Geosci. Model Dev., 5, 1023–1032, 2012
www.geosci-model-dev.net/5/1023/2012/
doi:10.5194/gmd-5-1023-2012
© Author(s) 2012. CC Attribution 3.0 License.

Geoscientific Model Development

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

M. Stockhause[1,2], **H. Höck**[1], **F. Toussaint**[1], **and M. Lautenschlager**[1]

[1]German Climate Computing Center (DKRZ), World Data Center for Climate (WDCC), 20146 Hamburg, Germany
[2]Max-Planck-Institute for Meteorology (MPI-M), 20146 Hamburg, Germany

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

Received: 22 March 2012 – Published in Geosci. Model Dev. Discuss.: 13 April 2012
Revised: 19 July 2012 – Accepted: 24 July 2012 – Published: 13 August 2012

Abstract. The preservation of data in a high state of quality which is 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 Phase 5 (CMIP5). The most important part of the data is to be archived, assigned a DOI, and published 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 workload 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. The distributed quality control approach is capable of identifying data inconsistencies and to make quality results immediately available for data creators, data users and data infrastructure managers. Continuous publication of new data versions and slow data replication prevents the quality control from check completion. This together with ongoing developments of the data and metadata infrastructure requires adaptations in code and concept of the distributed quality control approach.

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 IPCC Data Distribution Centre (IPCC-DDC) has been established to facilitate the timely distribution of a consistent set of up-to-date scenarios of changes in climate and related environmental and socio-economic factors for use in climate impacts assessments. DDC is a shared operation between the British Atmospheric Data Centre (BADC; web pages and data products), the Center for International Earth Science Information Network (CIESIN; socio-economic data), and the World Data Center for Climate (WDCC; global climate model data reference archive).

The results collected within the Climate Model Intercomparison Project Phase 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 non-formalized metadata information, which was imprecise with respect to model and simulation descriptions. The data volume for CMIP5 is expected to reach nearly 100 times that of CMIP3 (Taylor et al., 2012). Over 35 TB of data were collected for CMIP3 (Williams et al., 2008). Estimations for CMIP5 were corrected from early estimates of about 2 PB (Williams et al., 2008) up to 3–3.5 PB (Taylor et al., 2012) for the final CMIP5 data archive.

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, BADC, and WDCC. These centers committed to hold replica of the most important part of the CMIP5 data called *output1*, i.e. IPCC relevant data, on hard disks for quick access and data security.

- Information on models and simulations is enlarged significantly. The underlying metadata schema, the Common Information Model (CIM), was developed by METAFOR. Metadata is collected via a web-based questionnaire (Guilyardi et al., 2011).

- Data curation was improved by introducing a versioning concept and a quality assessment process. The related DataCite data publication provides besides a data citation reference uniform identification of datasets with a persistent identifier DOI (Digital Object Identifier) for data citation in scientific publications. The data citation reference like a citation reference 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 as requirement for DataCite data publications 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). Abbreviations and special expressions are explained in detail in the glossary in Appendix A.

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 data creators exist any longer. However, contents of the quality checks as well as definitions of quality levels and the overall quality assessment 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 repository managers 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) of the data creators has to be accompanied by a technical quality assurance (TQA). The TQA checks data and metadata consistency among the distributed data and metadata repositories, i.e. within the data infrastructure. It might include checks 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 (see e.g. Callaghan et al., 2012). Lawrence et al. (2011) postulate a generic check-list for SQA and TQA issues within a data review procedure. Data and metadata quality aspects are treated separately. In the case of metadata provided along with the data (self-describing data formats) as well as independently by a metadata repository, this metadata information has to be additionally cross-checked for consistency.

Quality control procedures of high volume data should 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).

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 creators. 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 creators' 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 is a registration agency of the IDF (International DOI Foundation). Analogous to the publication of an article in a scientific journal, the data publication makes the data citable and irrevocable. 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 has committed itself to grant persistent data access via the assigned DOI.

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. The degree of quality check distribution is a compromise between granting homogeneous QC application by a small number of QC managers and QC locations, and minimizing data transfer efforts by a high number of QC managers and QC locations. 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, on quality check performance, and on quality results as well as on provenance and on 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 documentation and the communication between DOI Publication Agency and data creator (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.

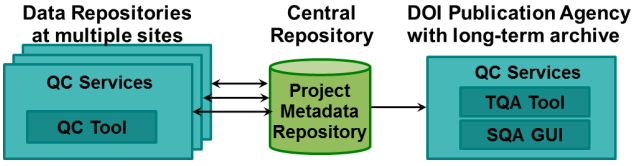

Fig. 1. Infrastructure components of a distributed quality assessment procedure.

The QC service package consists of different services to support:

– 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.

An alternative approach to the described central approach of storing QC results in a central repository is the storage of QC information in the data headers of self-describing data formats such as netCDF. This is not applicable for the described quality procedure, where the QC is performed by data intermediaries and data changes are tasks of the data creators. If the quality information is stored within the local data node along with the data, re-publication of data of a specific version is required for every finished check for QC level 2, increasing the workload of the local data managers significantly. The approach is less flexible compared to the described central approach for QC result storage, because it requires the QC checks to be performed on the original data at every data node. It is not possible to apply the QC on data replica by a selected group of QC managers located at few data nodes.

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

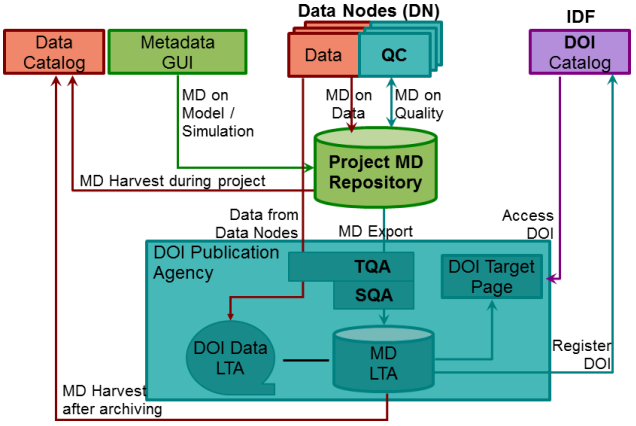

Fig. 2. Distributed quality control in a federated data infrastructure (MD: metadata, LTA: long-term archive).

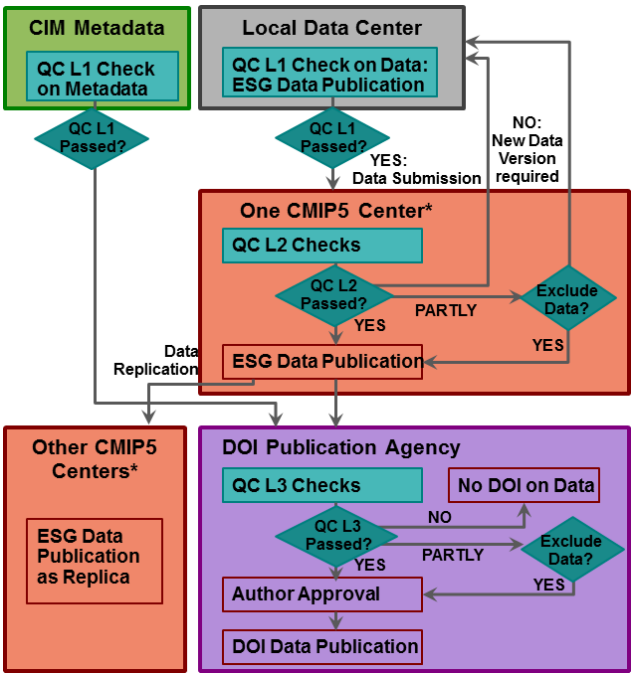

Fig. 3. Workflow of the quality control procedure (*: Current primary CMIP5 data centers are PCMDI, BADC, and DKRZ/WDCC).

on quality of the different QC level checks are added. These metadata are 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 data 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.

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 CMIP5 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. Examples for QC L1 checks are size > 0, correctness of DRS (Data Reference Syntax) identifier components, and monotonic time values. 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. Description of models and simulations are provided by the CMIP5 modeling centers via the CMIP5 questionnaire. During metadata publication the metadata is quality checked in regard to completeness and CIM conformance.

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 criterion is the continuity of the time axis and the usage of the accurate CF (Climate and Forecast) standard name for the variable. Data of QC level 2 is accessible for the CMIP5 research community. Version control of the ESG published datasets enables users to identify data, which they downloaded as latest data version at a certain time before QC procedure completion (DOI data publication).

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*)

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 (TDS), 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. Data access checks at the primary CMIP5 data centers are also part of the TQA. 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 like for CMIP3 and reasons of data security. Data discovery functionality in the gateways was enhanced from the search on essential data information to information on data, model, simulation, and platform (CIM content). 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 (Guilyardi et al., 2011). The CIM repository is meant to provide detailed and reliable long-term metadata information for different

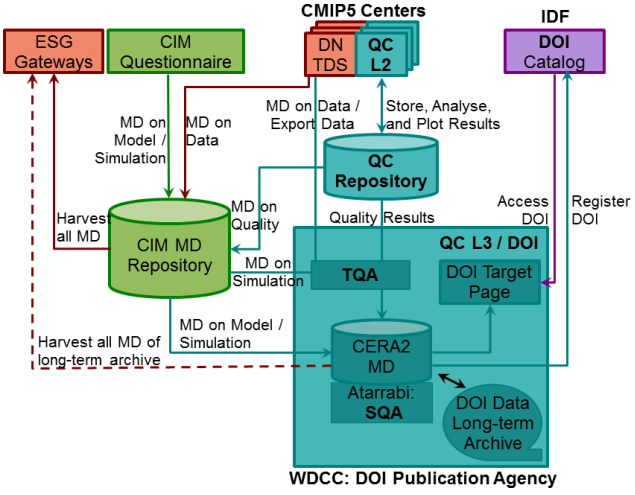

Fig. 4. Implementation of the distributed quality control approach for CMIP5 (MD: metadata, DN: data node, TDS: THREDDS data server).

portals like the ESG portal or the European IS-ENES portal. The CIM metadata format was chosen within CMIP5 as exchange metadata format. The ESG data nodes incorporate a THREDDS Data Server 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). 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.[1] For the final author approval of the data by the data creator, a graphical user interface is used to support 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, 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.

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 along with 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 CMIP5 internal 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 act 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 of serving 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]The CIM metadata repository is still under development.

Administrative concept issue of a deadline for the IPCC-AR5 reports but none for the CMIP5 data submission

Most modeling centers are still in the process of ESG publishing of 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 for several iterations. 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. For example, whole records with filling values 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.

Technical concept issue of data replication

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 managers at PCMDI, BADC or DKRZ remain responsible for the QC L2 assessment and thus the QC level 2 assignment. QC managers can either delegate the QC L2 application to local data node managers or perform the checks themselves. 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, scientists of WG I had to be supported in the citing of CMIP5 data, especially of data of QC levels 1 and 2, i.e. data without DOIs. WDCC set up its citation service for that purpose.

Ongoing technical developments

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 yet integrated in the data discovery functionality, but remain separated pieces of information. To bridge this intermediate invisibility of QC status and DOI, WDCC set up the QC status and result services for data users. These need to be linked from the ESG gateways. The switch from the ESG gateway to the ESGF gateway, which will support different data versions, is scheduled for end of July 2012 at the primary CMIP5 data centers. For an intermediate time period, both data infrastructures will exist within the data federation. The joint international initiative ES-DOC-Models (Earth System Documentation-Models) develops software for a project metadata repository of CIM documents. CIM quality documents will be harvested along with CIM documents resulting from the CIM questionnaire. Necessary adaptations of the QC procedure for CMIP5 have been started.

Application issues at local data nodes including identifiers
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 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) DRS syntax. DRS names for institutions and models are defined by the modeling centers twice, in the CIM questionnaire and in the file directory. Two slightly different controlled vocabularies are used within CIM and in the ESG data nodes. The QC relies on the names used in the data nodes. 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 using the same DRS_id, 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* calculated 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 encountered outdated checksums which were not recalculated and re-published after data changes producing 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. Because of these inconsistencies in the usage of unique identifiers, the DOI publication agency performs a second, slightly simplified TQA cross-check before the data creator starts the final author approval in the SQA GUI. 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 *MD5 checksum* might not be updated during data changes, their usability is restricted. The lack of a central controlled vocabulary of the DRS name components, which is up-to-date and accessible 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 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 co-operation like the CMIP5 infrastructure, the ideal single project metadata repository might not be achievable. However, a central metadata repository for metadata exchange can be established, preferably 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 repository, 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 has introduced these issues into the joint international initiative ES-DOC-Models (Earth System Documentation-Models), which aims to develop metadata services for climate projects. A CIM metadata repository for metadata exchange within CMIP5 is under development.

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. The 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. This CMIP5 quality control procedure has proved its value as it has detected several inconsistencies in the delivered data. 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 in future.

The identification of data inconsistencies within the CMIP5 data and metadata infrastructure by the presented QC procedure, enables the data infrastructure managers to analyze inconsistencies in order to re-establish a consistent distributed CMIP5 long-term data archive.

The roadmap of future developments 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 the integration of the QC result documentation. 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.
- *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 other QC level 2 checker tools, developed and established by 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 WDCC/DKRZ.
- *Peer Review:* Though data is thoroughly reviewed by WDCC's Publication Agents, a peer review process is still missing. A peer reviewed entry in a scientist's list of publications presently requires the additional publication of DOI data in a data journal, e.g. ESSD (Earth System Science Data; http://www.earth-system-science-data.net). The integration of a peer review process into the DOI data publication process in cooperation with peer reviewed journals as well as a closer relation of DOI data and DOI print publications is desirable. Possible concepts are discussed by Lawrence et al. (2011).

Appendix A

Glossary

CIM and METAFOR:	The Common Information Model (CIM) is a metadata schema for the description of numeric models, scientific experiments, simulations, and platforms of Earth System Science data. This metadata schema is used for the collection of information within CMIP5. A questionnaire was developed for CMIP5, in which the modeling centers provide their information. CIM was developed by the European project METAFOR (http://metaforclimate.eu). The current international joint initiative ES-DOC-Models continues the development of CIM services.
CMIP5 and IPCC-AR5:	Under the World Climate Research Programme (WCRP) the Working Group on Coupled Modelling (WGCM) established the Coupled Model Intercomparison Project (CMIP) as a standard experimental protocol for studying the output of Earth System Science models (http://cmip-pcmdi.llnl.gov/). CMIP provides a community-based infrastructure in support of climate model diagnosis, validation, intercomparison, documentation and data access. Phase five of CMIP (CMIP5) includes different climate projections on long-term as well as on decadal future climate changes. CMIP5 data are one source of information underlying the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5). Phase three of CMIP (CMIP3) was of similar importance for the IPCC-AR4. The technical infrastructure for CMIP data access was changed from a central approach for CMIP3 to a federated one for CMIP5. The expected data amount for CMIP5 is estimated to be about 100 times larger than that for CMIP3.
Primary CMIP5 data centers:	The primary CMIP5 data centers are a consortium of currently the three data centers: PCMDI, BADC, and WDCC/DKRZ. They committed to store a replica of all *output1* data on hard disk for quick access, to set up a gateway, and to take part in the CMIP5 quality control checks for level 2 with quality managers. They are also called CMIP5 archive centers or primary CMIP5 portals. QC level 2 checks are additionally performed by NCI (National Computation Infrastructure) for the national Australian CMIP5 data.
CMOR2:	Climate Model Output Rewriter Version 2 (CMOR2) is a software package recommended for the CMIP5 data preparation by the modeling centers. Among other issues CMOR2 writes DRS_id components and tracking_ids (uuids) into the netCDF file headers.

DataCite: The International Data Citation Initiative (DataCite; http://datacite.org) is an international association, which is a member of the IDF undertaking the function of a Registration Agency for the DOI publication of scientific data. The national DataCite members have contracts with national DOI data publication agencies. WDCC/DKRZ is a publication agency of the TIB (Technische Informationsbibliothek/German National Library of Science and Technology). Data Publication Agencies have to grant persistent access to DOI data and are responsible for the necessary data curation. Data is introduced in the long-term archive (LTA) of the publication agency for that purpose. CMIP5 data collections of simulations are assigned DOIs, i.e. data collections of all data belonging to all realizations of a CMIP5 experiment run by a certain modeling center with a certain climate model. The DOI assignment is the final step of a quality control procedure consisting of three quality levels.

DOI and IDF: The Digital Object Identifiers (DOI) are globally unique and persistent identifiers for data or print publications. They are suitable for use in reference lists to cite data or papers. The DOI handle system (http://dx.doi.org) redirects the DOI to a page maintained by the publication agency. These DOI landing pages could also be the sources themselves, like a print publication. In cases of data DOIs on data collections, they are entry pages for data access and data-related information. An example of a DOI landing page for CMIP5 data is DOI:10.1594/WDCC/CMIP5.MXELAM (http://dx.doi.org/10.1594/WDCC/CMIP5.MXELAM). The DOI handle system was developed by the International DOI Foundation (IDF; http://www.doi.org).

DRS: The Data Reference Syntax (DRS) provides identification of CMIP5 data. It has the two dialects: CMOR2 or production DRS without data versioning and ESGF DRS with the added components MIP table and version compared to the CMOR2 DRS (http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf). DRS directory structure is used for data storage; DRS identifiers are important for the technical components of the CMIP5 data infrastructure. CIM uses reduced DRS identifiers with less components. The DRS component *Product* is used to distinguish the more important part of the data *output1*, which is relevant for the IPCC-AR5, from less important data *output2*. The *output1* data is replicated among the primary CMIP5 data centers. Data is published in the ESG in ESG publication units of data collections on the hierarchy level of the DRS component *Ensemble member*. QC levels and DOIs are assigned to larger data collections of DRS *experiments*.

ESG: The Earth System Grid (ESG) is a productive grid infrastructure for global climate model and observation data access (http://www.earthsystemgrid.org). The data infrastructure consists of ESG gateways for data discovery and ESG data nodes for local data access and storage. A THREDDS catalog is part of the ESG data node infrastructure.

ESGF: The Earth System Grid Federation (ESGF) is an international collaboration with a current focus on serving the World Climate Research Programme's (WCRP) Coupled Model Intercomparison Project (CMIP) and supporting climate and environmental science in general. It reflects a broad array of contributions from the collaborating partners. An important part of its current work is the development of a P2P (peer-to-peer) architecture, which supports data replica for CMIP5.

ES-DOC-Models: The joint international initiative ES-DOC-Models (Earth System Documentation-Models) develops software for CIM metadata services (http://earthsystemcog.org/projects/es-doc-models/). It continues the development initiated by the European project METAFOR.

IS-ENES: The European Network for Earth System Modelling (ENES) is a cooperation of European partners working for the development of a network for Earth System Modelling. It includes university departments, research centers, meteorological services, computer centers and industrial partners. One of the aims is to develop an advanced software and hardware environment in Europe, under which high resolution climate models can be developed, improved, and integrated. The infrastructure project IS-ENES (http://is.enes.org/) focuses on that.

NetCDF CF: Network Common Data Form (netCDF; http://www.unidata.ucar.edu/software/netcdf/) is a self-describing binary data format, which is widely used in Earth System Sciences, esp. as data exchange format. Within netCDF files a standard_name attribute can be defined for a variable or coordinate, which should comply with the Climate and Forecast (CF) standard (http://cf-pcmdi.llnl.gov/). The CF conform netCDF format is used for CMIP5 data in netCDF3 classic and CF-1.4 versions.

SQA: The Scientific Quality Assurance (SQA) consists of data content and cross-data content checks. It is performed by the data creator and documented during the CMIP5 QC checks of level 3 during the final author approval step. The communication between data creator and the DOI Publication Agency WDCC is supported by the web application package atarrabi (http://cera-www.dkrz.de/atarrabi2/).

TDS: The THREDDS Data Server (TDS; http://www.unidata.ucar.edu/projects/THREDDS/) is a web server that provides metadata and data access for scientific datasets, using common remote data access protocols, like OPeNDAP, HTTP, and OGC WMS and WCS. It is widely used for geo-referenced data distribution and access. It is part of the data infrastructure which is used by CMIP5.

TQA: The Technical Quality Assurance (TQA) consists of consistency checks, which are not data content checks. TQA checks grant consistency of data and between data and metadata in the long-term archive. The TQA is part of the QC level 3 checks of the CMIP5 quality control procedure.

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.

Edited by: S. Easterbrook

References

Callaghan, S., Lowry, R., and Walton, D.: Data Citation and Publication by NERC's Environmental Data Centers, Ariadne, 68, available at: http://www.ariadne.ac.uk/issue68/callaghan-et-al, last access: 15 May 2012.

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

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

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, 42–46, 2011.

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 Sci. J., 5, 79–83, doi:10.2481/dsj.5.79, 2006.

Lawrence, B., Jones, C., Matthews, B., Pepler, S., and Callaghan, S.: Citation and Peer Review of Data: Moving Towards Formal Data Publication, Int. J. Digi. Cur., 2, 4–37, 2011.

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, 2011.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, B. Am. Meteor. Soc., 93, 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.

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., and Su, M.: Data management and analysis for the Earth System Grid, J. Phys., Conf. Ser., 125, 012072, doi:10.1088/1742-6596/125/1/012072, 2008.

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.

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, 40 pp., 2011.