is wrong to say that it could be removed by running all the models on the same machine. If we can't select a model today on climate-based criteria and comparisons with observations, I don't see any chance to select one computer. Running all the models on the same computer is not reducing or removing machine dependence: it's ignoring it (what we do today), which is fundamentally different.

Page 6, lines 13-20: I got really confused by this paragraph. You say Âń [. . .] using a chronological order in the strictest sense is meaningless because every 10 years segment is equally representative of the pre-industrial climate variability Âż. And the last two sentences of the paragraph (lines 19-20) contradict this statement. Starting from the same initial conditions, PIMO and PIAR take different trajectories after a couple of time steps, when the chaotic nature of the model takes over. The atmosphere is the first component to diverge (less than one day, much less than the 1 to 2 years mentioned page 8 line 16), and then the slow components of the climate model (namely the ocean and the soil) will keep similar trajectories for as long as their correlation time (depending notably on the geographical region considered) and then will diverge. Studies of the potential decadal predictability of the climate system in a so-called perfect model framework have shown that this correlation time is around a decade (model dependent). After around one decade, the memory of the initial conditions will be lost. There is thus no justification to compare the same dates. Therefore, in absence of more precise explanation, it is not Âń helpful in investigating the impact of machine dependance on model results Âż. I would add that this kind of vague formulation doesn't have it's place in a scientific paper.

Section 3.2: the SNR measure you use to estimate if the mean of PIMO is different from the mean of PIAR should come with a test to determine more objectively when the difference becomes significant. For instance, if you compute your SNR on n years, you could sample 100 couples of n random years (bootstrap) in the same simulation to check the distribution of the SNR when computed between two (random) periods of the same simulation. This would give you an estimate of the influence of internal variability on your SNR (but would also need a longer simulation for this). You can then compare the SNR computed between PIMO and PIAR with the distribution of the SNR within the same simulation and estimate the probability to have the same value between two periods of the same simulation. An alternative would be the use of the Student t-test on the difference of mean, and the Fisher F-test on the difference of variance. The paragraph page 7 lines 8-14 is totally confusing for me. Extract only the information that brings concrete elements to the debate, and remove the rest. Last point: I don't understand where the sqrt(2) comes from in the SNR formula... I would need a reference or an explanation for that.

Section 4.1: the presentation of your results is an illustration of the use of the SNR without knowing when it actually shows a true difference: you say that there are values of the SNR close to 1. Fair enough, as long as those differences are lower than one we should not care about them because they do not show any significant difference (according to your definition of the SNR). And then you talk about SNR values of 0.9, 0.7, that are supposed to be close to 1. How close from one are those results? From

which threshold are we supposed to take these results as proof of a true difference due to machine dependence? Following your definition of the SNR, you should not even mention it and just conclude that the difference in mean between PIMO and PIAR for the concerned variables is not significant. That's it. And not care about the physical explanations of this result.

Page 8, line 30 : you conclude much more than your results show ! Your results show that there is no difference only for the diagnostics that you've done, and those diagnostics have limited implications. You can only conclude that Âń according to our analyses of the SNR, the mean climate for the variables assessed here is not different between PIMO and PIAR. The porting did not affect the climatology of the models for those variables."

Page 8 line 31 to page 9 line 4: I'm afraid you came to this conclusion because you don't understand the problem, or try to sell your results way beyond what they show. Same for last paragraph of section 4.2 : you may just have a SNR larger than one (how larger? Is it significant?) because of the low-frequency internal variability in your model. This means that you need longer simulations to properly assess this question.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-83, 2019.