Geosci. Model Dev. Discuss., 5, C187–C190, 2012
www.geosci-model-dev-discuss.net/5/C187/2012/

**Geoscientific
Model Development
Discussions**

Interactive comment on "Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data" *by* M. Stockhause et al.

Anonymous Referee #1

Received and published: 20 April 2012

This paper is about a distributed quality control scheme for climate model inter comparison. It deals with the quality assurance process for models containing petabytes of data replicated globally, to my mind a fairly unique challenge. The lessons here should hold for other disciplines such as astronomy.

Major comments

Why is distributed QA so important? (e.g., as discussed in p6 l6). It seems to me that if the data is generated at (say) modeling centre 1, then it should be possible to make nearly all QC checks at the point of creation. We would then expect it to be distributed among the repositories, along with the metadata and quality assurance files. I am not

convinced by the argument that a 'distributed' QA process is necessary. Certainly, one would need a mechanism to gather feedback from the widening pool of data users as the dataset is distributed. But I do not understand why one cannot simply distribute high quality data to begin with. If we compare to commercial data distribution efforts, eg. Google or Facebook, a query at data centre X should return the identical result to data centre Y. Doing distributed QA checks seems like a recipe for inconsistent quality results, or are you claiming that automated checks always produce the same results (I am skeptical this can be done).

I gather that the main reason for distribution is that the data is very important to other modeling centres, and so waiting until it is fully QA-3 is not feasible. In this case, you would distribute the early QC-1 data, and then each data set is quality checked at each node it has been distributed too. This seems like it would be worthwhile commenting on in more detail. How is consistency maintained - hashing the contents? How can an external user guarantee that the files at node 1 are the same as node 2? On page 8, l22, you mention the ESG Quality Curator is responsible for possibly doing the QA at the local data centre. How and why is this decided? Finally, why not 'release' the dataset again once it has been quality controlled? It seems like this may be an overly process intensive approach, and this seems to be demonstrated in the lessons learned section..

I think my biggest concern is with the introspective nature of the definitions of quality. Nowhere in the paper is the term technical or scientific QA defined or defended. It seems like these two would produce quite different mechanisms for quality control: at the technical level, presumably the concern is reproducibility and accuracy; at the scientific level, goodness of fit etc. Quality of climate models is likely beyond the scope of this paper, but I think it merits at least some discussion or exemplars.

It may be out of scope, but I was curious how likely you felt it was that scientists would actually use the tools provided. I understand there are some process requirements that mandate certain QA steps be taken, but it seems to me that tools like Attarrabi might

see little use, or if used, incomplete records are entered.

I really appreciate the addition of lessons learned from the initial pilot. I think this is possibly the most valuable contribution of the paper. Indeed, further to my immediately preceding remark, I would have enjoyed more detail on the e.g., 'several findings' from p11 l26.

Finally, there was little to no discussion of related work (possibly because there isn't much). Still, I think it would behoove the authors to consider other 'big data' approaches, such as in astronomy, to compare their techniques. And, as mentioned, surely the data replication approaches of commercial entities bears comparison, e.g. http://research.google.com/archive/bigtable.html, to name but one.

Minor remarks

p2 l1 - 'high state of quality and suitable for' reword as 'high state of quality, which is suitable'

p2 l3 - 'data replica' - why separate 'data' from replicated data? Presumably it is all the same data (refer to CAP theorem on consistency, availability, and partitioning)

p3 l20 - maybe a footnote or reference to netCDF somewhere, for those unfamiliar with this format.

p7 l15 These metadata is -> these metadata are or this metadata is

p8 l9 - the table is pages away from the first reference. Not a problem with hyperlinks, but presumably the final version will have it embedded? Or maybe this is an artifact of GMD's formatting.

- Given the abundance of acronyms a glossary would have been useful.

- p10 l1 the ThREDDS acronym is defined here but first occurs on p9

- p10 l8 it is considered bad form to begin a sentence with "Especially"

- - I think it would greatly help readers if the abstract gave at least one sentence explaining the findings. Is the concept successful?

- You refer in the abstract to the "most important part of the data" - but what is this, and how is it determined? In the introduction you then explain it is "IPCC relevant data" but again, this is somewhat nebulous.

- I appreciated that the data (in the sense of the tools) was already available and the web links all worked. It was a good way to understand the mechanics of the paper.

Interactive comment on Geosci. Model Dev. Discuss., 5, 781, 2012.