Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-52-RC1, 2018
© Author(s) 2018. This work is distributed under
the Creative Commons Attribution 4.0 License.

Geoscientific
Model Development
Discussions

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

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

Printer-friendly version

Discussion paper

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.

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.

Specific Comments

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.

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

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.

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

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

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.

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

cc BY

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

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

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

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.

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

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?

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.

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

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

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

18. Page 10, line 19 – smaller – I think “larger” is correct. . .nearer to 1. The exponent is larger.

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.

Printer-friendly version

Discussion paper

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.

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.

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.

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

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.

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

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. . ..

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

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

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.

Printer-friendly version

Discussion paper

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?

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".

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

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.

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?

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

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

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

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

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

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

CC BY

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?

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. . .."

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

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

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

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

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?

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

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

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

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

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

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

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

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

CC BY

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

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, pre-process the model data – generating new data sets, subsets, averages and so forth. These new data sets should not have model PIDs, but their own.

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.

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

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

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

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

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

64. Page 24, line 25 – our data – Change to "climate" or "CMIP" data.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52, 2018.

cc BY