Geosci. Model Dev. Discuss., 5, C200–C206, 2012
www.geosci-model-dev-discuss.net/5/C200/2012/

**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.

M. Stockhause et al.

stockhause@dkrz.de

Received and published: 2 May 2012

Major remarks

1. Why is the distribution of the QA so important?
The CMIP5 data itself is created by multiple modeling centers worldwide and generally stored in national data nodes. There are 17 data nodes publishing CMIP5 data at the moment (cf. http://esgf.org/wiki/Cmip5Status/ArchiveView). The distributed QC control approach checks the data locally at each data node and stores the QC results centrally in a database. Thus the CMIP5 QC process is distributed in respect of the locations of QC checks but not in respect of distributing an individual QC check over multiple sites.

Data replication is slow due to narrow band widths. The advantage of a distributed QC approach is the possibility to check the data up to QC level 2 at the data nodes prior to data replication. Moreover, with the central repository for QC result storage, the results get immediately accessible by data creators, data users, and all QC managers for discussion, even prior to QC level 2 assignments.

2. Performing a distributed QC is a recipe for inconsistent quality checks. What was the reason for moving the QC L2 checks from the CMIP5 archive data nodes to the original data nodes?
Inherent in such a distributed approach is the danger of different applications. On the other hand, we wanted to share the QC efforts and make the QC results accessible as early in the QC process as possible. As a compromise the number of QC managers was reduced to three. These closely discuss application matters and use the same configuration template for QC L2 checks.

3. Why don't you store the QC results together with the data?
We discussed this option with the data infrastructure developers at PCMDI. The QC results are undoubtedly metadata but the argument against it were the following: The data publication to the ESG on a local data node (ESG publication) is connected with new data and with a considerable effort for the technician, which is too high to do it three times for the same data versions. An organizational disadvantage of such an approach is the fact that the CMIP5 archive centers have no control over the local data nodes. Thus, the publication of QC information cannot be enforced and might miss for a considerable part of the data. Therefore we followed the approach of collecting all metadata in CIM formats in the CIM repository for metadata exchange and harvesting by the ESG gateways and other portals. Currently, this repository is under development it but already harvests the QC CIM documents stored at WDCC along with the CIM questionnaire documents.

4. How are SQA and TQA defined?
The scientific quality assurance (SQA) is performed by the data creators. It consists

of data content and model-specific checks. The CMIP5 quality checks of level 2 are general checks applied to all variables published. These are statistical checks on data consistency and checks against the project requirements, i.e. Taylor's definition of the model output (http://cmip-pcmdi.llnl.gov/cmip5/docs/standard_output.pdf). During the final author approval step in QC level 3 within the GUI atarrabi, the data creator provides information on the scientific quality assurance procedure and results, which are added to the CMIP5 quality results.

5. Why do you perform QC L1 for replicated data files again?
The answer is that the QC L1 checks are integrated in the ESG data publication step, i.e. in the ESG publisher. QC L1 check criteria are requirements for the ESG data publication. Examples for QC L1 checks are readability, size>0, correctness of DRS_id components, and monotonic time values (cf. https://redmine.dkrz.de/collaboration/projects/cmip5-qc/wiki/Qc_l1).

6. How is data consistency maintained?
The background for the QC process is the conflict between a hard deadline for the scientists contributing to the IPCC-AR5 report and the continuing data changes/ESG data publications by the modeling centers. The national data centers decided to support their scientists in CMIP5 data access by replication of the latest data versions at a certain time. QC L1 data get replicated. As long as the data nodes follow a strict data versioning policy, the scientists have to deal with different but consistent data versions of replicated and original data. The identity of data is checked during data replication using MD5 checksums.

The QC approach simply has to deal with that fact and check for possible inconsistencies. This is part of the technical quality assurance (TQA) of QC level 3 checks. It is of special importance in distributed data infrastructures. Because of the non-reliability of individual identifiers for individual data, the cross-checks of the TQA use all available metadata on data identification in the different infrastructure components: DRS_id (including data version), tracking_id, MD5 checksums, and file size. The QC cannot

solve the problems of the data infrastructure but it has to ensure that the archived data assigned a DOI are consistent.

7. Usage of QC tools
To establish the QC tool of QC level 2 checks within the climate modeling community, it has to be improved in several perspectives: documentation, easy to install, easy to configure/apply, modularization of the code, and opening the code for community development. Within CMIP5 the QC L2 tool is applied as part of the CMIP5 QC procedure by the three QC managers in the archive centers or at least under supervision of them. Additionally, scientists at three modeling centers decided to apply the QC L2 tool to check their data before versioning and ESG publication in order to speed up the CMIP5 quality control process.

Atarrabi is already in production at WDCC for the DOI data publication of all kind of long-term archived data. It is WDCC's GUI for the communication between the publication agent at WDCC and the data creator / data author, who confirms the correctness of the data and stored metadata. In the case of CMIP5 we added the confirmation of CIM questionnaire metadata correctness. The author has to run through the whole process in atarrabi and explicitly finish it, before the publication agent finishes the QC L3 process by a DOI assignment to the data. Therefore incompleteness is impossible. The only varying content is the detail of the documentation of the scientific quality control. Required is a text comment. However, we encourage the authors to enter detailed assessment descriptions and results by uploading documents.

8. Examples for 'several findings' for which a QC L2 evaluation needs a statement of the data creator.
Whole records with filling values (if not commented in the data header) might indicate data loss or errors in the data post-processing procedure. Constant value records might be errors or might be caused by the specific model physics or model application.

9. QC approaches in other big data disciplines

We will continue the search for specific QC approaches in other disciplines. There seem to be little documentation on technical data checks (TQA) and on the different roles of data creators and data repositories within the data quality assessment or data curation procedure.

Minor remarks

1. Why do you distinguish between (original) data and data replica? Explain the relevance of the CAP theorem for the CMIP5 infrastructure.
We separated the data replica from the original data because of differences in the ESG data publication. Replica are assigned a flag. The other reason is that within CMIP3 and CMIP5 the original data at the national data centers might have higher latest versions than the data replica stored at the archive centers due to the time consuming replication process.

CAP theorem (A distributed system can have only two of the three characteristics consistency, availability, and partition tolerance.): The topic of this article is the quality control infrastructure for CMIP5. Within the QC we follow a central approach. All QC information is stored in a central database hosted at WDCC. The CAP problem is not applicable. The explanation of the data infrastructure is beyond the scope of the article. However, a few thoughts on the relevance of the CAP theorem in the data infrastructure on our QC approach: The traditional ESG data infrastructure does not store replica. Within the US all ESG gateways redirect the user to the data node storing the data for data access. The replication is a feature added for CMIP5. For the existing worldwide data nodes and worldwide the data access from some far away data nodes becomes cumbersome for data users. The data replication to a national data node gets advantageous. Applying the CAP theorem to the ESG infrastructure for CMIP5, the consistency is the non-solvable part. Every data node is independent of all the other nodes. Therefore data remain available if one data node is inoperative. Data node functionality continues if the communication to the other data nodes is lost. Total consistency in the respect that all data nodes host the same latest data versions cannot be achieved in a

system, replicating data in the order between 10s and 100s of TByte from several data nodes. However, a versioning was added to the ESG data publication process and to the DRS_id of the data. Thus, the data infrastructure can distinguish data published at different times using the version number or the DRS_id. There are no updates but only inserts of data into the distributed database. In theory this CMIP5 data infrastructure avoids data identification problems. The visibility of the version to the ESG gateway users can be improved, though.

For the QC infrastructure such a data infrastructure is appropriate, since for a QC procedure it is essential to be able to identify different data versions. However, the independence of the data nodes causes problems for the QC procedure, because of differences in versioning policies and variations in the data node configurations, and in the organization of responsibilities introduce inconsistencies. E.g. a data node manager decides to let the data creators maintain their data and does only link to the original data location (for a specific versioned data file). Then the data creator can do minor changes and decide not to inform the data node manager to assign a new version to it. This leads to inconsistent versions for this file and its replica (and inconsistent DRS_ids). The checksum in the Thredds catalogue is wrong, the tracking_id might be wrong. Then the file size is the only indicator left for the data inconsistency within the QC level 3 cross-checks.

2. Glossary
As your recommendations for additional examples, we clearly see the need of a glossary to help readers not familiar with the technical components of the CMIP5 infrastructure. We will add one in the next paper revision.

3. "IPCC relevant data" vs. "output1 data" vs. "replicated data"
You are right that the definition of the replicated part of the CMIP5 data is somewhat nebulous. They consist of comments in Taylor's model output requirements (cf. link at Major comments item 4.). We recently analyzed that nearly 90

4. Thank you for the other valuable minor remarks. We will consider them in our next revision of the article.

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