gmd-2018-52-RC1

Interactive comment on “Requirements for a global data infrastructure in support of CMIP6” by Venkatramani Balaji et al.

Anonymous Referee #1 Received and published: 23 April 2018

Overview

RC1-Overview-1 This paper reviews the infrastructure requirements needed to make CMIP6 successful. There are some attempts at charting a path towards the future. Overall, in spite of my numerous specific comments below, the paper is well presented with a few notable exceptions. My biggest complaint is that after reading the paper, I am not sure who the target audience is for this paper. This makes my job as a reviewer much harder, since I am guessing at the answer to that question. I have assumed that the audience are those who want to know something about how the networking/software part of CMIP works. This includes some of the modelers and folks in the large climate modeling institutions and a subset of the more comp-sci oriented users of the CMIP data. If other audiences are in view then my review would be very different. This paper is fairly technical. My second big picture issue is that references are needed in many, many places to either point the reader to supporting documentation or to find web sites that explain in more detail what the functions are of the various groups/position papers mentioned in the paper. Finally, references are also needed to support the statements made in the paper. My specific comments below highlights many of the missing references.

We thank the reviewer for a thorough and knowledgeable reading of the paper. In the revised text, we have addressed explicitly the question of intended audience, see page 4, line 14. Re the point about references, see below the answer to RC1-15. Several additional citations have been added as well.

RC1-Overview-2 Lastly, Section 3.4 needs rewritten. It is very confusing. There are lots of recommendations. In places, the language reads like these are a requirement. In other places, the prose basically say that the recommendations can be ignored. There needs to be some priority applied to the discussion. The readers need to know at the beginning of the section what is coming – requirements, recommendations, best practices or what. Each item discussed needs to be clearly defined in one of the bins – requirements, recommendations, etc. Some parts may be able to be deleted.

The distinction between findings, requirements, and recommendations is made clear now in the introduction, see page 4, line 9.. The section has been considerably edited in response to this comment, and comments below, RC1-20–24.

Specific Comments

RC1-1 1. Page 1, line 11 – purpose of assigning credit – This seems awkward/backwards to me. The tracking is so that the credit is clearly assigned, not the reverse.

Agreed, awkward wording removed, see page 1, line 11.

RC1-2 2. Page 2, line 6-8 – A references is needed for this statement.

Agreed, added, see page 2, line 7.

RC1-3 3. Page 2, line 11 – capable – Wrong word. “Available” is a better word. There were other climate models available around in the world at that time.

Agreed, see page 2, line 12.

RC1-4 4. Page 2, line 15 – Add “group” after Manabe.

Agreed, see page 2, line 16.

RC1-5 5. Page 2, line 16 – Add “group” after Hansen.

Agreed, see page 2, line 17.

RC1-6 6. Page 2 Line 17 – 24 – The role of AMIP is missing here in the formation of CMIP. I agree that the IPCC also played a role, but Larry Gates and AMIP was a necessary step to have CMIP formed.

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

RC1-7 7. Page 2, line 23 – I believe there are now 23 MIPs.

Agreed. We note 2 new MIPs have been added since the first draft of this paper, as well as the canonical citation for CMIP6, Eyring et al. (2016a), which is used in the text. The new wording reflects this evolution, see page 2, line 27.

RC1-8 8. Page 3, line 10-17 – References for CMIP3 and 5 are missing.

We believe this is covered by earlier references to Eyring et al. (2016a).

RC1-9 9. Page 5, line 4 – Reference needed for IPCC.

Added at first reference to IPCC, see page 2, line 20.

RC1-10 10. Page 7, line 2 – consumers – Is “society” a better word choice here?

We believe “value to society” of individual datasets is hard to assess, but value to actual data users/consumers – for example by citation counts – is a measurable quantity.

RC1-11 11. Page 7, line 8 – Designing – I think the CMIP Panel understands the cost of participating in CMIP since it is mainly made up of modelers. It could be argued that some of the new MIP chairs in CMIP6 do not understand. Certainly, most users do not understand. Reword.

Reworded, see page 8, line 3.

RC1-12 12. Page 7, line 9 – Add “data archived in” before CMIP experiments.

Reworded, see page 8, line 4.

RC1-13 13. Page 7, lines 7-10 – This section is vague. Expand and define exactly what is in view here. I assume it includes model development, cpu and storage costs, people time and etc. What is in view? Exactly what costs are in view?

Some explanatory text added, see page 8, line 8.

RC1-14 14. Page 7, line 19 – machine readable experiment design – This needs to be explained here. Page 8, line 14 has a similar problem. It needs noted that this is a goal of this effort.

Some explanatory text added, see page 8, line 18.

RC1-15 15. Page 7, line 29 – A reference and location is needed for the fact sheet.

In the first draft we used embedded URLs, which were not visible by editorial decision on coloured text. All URLs have now been made visible as footnotes, including this one, see page 8, line 30.

RC1-16 16. Page 8, line 5 – Where are these position papers found??? Are they peer reviewed, citations?

As noted, these are listed in Appendix A as noted. The citations are there, but as above the embedded URLs were invisible. They are now visible in footnotes. The position papers are not themselves peer-reviewed, though publicly available for comment: this paper is in fact their peer review.

RC1-17 17. Page 8, line 13 – machine readable – This needs defined. Anything stored in a computer is machine readable. . .by definition. More is needed.

It is explained just below, as being encoded in structured text documents in XML or JSON format for example.

RC1-18 18. Page 10, line 19 – smaller – I think "larger" is correct. . .nearer to 1. The exponent is larger.

Explanatory text added clarifying why it is in fact smaller, not larger, see page 11, line 22.

RC1-19 19. Page 10, line 24 – Add "the first part of complexity" somewhere near here. The second paragraph starts with the "second component of complex" which is confusing given the prose in the first paragraph.

Fixed, see page 11, line 29.

RC1-20 20. Page 11, line 3 – WIP has recommended – This seems in conflict with line 11 and page 12, line 32. As I note in my general comments section, this section is not well written or thought out. What message do the authors want to convey to the readers? Rewrite.

We have considerably rewritten Section 3.4 for clarity, following this reviewer's and others' recommendation.

RC1-21 21. Page 11, lines 4-24 – Regridding – I understand the Griffies papers have a long discussion of the advantages and disadvantages of regridding, but a summary of those papers need to be presented here. The whole discussion of the disadvantages of regridding is missing here.

Discussion added, see page 12, line 20..

RC1-22 22. Page 11, lines 4-24 – Common grid – So what are the authors recommendations for a common grid or regridding? If there are none, then delete this discussion to just a summary of the Griffies papers.

The distinction between findings, requirements, and recommendations was made explicit in the introduction, see page 4, line 9.. Furthermore, we have made explicit that where there is no consensus, we can but present arguments for and against, as this reviewer has requested. We have duly represented here the debate around regridding, calendars, and data deflation, and noted the lack of consensus. The debate needs to continue, in the literature and in other forums of experimental design, until a compromise is achieved. We have also described, in Section 5.2, a tracking mechanism to provide data for this debate, in terms of what user preferences are with respect to these issues.

RC1-23 23. Page 11, lines 32-33 – Again, what is the recommendation? If none, what is the justification for keeping the text?

See above.

RC1-24 24. Page 12, lines 4-10 – What is the recommendation? If any, it needs highlighted. Has the WIP surveyed CMIP users in regard to these recommendations? I am worried that many users will not be able to handle compressed files or shuffled data files.

See answer above. There has been no explicit survey of users in this regard. Shuffling and reinflation are automatic and transparent to the user if using netCDF4 libraries.

RC1-25 25. Page 12, line 8 – coupled model – Define. There are many types coupled models in climate. I assume AOGCM and ESMs are in view.

Fixed, see page 13, line 29.

RC1-26 26. Page 12, line 15 – I do not see what the advantages are of a modeling center having this tool. Please explain. The center should know its model's grid and variables to be archived. . ..

Explanatory text added, see page 14, line 15.

RC1-27 27. Page 12, line 18 – Add "compressed" before "data volume".

see page 14, line 8.

RC1-28 28. Page 12, line 20 – Add "current CMIP 3 and 5" before archive size.

see page 14, line 8.

RC1-29 29. Page 12, line 21 – 25 – The sentences that start with "The more dramatic . . .." And end with "in years simulated" seems out of place and should be moved much earlier.

Agreed, some lines have been a few paragraphs above, see page 12, line 2. and see page 14, line 10..

RC1-30 30. Page 12, lines 26-27 – an attempt to impose rational order on CMIP5, rather than a qualitative leap" – What is the unit of measure here? Be careful to fully explain this phrase. As is it could easily be misused or misunderstood. If CMIP6 is just imposing order, why the large expenditure of resources?

The sentence states that CMIP6's structural innovation (DECK+endorsed MIPs) imposes order, not CMIP6 itself. We believe this sentence should be allowed to stand.

RC1-31 31. Page 12, line 32 – merely recommendations – As noted in my general comments, this paper needs to be much clearer what is meant by "recommendation".

RC1-32 32. Page 13, fig. 2 caption – data usage pattern – It seems to show data access, not usage.

Agreed, caption fixed.

RC1-33 33. Page 13, line 4 – Add "third party" in front of "copies". Also delete rest of sentence after "copies". It is not clear what is meant and seems redundant with first half of sentence.

We have added "third party" as suggested, but believe that the reference to the snapshots should be allowed to stand, as they were a notable community contribution to CMIP5.

RC1-34 34. Page 13, line 16 – More is needed here. How will a modeling center know when somebody is misusing its data? Is their any software existing or planned to help a center track its data? If so, it needs mentioned here. Furthermore, how can the license change in time in this scheme? Many centers make their data public after a period of time. It seems that the data files will need to be rewritten to change the license agreement. Is this the plan?

It is not possible to know when someone is using, let alone misusing, data, until someone notices and informs the data provider. We assume here the reviewer means by "misuse", a contravention of the license terms. If "misuse" is intended to mean mis-interpretation, we rely on the journal peer review process to prevent that.

Even if such tracking technologies were available, they would be quite intrusive, and quite surely involve privacy violations. However, when data is properly cited following the findings outlined in Section 5, data providers will be able to assess the utility of their data. We believe this will be a substantial advance over current practice.

As regards a center changing the terms of their license after the data has been published, that will require the issuance of a fresh PID. The terms of use require the user to adhere to the license associated with the PID used.

RC1-35 35. Page 14, line 1 – Reference needed (location) of the . . ..4.0 International License.

References visible now, see answer to RC1-15.

RC1-36 36. Page 14, line 13 – Consortium – Reference, web site?

References visible now, see answer to RC1-15.

RC1-37 37. Page 14, line 28 – Handle System – Reference.

Reference added, see page 16, line 23.

RC1-38 38. Page 15, line 4 – position paper – Where is this found?

References visible now, see answer to RC1-15.

RC1-39 39. Page 15, line 11 – DataCite infrastructure – Reference and location.

References visible now, see answer to RC1-15.

RC1-40 40. Page 15, line 22 – informally peer reviewed – This needs better defined. Unclear what this is.

Clarified, see page 17, line 21..

RC1-41 41. Page 15, line 27 – collections are static – How will groups correct errors found after the DOI is set? How will corrected data be made available? How will users know there are corrections?

We have clarified the treatment of errors, see page 17, line 27.. Users can discover if the data (PIDs) they are using are superseded using the errata service, Section 7.1.

RC1-42 42. Page 16, figure 3 caption – PID architecture . . . – PID is not found in the figure. How/What things in figure gets a PID? The current figure caption should read “A cartoon of data generation. . ..”

Caption to Figure 3 updated.

RC1-43 43. Page 16, line 5 – global Handle registry – Reference, web site needed.

Added above, see page 16, line 23.

RC1-44 44. Page 16, line 9 – CMIP6 Handle service – Reference, web site location needed

Added above, see page 16, line 23.

RC1-45 45. Page 16, line 11 – Add “for all simulation times” after “a single experiment”. . . if correct. If not, add details.

Clarified, see page 18, line 11.

RC1-46 46. Page 16, line 13 – position paper – Location?

References visible now, see answer to RC1-15.

RC1-47 47. Page 17, line 1 – Is there software to generate such a list? Seems like in multimodel studies such a list could be very long. Will journals publish a long list?

Indeed, a list of PIDs could be very long. In general, journals (including even leading ones such as *Science*) do not count supplementary material against page count limits or costs, nor do they include them in print versions, so the length should not be an issue.

If the reviewer is asking if the WIP is providing software for this purpose, the answer is no. But as the PIDs are in the netCDF files, it cannot be seen as difficult for scientists to harvest them from the files they use in their research.

Text unchanged.

RC1-48 48. Page 17, line 4 – RabbitMQ – Reference needed.

References visible now, see answer to RC1-15.

RC1-49 49. Page 17, line 20 – CMOR – Reference and web site needed.

References visible now, see answer to RC1-15.

RC1-50 50. Page 17, line 21 – PrePARE – Reference and web site needed.

References visible now, see answer to RC1-15.

RC1-51 51. Page 18, line 4 – QA nodes – I assume this is software. As written seems like hardware. More is needed.

It is indeed hardware. Text updated, see page 20, line 8..

RC1-52 52. Page 19, line 6 – realms – Define.

see page 22, line 2.

RC1-53 53. Page 19, line 7 – a set of tables – More is needed or delete.

Added, see page 22, line 4.

RC1-54 54. Page 19, line 13 – version-controlled code – Add "software that generates versioncontrolled code". It's all code. . .

Clarified, see page 22, line 9.

RC1-55 55. Page 20, line 21 – embedding – By whom? Modeler?

Clarified, see page 22, line 33.

RC1-56 56. Page 20, line 26 – position paper – Location?

References visible now, see answer to RC1-15.

RC1-57 57. Page 20, Replication section – I did not see any way for 1-off data sets to be issued PIDs. I appreciate that this is hard to enforced but the major impact user distribution sites should be required to issue PIDs in this framework. Numerically, the impact users are the single biggest group using CMIP data. Many of the sites serving them, preprocess the model data – generating new data sets, subsets, averages and so forth. These new data sets should not have model PIDs, but their own.

This is an excellent point, and we have added clarifying text, see page 25, line 22.

RC1-58 58. Page 21, line 4 – This statement implies that there are some CMIP data sets NOT accessible across ESGF. Is this true? More needed here. It is not clear what is meant.

Clarified, see page 23, line 14.

RC1-59 59. Page 21, line 11 – ICNWG – Reference, web site needed.

References visible now, see answer to RC1-15.

RC1-60 60. Page 21, line 13 – synda – Reference, web site needed.

References visible now, see answer to RC1-15.

RC1-61 61. Page 22, fig. 7 caption – CMIP6 replication team – It says CDNOT does this on the previous page. Correct.

Clarified in the caption to Figure 7, also see page 24, line 6..

RC1-62 62. Page 22, lines 3-6 – Does this break the data chain (PID and etc.)? More needed.

More explanatory text added, see page 25, line 8.

RC1-63 63. Page 23, Errata section – Are the replication nodes inside or outside of CMIP? This is not clear.

It is not clear what text the reviewer is referring to, as there is no reference to replication nodes in the Errata section. Nonetheless, as a general comment, we have attempted to move away from the notion of “inside” and “outside” nodes: for instance, see page 25, line 8.

RC1-64 64. Page 24, line 25 – our data – Change to “climate” or “CMIP” data.

Corrected, see page 28, line 10.