gmd-2018-52-RC2

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

RC2-1 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 openaccess, 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.

The reviewer raises several excellent points, and addressing those has considerably clarified the text. To begin with, we have (hopefully) addressed some of the stylistic issues, such as the "abuse" of bullets, point well taken. Second, we hope the text makes clear that the WIP has indeed taken the temperature of many of the players in this arena

(through in-depth consultations, not mass email): we note in particular that data *users* alone are not the target of this temperature-taking: data providers, managers of repositories (official and dark), and users have all been taken into account, and indeed are represented on the WIP itself. We have restructured the document to provide more context, including historical; and considerably rewritten the conclusions with more "love", we hope.

Also, as the reviewer notes, the findings here can be challenged if they are technically incorrect, but if they reflect the current community consensus (or lack thereof), we can but report that in this article, which we have done. The distinction between findings, requirements, and recommendations is made explicit, see page 4, line 9. We have strictly followed that nomenclature in the subsequent text.

Detailed comments

Page | Line | Comment

RC2-2 1 7 "data as a commodity in an ecosystem of user" what does this mean exactly?

Clarified, see page 1, line 7.

RC2-3 1 11 dataset-centric: Shouldn't the objective be for the system to be user-centric?

Of course, the intent is always to be "user-centric". In practice however, we believe we cannot anticipate all possible user needs, as the users of climate data are very diverse, and the science continues to evolve. This is why we have tried to introduce the notion of a data ecosystem, see page 1, line 7..

The distinction we are trying to make here is that there is no giant software infrastructure that is itself a single point of failure. Once users have datasets in their hands, or even their PIDs, they can continue to perform data transactions peer to peer even if, for instance, some key ESGF nodes go down.

This point is repeatedly brought up throughout the text, and here in the abstract in the phrase "less prone to systemic failure". No changes in response to this reviewer comment.

RC2-4 2 9 prescient: maybe a bit strong

see page 2, line 10.

RC2-5 2 15 3 -> three. As a general rule, spell numbers < 10

see page 2, line 16.

RC2-6 2 18 5 -> five

see page 2, line 21., see page 2, line 23.

RC2-7 2 18 "formalized" used in last sentence and sentence is unclear. Mix of historical and current (DECK) denominations is confusing.

reworded, reference added, see page 2, line 21.

RC2-8 3 6 in in Figure 1

see page 3, line 11.

RC2-9 3 6 (some of) remove parentheses

see page 3, line 10.

RC2-10 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.

Replaced “component” with “artifact”, see page 3, line 13.. We agree that the ESGF response to WIP requirements is out of scope for this paper, and a separate paper on the ESGF itself is warranted, once the software stack is finalized, and the system operational.

RC2-11 3 12 upon , a proposal

see page 4, line 2.

RC2-12 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.

A more appropriate figure has been substituted, see Figure 1 and see page 3, line 10.. Here the nodes are mapped to their geographic locations rather than relative to national boundaries.

RC2-13 5 6 It’s not clear to what "which are summarized here" make reference to, "fundamental changes" or the "evolving scientific and operational requirements"?

The clunky phrasing has been removed in the rewrite of Section 2.2.

RC2-14 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.

We thank the reviewer for this useful guidance, and the entire Section 2.2 has been rewritten as suggested, without bullets. Also, re “changes”, see page 5, line 26.

RC2-15 6 9 review sentence syntax, second clause seems incomplete. Again, the bullet format feels innapropriate for dense and elaborate content.

Entire section rewritten as noted above.

RC2-16 6 21 The first bullet is the context, and the second the requirement. Please maintain some uniformity in the organization of ideas.

Bullets removed, see see page 7, line 12.

RC2-17 7 11 Idem

Bullets removed, see see page 8, line 12.

RC2-18 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.

Context provided, see page 9, line 17.

RC2-19 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.

We have rewritten Section 3.1 at an appropriate level of detail, and hope, with the added context, its key role is now readily understood.

RC2-20 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.

The data volume discussion has been rewritten, see Section 3.4. As regards the second point, it is no doubt true that many of the variables exhibit considerable covariance, and are not statistically independent. But this remains still a matter for analysis and discovery. The current 1-variable-file remains a useful unit of analysis, a compromise for most users between too large files and too many files. Future infrastructure may indeed move in other directions based on the outcomes of CMIP6, and indeed POSIX "files" may themselves become obsolete, under certain technological evolutionary pathways currently at the cutting edge. We have added some discussion of these issues in the Conclusion.

RC2-21 11 4 The use of a numbered list here makes little sense.

List removed, see page 12, line 13.

RC2-22 11 4 Please start the paragraph with the recommendation itself. Same suggestion applies to second recommendation.

Section 3.4 has been considerably rewritten, and the recommendations restated at the end of the section.

RC2-23 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.

We have removed the excess detail in the rewrite of Section 3.4, see page 14, line 1..

RC2-24 12 20 CMIP archive size. Are you referring to CMIP5? Please clarify.

see page 14, line 9.

RC2-25 12 21 Sentence is confusing : "same causes, but with a much larger change"

Reworded, see page 14, line 10.

RC2-26 13 Fig 2. Why "!" after local cache ?

Gone, see caption to Figure 2.

RC2-27 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.

Clarifying text, and connection to replication discussion added, see page 15, line 14.. Replication is peer-to-peer in this system but based on units of atomic datasets, not packets, as in say BitTorrent. This does not preclude future development of P2P replication at the packet level.

RC2-28 13 15 Review syntax.

As this is a quote from another document, it doesn’t seem appropriate to change the syntax. The meaning seems fairly clear to us. The editors should let us know if they disagree.

RC2-29 13 18 I don’t understand what this sentence means and how it relates to the preceding text.

Some of this text is indeed out of place, rewritten. see page 15, line 4.. Some of the text is displaced to a discussion of the role of cloud analysis platforms in reducing data movement, see page 7, line 21..

RC2-30 13 20 Idem.

With the changes, we believe there is now continuity in the licensing discussion, see page 15, line 15..

RC2-31 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.

Reference for Handles added, see page 16, line 23.. PID issuance introduced here, see page 16, line 24..

RC2-32 20 17 Close parenthesis

see page 22, line 28.

RC2-33 21 5 Item 4 in section 2 only discusses model evaluation, not general data analysis.

This section has been generalized as suggested, see page 7, line 21..

RC2-34 Figure 7 It’s not clear what this figure adds to the explanation.

While we encourage ad-hoc replication, we wish to also underline the concerted effort to make sure high-value data is not repeatedly moved across geographic domains. This figure illustrates the efforts to coordinate replica nodes with sufficient storage, as well as the involvement of the network provisioners (ICNWG). We believe the figure should stay.

RC2-35 24 24 Bullet list with no proper introduction. Please write a proper conclusion.

Section 8 has been rewritten following reviewer's suggestion.

RC2-36 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.

See answer above.