Geosci. Model Dev. Discuss., 5, C325–C327, 2012
www.geosci-model-dev-discuss.net/5/C325/2012/

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: 29 June 2012

We extracted four questions:

1. Could you provide actual numbers for the data volume for CMIP5?
For CMIP3 / IPCC-AR4 over 35 TB of data were collected (Williams et al., 2008). For CMIP5 2-3 PB are expected. Including additional data the estimations reach several PBs. The current data amount is 1.4 PB for the latest data versions (see http://www.esgf.org/wiki/Cmip5Status/ArchiveView). The amount of CMIP5 data in all ever published data versions can be estimated as about 1.5 to 2 times that size.

2. Why can the TQA only be applied by the data intermediaries?

The Technical Quality Assurance (TQA) is a consistency check of metadata and data at the archiving center, which is in this case the DOI Publication Agency WDCC. The metadata in the archive is checked against the metadata of the ESG data node, the metadata of the CIM questionnaire, and the metadata in the QC database. Since all the processes of metadata and data collection are independent of each other, the data identifier and checksums are stored along with the data. In plain words, we check if the archived data, the data in the ESG data node and the quality checked data are the same.

3. What are the causes slowing down the data publication process if it is not the QC procedure itself?

a. There is no timeline for the CMIP5 data submission. Hard deadlines exist only for the scientific publications which enter the IPCC-AR5 report. For the report *The Physical Science Basis* (Working Group 1) the paper submission deadline is July 31, 2012. The underlying CMIP5 data for those scientific publications which enter that report can be revised and be published as new versions any time. CMIP5 as a scientific model intercomparison project has not set a deadline for data changes because data revisions are regarded as data improvements. The direct consequence for the QC process is that the modeling centers as authors for the DOI data publication do not wish to archive their data in a state of expected further data corrections. Indirectly, the still not stable data slows down the replication to WDCC for data archiving.

b. Though the estimations for the data amount were quite accurate, the time needed for data replication was underestimated in the data infrastructure development. As the presence of a data replica in the data archive of the Publication Agency WDCC is a precondition for the DOI assignment, the QC process cannot start with QC level 3 checks after reaching QC level 2. Thus the distributed and federated QC approach is capable to check the data in level 2 as it is published in revised data versions. The described QC infrastructure provides the Publication Agency immediate access to all QC results of level 2 and access to other metadata for the cross-checks of QC level 3.

GMDD

5, C325–C327, 2012

4. Is it a failure for the designed QC process, if a large portion of the data never makes it through QC level 3?

First of all the QC checks of level 2 add to the reliability of the published data comparing with the situation for the CMIP3 / IPCC-AR4 data. The basic and automated QC level 2 checks cannot be sufficient to prove the scientific quality of the data, though.

We would rather call it a success if the cross-checks of QC level 3 identify data inconsistencies, e.g. differences in unique identifiers of the same data version between original data and the data replica in the archive of the Publication Agency.

We expect that in the future national Publication Agencies will provide archiving and DOI services. That will solve the data replication problem for the data DOI assignment. However, it won't help the scientists accessing the data via the internet because they will still suffer from narrow bandwidths for data stored on a different continent.

Reference

Williams, D. N., Ananthakrishnan, R., Bernholdt, D. E., Bharathi, S., Brown, D., Chen, M., Chervenak, A. L., Cinquini, L., Drach, R., Foster, I. T., Fox, P., Hankin, S, Henson, V.E., Jones, P., Middleton, D.E., Schwidder, J., Schweitzer, R., Schuler, R., Shoshani, A., Siebenlist, F., Sim, A., Strand, W.G., Wilhelmi, N., Su, M. (2008): Data management and analysis for the Earth Sytem Grid, J. Phys., Conf. Ser., 125, 012072, doi:10.1088/1742-6596/125/1/012072, 2008.

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