Geosci. Model Dev. Discuss., 5, 781–802, 2012
www.geosci-model-dev-discuss.net/5/781/2012/
doi:10.5194/gmdd-5-781-2012

Geoscientific Model Development Discussions

Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data

M. Stockhause, H. Höck, F. Toussaint, and M. Lautenschlager

German Climate Computing Center (DKRZ), World Data Center for Climate (WDCC), 20146 Hamburg, Germany

Received: 22 March 2012 – Accepted: 23 March 2012 – Published: 13 April 2012

Correspondence to: M. Stockhause (stockhause@dkrz.de)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

The preservation of data in a high state of quality and suitable for interdisciplinary use is one of the most pressing and challenging current issues in long-term archiving. For high volume data such as climate model data, the data and data replica are no longer stored centrally but distributed over several local data repositories, e.g. the data of the Climate Model Intercomparison Project No. 5 (CMIP5). The most important part of the data is to be published as DOI according to the World Data Center for Climate's (WDCC) application of the DataCite regulations. The integrated part of WDCC's data publication process, the data quality assessment, was adapted to the requirements of a federated data infrastructure. A concept of a distributed and federated quality assessment procedure was developed, in which the work load and responsibility for quality control is shared between the three primary CMIP5 data centers: Program for Climate Model Diagnosis and Intercomparison (PCMDI), British Atmospheric Data Centre (BADC), and WDCC. This distributed quality control concept, its pilot implementation for CMIP5, and first experiences are presented.

1 Introduction

The International Panel on Climate Change (IPCC) aims to establish one common climate model data archive to advance the knowledge of climate change and variability. The results collected within the Climate Model Intercomparison Project No. 5 (CMIP5) are intended to underlie the coming fifth assessment report (IPCC-AR5). CMIP3 data for the last report IPCC-AR4 were collected and provided centrally by the Program for Climate Model Diagnosis and Intercomparison (PCMDI) without version control and with compact unformalized metadata information, which was imprecise in respect of model and simulation descriptions. The data volume for CMIP5 is expected to reach nearly 100 times that of CMIP3 (Taylor et al., 2012).

GMDD

5, 781–802, 2012

These experiences from CMIP3 together with the expected data volume led to three main improvements for the CMIP5 data infrastructure:

- Data is stored in several decentralized data nodes connected by the Earth System Grid (ESG; Williams et al., 2009, 2011). Three of them located at major data centers have built a federated system of data archives (also called primary CMIP5 data portals, Taylor et al., 2012): PCMDI, the British Atmospheric Data Centre (BADC), and the World Data Center for Climate (WDCC). These centers committed to hold replica of the most important part of the CMIP5 data, i.e. IPCC relevant data, on hard disks for quick access and data security.
- Information on models and simulations is enlarged significantly. The metadata schema used is the Common Information Model (CIM) developed by METAFOR and collected via a web-based questionnaire (Guilyardi et al., 2011).
- Data curation was improved by introducing a versioning concept and a quality assessment process providing a uniform identification of datasets as well as a persistent identifier DOI (Digital Object Identifier) for data citation in scientific publications. The data DOI, like a DOI for printed papers, gives scientific credits to data creators for their work and allows for persistent and direct data access.

The quality assessment procedure for CMIP5 has to support the federated data infrastructure and incorporate all available metadata resources, especially CIM metadata and those stored in the self-describing data headers of the netCDF files. A general concept for a distributed and coordinated quality assessment procedure suitable to use in a distributed data infrastructure was developed (Sect. 2). This concept was altered and adapted for its pilot application within CMIP5 (Sect. 3).

GMDD
5, 781–802, 2012

2 Concept of a distributed quality assessment of high volume data

Quality control and description of data in repositories and especially in long-term archives are generally viewed as essential. Moreover, a demand for more efficient evaluation services to convert data into information and information into knowledge is detected (Overpeck et al., 2011). This is of special importance for open-access data of interdisciplinary use, where no direct contacts between data users and original data creators exist any longer. However, contents of the quality checks as well as definitions of quality levels and the overall quality procedure vary significantly between data types and scientific disciplines.

The ESIP (Federation of Earth Science Information Partners), a consortium of 120 organizations, formulated some principles on data stewardship and recommended practices (ESIP, 2011): Quality assessment and its documentation are tasks of the data creator. Data intermediaries like repositories should set time limits for quality control procedures in order to prevent it from delaying data accessibility. Data intermediaries additionally function as communicators between data creators and data users. ESIP (2011) focuses on the scientific content of the data in its principles for quality assessment. For scientific data distributed over several repositories this scientific quality assurance (SQA) has to be accompanied by a technical quality assurance (TQA). The TQA checks data and metadata consistency among the distributed data and metadata repositories and might include a check against data and metadata standards. This TQA can only be applied by the data intermediaries at the data repositories adding the TQA task including its documentation to their communicator role.

Quality control procedures of high volume data have to be carried out at the storage location before opening the repository for interdisciplinary data access and use. Together with the trend towards decentralized data repositories, quality control procedures have to become distributed/federated themselves and need to be coordinated and standardized (Sect. 2.2).

GMDD
5, 781–802, 2012

2.1 Data quality control procedure for model data

In general increasing quality levels of data correspond to increasing data suitability for a broader community which subsequently is given access. Roughly quality checked data of the initial quality level is suitable for a specialized scientific community. Other than for observational data, where quality levels are commonly connected with data changes or the derivation of new data products, quality levels of model data are generally not connected with data processing but only with data validation steps. Therefore model data is not altered during the quality procedure at the data repositories, but accepted or rejected. Model data is revised only by the data creator. The data delivered by the data creators is strictly version-controlled. For new data versions the quality control (QC) process is started over again.

A typical model quality procedure consists of three levels:

- *QC Level 1:* Quality checks on formal and technical conformance of data and metadata to technical standards,
- *QC Level 2:* Consistency checks on data and metadata to project standards,
- *QC Level 3:* Double- and cross-checks of data and metadata, check of data accessibility (TQA), and documentation of the data creator's quality checks (SQA).

After finalizing the quality assessment procedure with QC Level 3, the data is long-term archived and should be published according to the DataCite DOI regulations for publishing scientific data (DataCite, 2011; Klump et al., 2006). DataCite (http://datacite.org) is a registration agency of the IDF (International DOI Foundation, http://www.doi.org). Analogue to the publication of an article in a scientific journal the data publication makes the data citable and accessible. Thus, it can be included in a scientist's list of publications to give him credit for his efforts on data preparation and the SQA. This data publication is performed by a DOI publication agency which committed itself to grant persistent data access via the assigned DOI.

GMDD
5, 781–802, 2012

A DOI is assigned to collections of individual datasets, which are suitable for data citation purposes in the scientific literature. For climate model data the simulation is chosen, which includes all data of a model application or even all data of all realizations (ensemble members) belonging to a prescribed scenario or projection.

2.2 Distributed quality assessment approach

For distributed repositories of high volume model data the identical quality assessment is performed at different locations. Basic preconditions for such a distributed QC are a uniform and coordinated QC check procedure with a uniform QC result evaluation. These have to be independent of the QC manager performing the QC. Furthermore, an appropriate infrastructure has to be built to support result analyses and result sharing. The final data checks for DOI data publication rely on the results of the preceding quality checks.

The technical infrastructure of the distributed quality control approach consists of three main components (Fig. 1):

1. *A central project metadata repository* used for quality information storage: The project metadata repository provides information on quality check configurations and other input data if used, quality check performance, and quality results as well as on provenance and status.

2. *Locally-installed QC service packages* supporting the overall QC procedure by adding a service layer on top of established QC checker tools

3. *User interface to support the data creator's SQA (SQA GUI):* The SQA GUI supports the data creator in inserting quality procedure description and quality results as well as in reviewing the basic metadata, i.e. basic data citation information. This is part of the checks for DOI data publication.

The QC service package consists of different services to support:

GMDD
5, 781–802, 2012

- the *analysis* of QC results by exception statistics, provenance information, and plotting,
- the *insert* of QC tool application information and results into the Project Metadata Repository,
- the *assignment* of QC levels (including a possibility to exclude certain data from the assignment to a data collection), and
- the final data checks for DOI data publication by *providing information* on project metadata.

2.3 Embedding the distributed quality control into a federated data infrastructure

Quality Control procedures rely on data and metadata accessibility. Data is stored in different local data repositories or Data Nodes (DN). Metadata (MD) is created during the whole project life time, starting with the description of model and model application (MD on model/simulation) provided by the data creator and inserted via a Metadata GUI (Fig. 2). Later metadata on the data in the data nodes and metadata on quality of the different QC level checks are added. These metadata is collected and stored centrally in a project metadata repository.

The available metadata information in the project metadata repository is used within the cross- and double checks (TQA) of the DOI publication process (QC Level 3). At the end of the project and the QC procedure, data and metadata are long-term archived, i.e. a data copy is stored at a data center along with all available metadata out of the project metadata repository. DOI published data is long-term archived at the long-term archive (LTA) of the DOI publication agency. The DOI is assigned via the registration agency DataCite to the IDF and is integrated into the global handle system. The DOI resolves to an entry metadata page hosted by the DOI publication agency.

GMDD
5, 781–802, 2012

Data catalogs support data discovery by harvesting the metadata information of the project metadata repository or the LTA, and the DOI handle system supports data discovery of the long-term archived data after the end of the project.

3 Application of the distributed quality control in CMIP5

The distributed quality assessment procedure was adapted for the pilot implementation in the CMIP5 project. Detailed information on the quality control procedure within CMIP5 is available at http://cmip5qc.wdc-climate.de.

3.1 Quality control procedure within CMIP5

The definition of quality control levels and its implications are summarized in Table 1. The quality procedure workflow with its actors is sketched in Fig. 3. The data collection for QC level assignment within CMIP5 is a CMIP5 simulation, i.e. all data of all realizations of one experiment carried out with a certain model. CMIP5 data is ESG published at decentralized local data centers. Most of the modeling centers decided to either host their own data node or ESG publishes their data at a national data node. The data submission step is performed by the integration of the data node's metadata in the ESG gateway catalogs. During ESG publication at the data nodes, QC level 1 checks are performed, i.e. CMOR2 and ESG publisher conformance. Access of data of level 1 is restricted to selected users, who contribute to the QC process by reporting problems and errors to the data nodes or the ESG gateways.

In a second submission step the data is copied from the local data centers to one of the primary CMIP5 data centers (PCMDI, BADC, or DKRZ) for quality checks of level 2, followed by an ESG data publication at the CMIP5 data center. The QC Manager at the CMIP5 Center can alternatively decide to carry out the QC L2 checks at the local data center in parallel or prior to data replication. An example for a QC L2 check criteria is the continuity of the time axis and the usage of the accurate CF (Climate and Forecast)

GMDD
5, 781–802, 2012

standard name for the variable. Data of QC level 2 is accessible for the CMIP5 research community. By strict version control of the ESG published datasets, an identification of data is possible for users that downloaded the latest data version at a certain time before QC procedure completion (DOI data publication).

Data of QC level 2 is replicated among the three primary CMIP5 data centers (Fig. 3). The QC level 3 process is carried out by WDCC as DOI publication agency. The cross- and double checks (TQA) include metadata on data extracted from the THREDDS data servers, CIM metadata on models and simulations harvested from the atom feed at BADC, and quality results accessed from the QC repository. Examples for TQA check criteria are the identity of model names or identifiers like the tracking_id in all metadata repositories. QC on CIM metadata is carried out separately by BADC prior to the final QC level 3 checks.

3.2 Implementation of the distributed quality control for CMIP5

The existing data infrastructure of the Earth System Grid (ESG; Williams et al., 2009) was adapted to CMIP5 requirements (Williams et al., 2011). Data replication functionality was added to exchange identical copies of the most important data among the three primary CMIP5 data centers PCMDI, BADC, and WDCC. The federation approach is motivated by improved response times for users' data access via the internet compared to a single central repository and reasons of data security. Data discovery functionality in the gateways was enhanced from the search on essential data information to enhanced information on data, model, simulation, and platform. User registration, authentication, and authorization were changed from a central to a federated approach.

The enhanced metadata on model and simulation is collected via a web-based questionnaire (Fig. 4) and stored in the CIM repository in CIM metadata format (http://metaforclimate.eu; Guilyardi et al., 2011). The CIM repository is meant to provide detailed and reliable long-term metadata information for different portals like the ESG portal or the European IS-ENES portal in future. The CIM metadata format functions within CMIP5 as exchange metadata format. The ESG data nodes incorporate a

GMDD
5, 781–802, 2012

THREDDS Data Server (TDS, http://www.unidata.ucar.edu) that extracts metadata from the netCDF file headers. This metadata is mapped to the CIM format and added to the CIM repository. A similar mapping to the CIM metadata schema is performed for the QC information stored in the QC repository.

The distributed quality control approach outlined in Sect. 2 was adapted to include existing infrastructure components: the data nodes with ESG publisher and TDS as well as the CIM metadata repository hosted at BADC (Fig. 4). Especially the technical quality assessment (TQA) part of QC level 3 checks had to be altered to read the metadata schemas of TDS and CIM and significantly extended to include their contents in the checks. For the scientific quality assurance (SQA) and the final author approval of the data by the data creator, a graphical user interface is used for the interaction between data creator and DOI publication agent at WDCC: atarrabi (http://atarrabi.dkrz.de/atarrabi2; Fig. 4). Finally, a service to support the CMIP5 data centers in the prioritization of data replication was set up providing a list of the ESG publication units of QC L2 or L3 including a filter functionality.

Since the developments of metadata and data infrastructures are still ongoing and quality information is not harvested, yet, a couple of additional quality related services were established (http://cera-www.dkrz.de/WDCC/CMIP5):

- *the QC result service:* http://cera-www.dkrz.de/WDCC/CMIP5/QCResult.jsp,
- *the QC status services:* GUI for user access: http://cera-www.dkrz.de/WDCC/CMIP5/QCStatus.jsp, Java servlet for data replication control at the gateways,
- *the CIM quality document publication via atom feed:* http://cera-www.dkrz.de/WDCC/CMIP5/feed, and
- *the data citation service:* for data users: http://cera-www.dkrz.de/WDCC/CMIP5/Citation.jsp.

GMDD
5, 781–802, 2012

All QC L2 results stored in the QC repository become immediately accessible to the community via the QC result service. CIM quality documents are created and published via atom feed at QC level 2 and QC level 3 assignments. The data citation service provides a central entry to search by tracking_id for the current preliminary citation recommendation in case of data of QC level 1 or 2 and for the persistent citation of DOI published data of QC level 3.

Thus, the QC repository serves within CMIP5 not only as intermediate QC result storage facility to support the QC process but additionally as long-term source of quality results and quality-related information for the climate community.

WDCC plays a double role in the federated quality procedure of CMIP5 by performing QC L2 on their share of the CMIP5 data and functioning as DOI publication agency for all replicated long-term archived data (QC L3).

3.3 Experiences of the CMIP5 quality approach

First experiences of the federated quality assessment procedure in CMIP5 are encouraging. The federated QC approach is capable to serve as QC procedure for CMIP5. The QC has helped to find data inconsistencies in order to improve the CMIP5 data quality. First DOIs on data are assigned, e.g. doi:10.1594/WDCC/CMIP5.MXELAM. Though the QC concept and its implementation worked out fine, several problems occurred during the QC application:

1. Most modeling centers are still in the process of ESG publishing new data versions or additional data, e.g. cfMIP data. Since no deadline exists for the creation of data for the DOI data publication process, the QC L2 process for the CMIP5 simulations cannot be finished but has to be continued several times. Among other reasons QC L2 findings contribute to these data revisions. The contact between QC L2 manager and data creator is more intensive than expected. Several findings during QC L2 require the interpretation of the data creator to distinguish a real error from a minor model-specific issue.

GMDD
5, 781–802, 2012

2. The time necessary for data replication was underestimated. Reasons are narrow bandwidths together with the ongoing data changes. As the long-term data archiving at the DOI publication agency is a precondition for DOI assignment, the DOI data publication is significantly delayed. The former aim to provide DOI data citation references for scientists of IPCC Working Group I (WG I) was altered in order to provide those citations for WGs II and III in fall 2012. Additionally, the data aggregation for a data DOI had to be changed from including all data of a CMIP5 simulation into including at least the monthly and yearly data of a CMIP5 simulation. The data of higher temporal frequency will be published as a new DOI, related to the first DOI via DataCite relation *isSupplementTo* (DataCite, 2011). The DOI data publication decision is a compromise between data completeness and providing data citation regulations for scientists, especially for those contributing to the 5th IPCC assessment report.
   The data replication problem had three implications for the QC procedure: First, the QC L2 checks had to be distributed even more to enable QC L2 applications directly at the data nodes. The QC manager at PCMDI, BADC or DKRZ remains responsible for the QC L2 assessment and thus the QC level 2 assignment. However, he could either delegate the QC L2 checker tool run at the data node to the data node manager or run it himself. Secondly, the QC procedure for level 3 had to be altered to enable the exclusion of non-replicated data before starting the QC L3 process. And, thirdly, the scientist of WG I had to be supported in the citing of CMIP5 data, esp. of data of QC levels 1 and 2, i.e. data without DOIs. WDCC set up its citation service for that purpose.

3. In the data infrastructure data, replication as well as the inclusion of data replica and multiple data versions are not fully supported by the current ESG gateways. Thus, QC status or DOI data are not integrated in their data discovery functionality, but remain separated pieces of information. Even the available CIM quality documents are not harvested by the ESG gateways, so far, due to other more urgent development issues. To bridge this intermediate invisibility of QC status and

GMDD
5, 781–802, 2012

DOI, WDCC set up the QC status and result services for data users. These need to be linked from the ESG gateways.

4. The different unique identifiers in use for CMIP5 data turned out to be not strictly unique in every case. The infrastructure components use strict DRS (Data Reference Syntax; http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf) names only down to the granularity of an ESG publication unit. Moreover, there exist different dialects for the DRS syntax: Data production uses the CMOR2 DRS syntax without versioning and the data nodes use the ESGF (Earth System Grid Federation, http://esgf.org) DRS syntax. DRS names for institutions and models are defined by the modeling centers twice, in the CIM questionnaire and in the file directory. Therefore, these names potentially differ between data (TDS, QC data base) and CIM questionnaire documents. These differences require relatively high mapping efforts during QC L3 cross- and double-checks. Additionally, local data centers tend to publish the same data version again, in the case of minor changes in one variable within an ESG publication unit.
   The other unique identifiers are the *tracking_id* written by CMOR2 and the *MD5 checksum* written and published during ESG publication. In cases of files not written with CMOR2, tracking_ids of two different files might be equal. Regarding checksums, the data replication managers have found cases of checksums not re-calculated and re-published after data changes leading to a replication error message. These identifier problems result in the extension of the QC L3 cross-checks criteria to integrate consistency checks on all these identifiers plus the file size available in the different infrastructure components to ensure metadata and data consistency. A general problem lies in the identification of the CIM simulation document for cross-checking, if the data creator e.g. decided to use different model names in the file system and in the CIM questionnaire. The DOI publication agent has to identify the problem before starting the QC L3 process and ask the data creator to change the simulation description in the CIM questionnaire during the SQA. Thus, the DOI publication agency performs a second slightly simplified

GMDD
5, 781–802, 2012

TQA cross-check before the data creator starts the SQA. The documented final TQA follows after the scientist's approval. These additions and changes led to a significantly increased complexity and duration of the QC L3 procedure.

These problems illustrate the importance of the use of unique identifiers in a distributed and federated infrastructure. The current CMIP5 infrastructure has defined unique identifiers but does not enforce their usage enough. As long as *tracking_id* and *checksum* might not be updated during data changes, their usability is restricted. The lack of a controlled vocabulary of the DRS name components stored centrally and accessed by all infrastructure components, led to error-prone DRS name mappings.

Furthermore, a closer coupling of the different infrastructure components is desirable. The current technical infrastructure of CMIP5 consists of several technical components for similar purposes, e.g. metadata is stored in the data nodes, the CIM repository, the QC repository, and at the DOI publication agency. Apart from these CMIP5 components other portals like IS-ENES (http://verc.enes.org/) harvest and store their own metadata. This current infrastructure is more complex than necessary, which introduces additional metadata exchanges between the sites. In an international cooperation like the CMIP5 infrastructure the ideal single project metadata repository is not achievable. However, a central metadata repository for metadata exchange can be established, preferable storing metadata in a uniform format, e.g. CIM. Such a central repository would enable the QC L3 procedure to access only one metadata resource instead of currently four (TDS, CIM, QC DB, and metadata repository of the DOI publication agency). Furthermore, the ESG gateways and other portals like IS-ENES can use this metadata repository for harvesting, ensuring data discovery on identical metadata resources.

WDCC is going to raise these issues as partner of the recently started joint international initiative ES-DOC-Models (Earth System Documentation-Models, http://earthsystemcog.org/projects/es-doc-models/), which aims to develop metadata services for climate projects.

GMDD
5, 781–802, 2012

4 Conclusions

A concept of a distributed quality assessment procedure for high volume data is presented together with its pilot implementation for the international project CMIP5. Several adaptations of the concept had to be implemented for CMIP5 to integrate existing infrastructure components and to bridge the lack of planned but not yet realized ones. In CMIP5 QC level 1, checks are performed at decentralized data nodes. QC level 2 managers are located at the primary CMIP5 data centers to co-ordinate the QC checks for the data submitted to their gateways. Results of QC level 2 checks are collected in a central repository, which enables the DOI publication agency to start with QC level 3 checks during data replication among the primary CMIP5 data centers.

The QC approach is capable of supporting the overall QC procedure for CMIP5. First DOIs are assigned to CMIP5 simulations with finished quality control procedure. Due to inconsequential usage of identifiers and naming conventions, the QC procedure especially QC L3's cross- and double-checks became extremely complex and delicate. The quality control has detected several inconsistencies in the delivered data and thus shown its value. Moreover, the QC procedure is not slowing down the data publication process. However, delayed data delivery of the modeling groups and slow data replication rates, have led to a significant delay in the DOI data publication of CMIP5 data. The presence of several globally distributed DOI publication agencies with long-term archives could overcome the DOI data publication delay but not the data replication delay.

The roadmap for the distributed QC consists of:

- *Consolidation:* The distributed QC needs to be integrated into the ESG infrastructure more closely. The joint international initiative ES-DOC-Models can provide the necessary standards for that. Additionally, the long-term archive phase after the end of the project has to be clarified and supported by service level agreements between the metadata long-term archive CIM and the data long-term archive of the publication agency WDCC.

GMDD
5, 781–802, 2012

– *Transparency:* The used QC checker tools as well as the QC assessment workflow should be made available for the community, accompanied by improved tool documentations. Provenance aspects of the QC application have to be collected more precisely.

– *Application:* The approach is to be applied to data of other projects, which might require the integration of alternate QC level 2 checker tools, developed by and established or at least accepted within the scientific community. The WDCC is going to integrate the outlined quality control process into their local long-term archiving implementation. The aim is to establish a fully documented and quality proven long-term data archive at the WDCC of DKRZ.

Acknowledgements. This work was funded by the Bundesministerium für Bildung und Forschung (BMBF; German Federal Ministry of Education and Research) under the support code 01LG0901A. We thank the ESGF group members for their contributions and support, with special thanks to Bryan Lawrence and Karl Taylor for intensive discussions about the integration of the quality control into the overall CMIP5 process.

The service charges for this open access publication have been covered by the Max Planck Society.

References

DataCite: DataCite Metadata Scheme for the Publication and Citation of Research Data, Version 2.2, doi:10.5438/0005, July 2011. 785, 792

ESIP: Interagency Data Stewardship/Principles, http://wiki.esipfed.org/index.php/Interagency_Data_StewardshipPrinciples, last access: 20 March 2012, Federation of Earth Science Information Partners (ESIP), 9 September 2011. 784

Guilyardi, E., Balaji, V., Callaghan, S., DeLuca, C., Devine, G., Denvil, S., Ford, R., Pascoe, C., Lautenschlager, M., Lawrence, B., Steenman–Clark, L., and Valcke, S.: The CMIP5 model and simulation documentation: a new standard for climate modelling metadata, CLIVAR Exchanges, No. 56, 16, 42–46, 2 May 2011. 783, 789

GMDD

5, 781–802, 2012

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., Schindler, U., Sens, I., and Wächter, J.: Data Publication in the Open Access Initiative, Data Science Journal, 5, 15 June 2006. 785

Overpeck, J. T., Meehl, G. A., Bony, S., and Easterling, D. R.: Climate Data Challenges in the 21st Century, Science, 331, 700, doi:10.1126/science.1197869, 11 February 2011. 784

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, submitted, B. Am. Meteorol. Soc., doi:10.1175/BAMS-D-11-00094.1, 2012. 782, 783

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., Fraser, D., Garcia, J., Hankin, S., Jones, P., Middleton, D. E., Schwidder, J., Schweitzer, R., Schuler, R., Shoshani, A., Siebenlist, F., Sim, A., Strand, W. G., Su, M., and Wilhelmi, N.: The Earth System Grid: Enabling Access to Multimodel Climate Simulation Data, B. Am. Meteor. Soc., 90, 195–205, doi:10.1175/2008BAMS2459.1, 2009. 783, 789

Williams, D. N., Taylor, K. E., Cinquini, L., Evans, B., Kawamiya, M., Lautenschlager, M., Lawrence, B. N., Middleton, D. E., and ESGF contributors: The Earth System Grid Federation: Software Supporting CMIP5 Data Analysis and Dissemination, CLIVAR Exchanges, 56, 16, 40 ff., 2 May 2011. 783, 789

GMDD

5, 781–802, 2012

Table 1. CMIP5/IPCC-AR5 quality control levels and their implications.

	QC Level 1: CMOR2, ESG Conformance of Data and CIM Conformance of Metadata	**QC Level 2:** WDCC Conformance and subjective controls	**QC Level 3:** DOI Data Publication via DataCite
Data	Data preliminary; no user notification about changes; performed for all data; metadata may not be complete	Not finally agreed; no user notification about changes; performed for replicated data	published and persistent data with version and unique DOI as persistent identifier; performed for replicated data
Access	constrained to CMIP5 modeling centers	constrained to non-commercial research and educational purposes	constrained to non-commercial research and educational purposes or open for unrestricted use
Citation	no citation reference	informal citation reference	formal citation reference
Quality Flag	*automated conformance checks* passed	*subjective quality control* passed	*approved by author* (in case of newer DOI available: *approved by author, but suspended*)

GMDD
5, 781–802, 2012

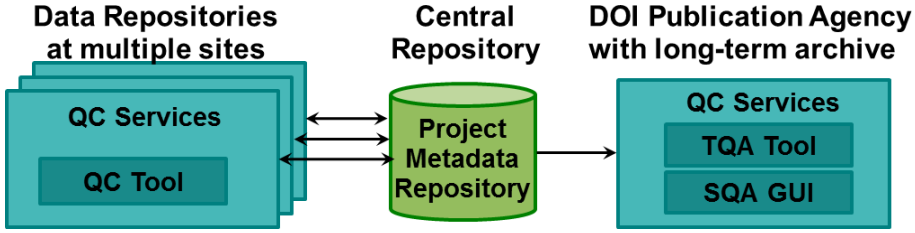

Fig. 1. Infrastructure components of a distributed quality assessment procedure.

GMDD
5, 781–802, 2012

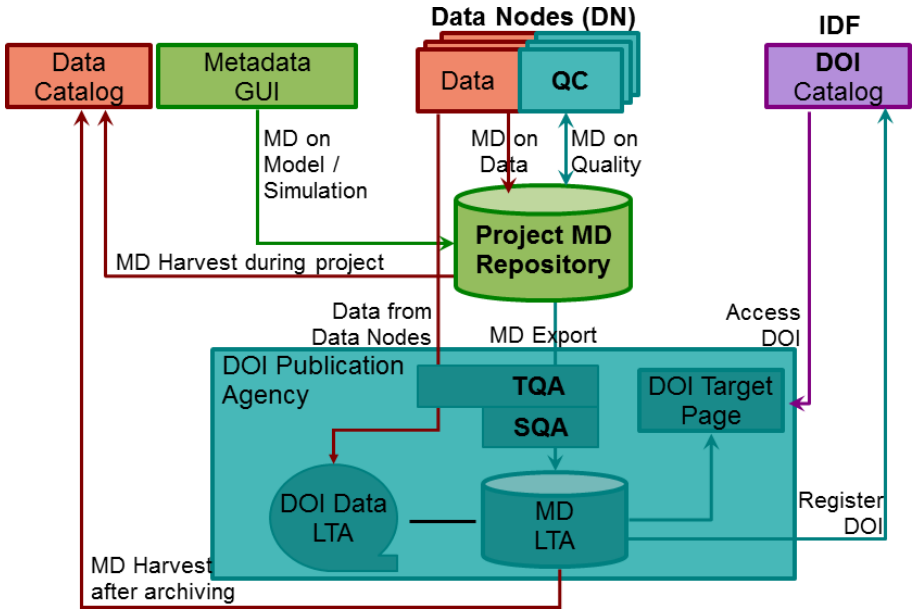

Fig. 2. Distributed Quality Control in a federated data infrastructure (MD: metadata, LTA: long-term archive).

GMDD
5, 781–802, 2012

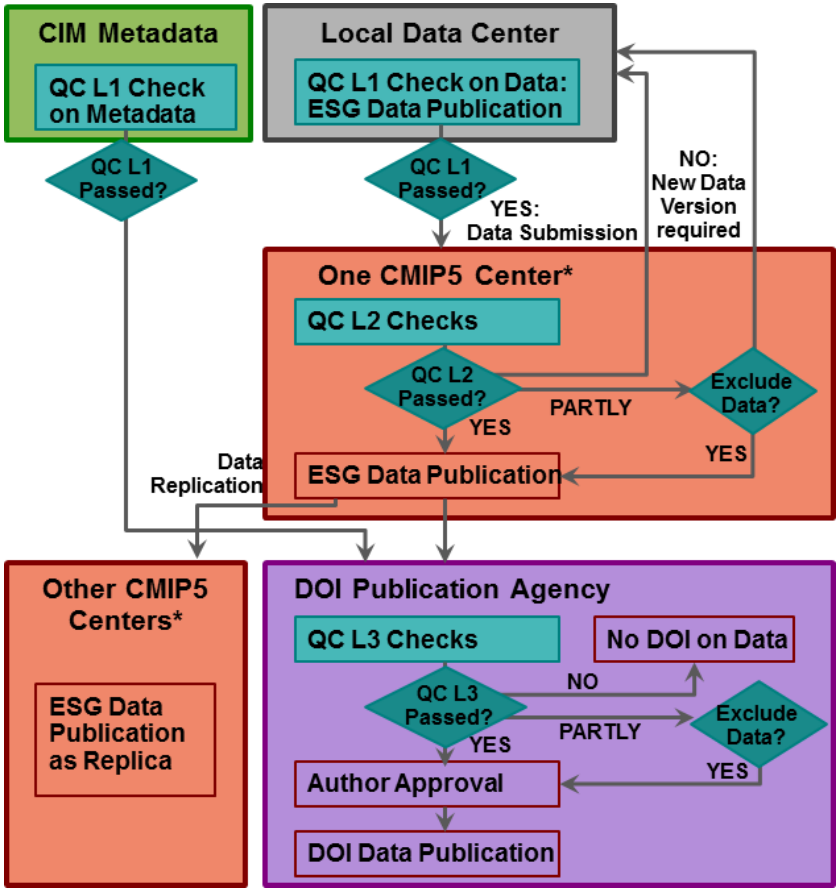

Fig. 3. Workflow of the quality control procedure (*: Current primary CMIP5 data centers are PCMDI, BADC, and DKRZ/WDCC).

GMDD
5, 781–802, 2012

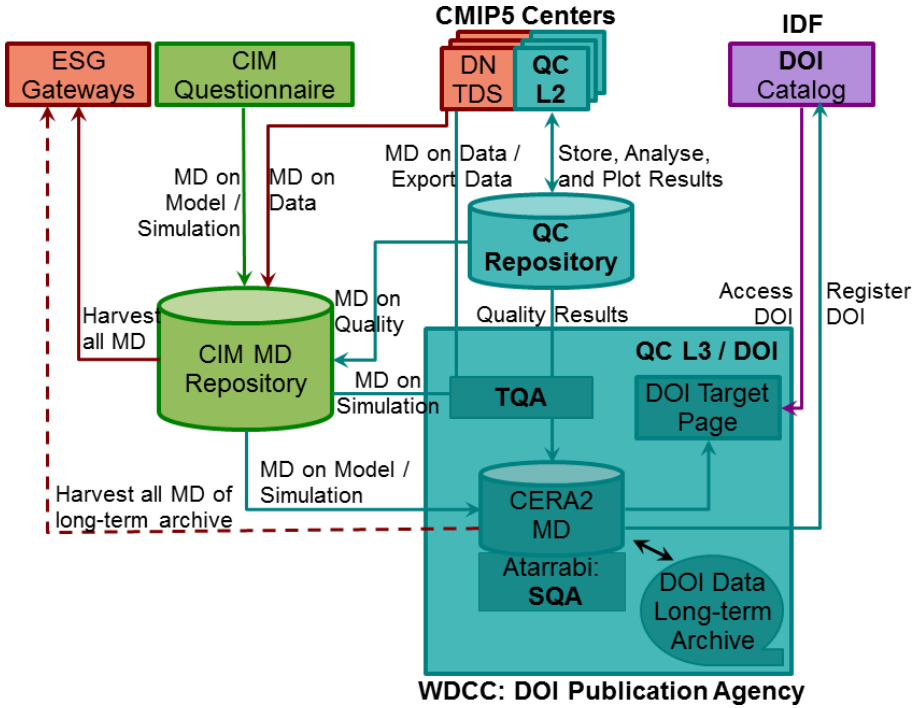

Fig. 4. Implementation of the distributed quality control approach for CMIP5 (MD: metadata, DN: data node, TDS: THREDDS data server).

GMDD
5, 781–802, 2012