gmd-2018-52-RC3

Interactive comment on "Requirements for a global data infrastructure in support of CMIP6" by Venkatramani Balaji et al. Anonymous Referee #3 Received and published: 8 May 2018

The paper describes the challenges for the global data infrastructure needed to support the ongoing efforts of the climate modelling community that are organised in the CMIP enterprise. The material presented is of great importance and should be published. However, its presentation does not meet the requirements of a journal article and requires major revision. There are two major issues that need to be addressed as the paper gets rewritten:

RC3-1 1. The paper is clearly the result of a lot of work within the WGCM Infrastructure Panel. Acknowledging this is important, but the paper completely goes overboard and as a result reads like a report to some steering committee, rather than a journal article. I counted no less than 46 (and I am sure I missed some) occurrences of statements like "The WIP recommends. . .," "The WIP did. . ." or "Based on what the WIP thinks . . .". This is simply not the style of a paper. I recommend removing all these references and telling us what the authors of this paper think. I realize they are the WIP, but the reader does not need to be told this every other paragraph. I suggest putting a clear statement that the suggestions of this paper are the result of deliberations by what is likely a temporary body in the long run, the WIP, and then present what are hopefully not temporary conclusions for the infrastructure needs. I also suggest avoiding repeated statements that more detail can be found in WIP reports. This can be said once and the reports listed in the Appendix, as is the case.

We thank the reviewer for their thorough and candid review. This stylistic recommendation has been followed throughout in the revision of the text (too many instances to call out here...) and we believe has greatly enhanced its readability.

RC3-2 2. Perhaps the more important question I struggled with is who the intended audience for this paper is, which will define its purpose and then structure. If it is scientists and users, the paper needs to significantly cut down on jargon (see minor comments below). If it is infrastructure communities outside climate, then this should be written as an example for what challenges the climate community is facing and what it is doing about it, so that others can learn from it. Or is it the modelling centres to instruct them on new procedures and tools? In that case, a paper is unlikely needed as they can be sent an email with the detailed position papers! At the moment, neither community will benefit from this paper as it isn't clear what it is trying to achieve. I realize that this is harder to solve than issue 1, but it is important to know this before rewriting the paper begins. Once it is clear, the goal should be stated in the introduction.

This is a good point raised by RC1 as well, and we have provided some context, see page 4, line 14.. The text and conclusions have been modified as well to make clear the audience and intent. A new section on Historical Context has been added, Section 2.1.

More minor but often typical issues in chronological ordere

RC3-3 Page 2 Line 17 – The statement that by the FAR of the IPCC modelling inter comparisons were formalised is untrue. The first formal model inter comparison was AMIP and was reported in 1992. (Gates, W. L., 1992: AMIP: The Atmospheric Model Intercomparison Project. Bull. Amer. Meteor. Soc, 1962–1970, doi:10.1175/1520-047773.12.1962.) CMIP started only after that. Please correct this

Agreed, reference added, see page 2, line 19..

RC3-4 Page 2, Line 18 – Please cite the appropriate paper when referring to the DECK

Citation added, see page 2, line 21.

RC3-5 Page 4, Figure 1 – Most dots on this figure are in the city the node is located. The one in Australia is in the middle of the desert. Canberra is not. Please correct.

See reply to RC2-12.

RC3-6 Page 8, Line 17 – What is "The data request". This needs an introductory sentence as only people in the know will know. Same line: What is the "DREQ" tool. This is an example for the frequent jargon with no explanation. Please be more careful as not every reader will already know these acronyms.

Good point, and we have added context, see page 9, line 17.. Some unneeded jargon removed.

RC3-7 Page 8, Line 28 – The sentence about the database allowing MIPs to do things is another example for jargon. It means nothing to someone who doesn't already know all this. Please explain it better. For instance, highlight that different MIPs will request different variables, but some will be common. You can't assume the reader to be a CMIP expert and if you do, why write this paper?

Section 3.1 has now been rewritten at what we hope is an appropriate level of detail and context.

RC3-8 Page 9, lines 1-3 – This list is very confusing and requires more context.

Addressed in the revised Section 3.1, bulleted lists removed.

RC3-9 Page 9, lines 16-20: A single paragraph does not deserve it's own subsection. Please correct. The last sentence is another example for a sentence from a report that makes little to no sense to an independent reader. Please remove those as you rewrite the paper.

While this section is indeed quite short, the input4MIPs and obs4MIPs efforts, and their coherence with the overall data design, is an important element and we believe this point will be lost if it is buried in a paragraph somewhere. This point is highlighted and the language on versioning clarified, see page 10, line 19.

RC3-10 Page 10, Line 6 – The statement on increasing data volumes overstates the case if its is not put into context. The Large Hadron Collider produces vastly larger data volumes than any set of climate models ever will! It is important to clarify that the challenge is that the data is both produced and used in a distributed network. If one place with all the resources

needed ran all the CMIP runs from all the models, archiving them would be simple! Distributing them might still be challenign though!

Context added, see page 11, line 5.

RC3-11 Page 10, lines 28-29 – What do you mean with "appear to have grown". Has it or not?

see page 12, line 8.

RC3-12 Page 11, line 11 – What is the "CMIP6 Output grid guidance document"? If you use it, you need to provide a reference/link to it.

References visible now, see answer to RC1-15.

RC3-13 Page 11, line 16-18 – To an outsider to the climate community this appears insane! There is only one real calendar and it has to do with the Earth going around the sun in a certain unit of time. It would be worth commenting on the future of this, as it implies a "laziness" in the climate community to do something simple (I understand it is not that simple).

Good point :-). An explanation of the calendar issue for "outsiders" has been added, see page 13, line 1..

RC3-14 Page 11, line 23-24 – Again, to an outsider this sounds strange. How can infrastructure that does relatively straightforward analysis be overburdened? Isn't that because the way this is funded is inadequate. If you agree, isn't important to point this out in this paper about the future?

Regridding of data is burdensome for many reasons: we have more explicitly pointed out this out now earlier in this section, see page 12, line 20.. We have added a discussion in the Conclusion section about the funding constraints.

RC3-15 Page 11, line 25 – By now I had now idea that there were two issues. Please remove this first and second bit. The first issue was several already!

The numbering has been removed, see page 12, line 13..

RC3-16 Page 12, line 5 – If the results are public, please cite where and how the reader can access them.

References visible now, see answer to RC1-15.

RC3-17 Page 12, line 14 – Where is the WIP website? Please add a link.

References visible now, see answer to RC1-15.

RC3-18 Page 12, line 29-30 – Jargon. We don't know what a Tier1 node is, let alone that it has a manager. Explain or remove!

see page 14, line 22.

RC3-19 Page 14, line 2 – PCMDI website – please cite properly by adding the link.

References visible now, see answer to RC1-15.

RC3-20 Page 14, line 20 – O(10^6), 10^6 what? Add units.

It is a dataset count as stated, no units.

RC3-21 Page 20, line 26 – A good example of overdoing the WIP(ping). The reader does not care where the replication strategy is covered. They want to know what there authors of this paper have to say about replication.

see page 23, line 5.. In general there is much less WIPping in the new draft, see answer to RC3-1.

RC3-22 Page 21, line 21 – Jargon. What is the CDNOT group? Please explain.

The CDNOT was introduced above, see page 6, line 14..

RC3-23 Page 23, line 21 – There seems to be only one subsection, so why have it? Please remove the heading.

The Errata section is related to versioning but an important independent piece, deserving of its own section, in our estimation.

RC3-24 Section 8 – I was disappointed by this section, as I was expecting a summary of the main challenges and recommendations for solutions. I feel some of the challenges need to be spelled out here. For instance, funding for the activities described here is pretty ad-hoc. This is disturbing given the attention the world pays to the data sets in question. Is there something here to discuss? Can the world continue to scramble its way through this? Thoughts? Other big issues: Doing more and more in CMIP (more models, more experiments, more users) cannot be sustained unless investment into the enterprise gets better coordinated – any role for international organisations to help with this? Many of the issues are discussed are the result of accepting the status quo in distributed climate modelling. Should we? Are there more sensible alternatives?

Section 8 has been considerably rewritten. In it we have mentioned that prior panels including at the US National Academies level, have indeed made this case, but so far to no avail. We have explained how the new dataset-centric design is indeed intended to reduce systemic risk due to infrastructure failure, and allows for a scalable system that is sized at a level appropriate to available resources.

RC3-25 Appendix A – Might be nice to add links to each of these reports (or the main page where they can all be found).

References visible now, see answer to RC1-15.