Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-52-RC2, 2018

Geoscientific
Model Development
Discussions

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

Anonymous Referee #2

Received and published: 23 April 2018

General comments

The manuscript provides an overview of WRCP’s Infrastructure Panel (WIP) work, discussions and recommendations regarding the evolution of CMIP6’ cyberinfrastructure. It discusses some of the limitations of the current system, projections for future requirements and the rationale for decisions made by the WIP. It also describes some of the systems that are being put in place in preparation for CMIP6, in particular to better support citations, errata and provenance information for datasets and large ensembles, as well as managing the increasing volume of information to be stored.

The paper would benefit from an in-depth editorial review. It abuses bullet lists and the level of technical detail varies considerably across sections and topics. The result is that although interesting and pertinent, the manuscript is at times confusing and hard

to decipher. I was sometimes left with the impression that the paper was composed by copy-pasting sections of various WIP reports. The big picture (data-centric system) only really became clear to me at the end of a second reading; many of its implications are scattered across and not properly merged and highlighted in the conclusion. Indeed, the conclusion deserves some love, as at the moment it consists in fairly disjointed bullet list items.

The figures would also benefit from some attention as they apparently have been created independently from each other, and their content does not always support very well the text around them.

Most of my suggestions below concern style, as I understand that the manuscript has to reflect the WIP’s finding and work, which can’t be modified to please reviewers. I think however that the paper should leave some room to discuss criticisms made here and elsewhere and possibly respond to those. Among these would be the relatively small attention given to server-side analytics (raised by another referee). I also wonder why the paper does not discuss user-feedback? Is this the responsibility of the WIP, ESGF or CDNOT? How does the WIP consult users, what do they think of the tools that are built and operated for them? The paper makes no mention of recommendation concerning the user interface of public facing services. Does the priority setting process involves non-IT scientific users? Does the WIP include representatives from institutions operating dark repositories? Clearly they are prime users of CMIP data, yet feel the need to duplicate functionalities, and I somehow doubt it is only a matter of bandwidth optimization. Other topics not addressed by the paper are software security and open-access, as many of the technical issues that have frustrated users and complicated the life of software developers had to do with access tokens.

I feel the paper would be stronger if it discussed the feedback it got from the downstream climate science community and used this paper as an opportunity to communicate with it. I think there is a need for such a communication exercice after the frustrating experience some have had with CMIP5 data access in the past.

Detailed comments

Page | Line | Comment

1 7 "data as a commodity in an ecosystem of user" what does this mean exactly?

1 11 dataset-centric: Shouldn't the objective be for the system to be user-centric?

2 9 prescient: maybe a bit strong

2 15 3 -> three. As a general rule, spell numbers < 10

2 18 5 -> five

2 18 "formalized" used in last sentence and sentence is unclear. Mix of historical and current (DECK) denominations is confusing.

3 6 in in Figure 1

3 6 (some of) remove parentheses

3 8 Is the ESGF a "component". It looks to me as a loosely structured organization, with a "soft leadership", which indeed poses a number of challenges in terms of planning and delivery of operational software. This is possibly out of scope for this paper, but consider adding a paragraph somewhere in the paper about how ESGF organizes to implement WIP recommendations and some of the challenges it faces.

3 12 upon , a proposal

4 Figure 1: There is a site that looks to be in James Bay. Also is it really necessary to include personal contact email? This is something that can get outdated very fast.

5 6 It's not clear to what "which are summarized here" make reference to, "fundamental changes" or the "evolving scientific and operational requirements"?

5 7 The presentation is a bit awkward here, with a numbered list nesting a bullet list. I feel that this could all be written in text form. Also, the text suggests that the following items are "changes", but some of the opening statements are not.

6 9 review sentence syntax, second clause seems incomplete. Again, the bullet format feels innapropriate for dense and elaborate content.

6 21 The first bullet is the context, and the second the requirement. Please maintain some uniformity in the organization of ideas.

7 11 Idem

8 15 The data request concept is not properly introduced. Please clarify what it is and what purpose it is intended to serve before providing implementation details.

8 16 I feel that the level of details given on Data Requests far exceeds that of other sections. Who are the intended users? Data managers or analysts? Is the level of detail really relevant to this paper? Frankly, I read it a couple of times and I still don't understand the role it plays.

11 3 If I understand correctly, the single most important factor in the growth of data volume between CMIP3 and 5 is the number of variables that are archived. Yet, this issue does not appear to be formally addressed by the WIP as a volume problem further down in the text. At the moment, my understanding is that data is saved using the 1-file-per-variable approach. With hundreds of variables to probably co-vary in time and space, I'm guessing there might be compression benefits in storing multiple variables in the same file.

11 4 The use of a numbered list here makes little sense.

11 4 Please start the paragraph with the recommendation itself. Same suggestion applies to second recommendation.

11 13 Is the reference to the name of the actual python file really necessary? I suggest putting links to tools and software in appendix B.

12 20 CMIP archive size. Are you referring to CMIP5? Please clarify.

12 21 Sentence is confusing : "same causes, but with a much larger change"

13 Fig 2. Why "!" after local cache ?

13 14 Is that really "embracing" the dark repository model? I believe embracing that model would entail something a lot more ambitious such as a P2P network between official and dark repos that lets ESGF leverage dark repo to replicate and disseminate data. This is discussed later with synda (as far as I understand), but would deserve discussion here.

13 15 Review syntax.

13 18 I don't understand what this sentence means and how it relates to the preceding text.

13 20 Idem.

13 26 Please define "handles".

Figure 4 Who issues the PID? The data producer? This is only discussed later on page 18. I think it should be explained earlier.

20 17 Close parenthesis

21 5 Item 4 in section 2 only discusses model evaluation, not general data analysis.

Figure 7 It's not clear what this figure adds to the explanation.

24 24 Bullet list with no proper introduction. Please write a proper conclusion.

25 8 Is that really the message you want to end with? I suggest ending with an invitation to the climate science commnuity to provide feedback and suggestions, and generally get involved in the WIP's activities.

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