Discussion Paper:

The CMIP5/AR5 Model Data Quality Control

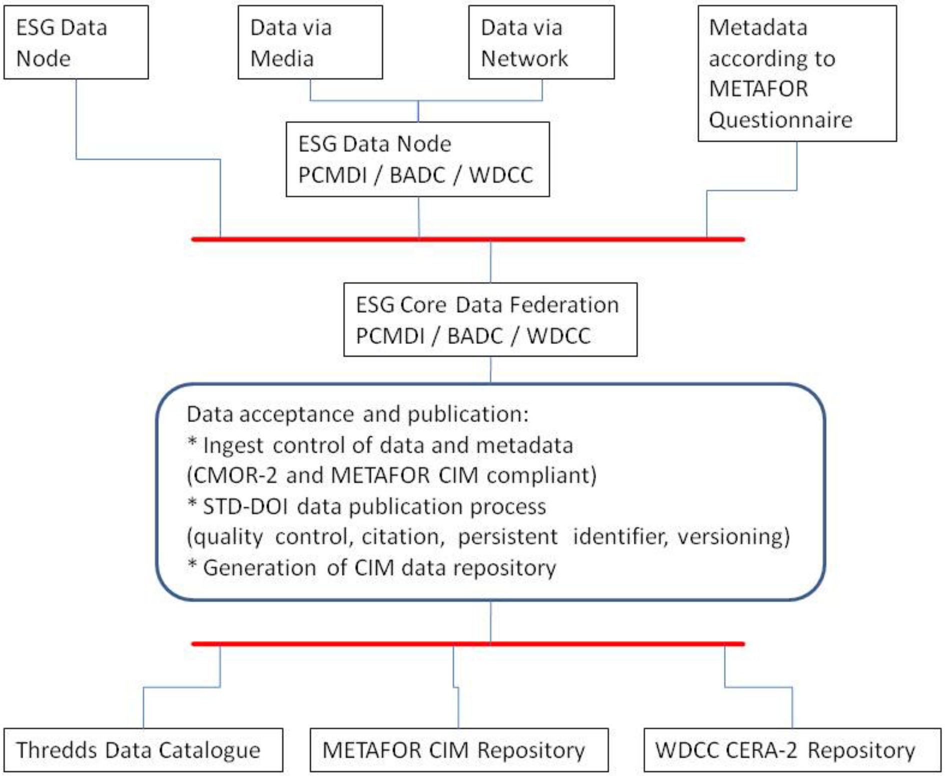
Michael Lautenschlager, Frank Toussaint
with contributions from Bryan Lawrence, Stephan Kindermann

January 11th, 2010

Background

Model output archives such as the IPCC and CMIP archives enable scientists to write papers based on runs done by others and to perform their own scientific research. Beside the definition of a proper method to give credit to the modeling groups while using their data (agreed climate model data citation direction) the responsible data archives have to define and to guarantee a certain level of data quality. This data quality assurance is especially important for climate model data usage in an interdisciplinary context like IPCC WG II and III.

An overall block diagram of the CMIP5 data ingest and publication together with tasks for data acceptance, documentation and quality control is provided in the following graph:


ESG Data Node
Data via Media
Data via Network
Metadata according to METAFOR Questionnaire
ESG Data Node PCMDI / BADC / WDCC
ESG Core Data Federation PCMDI / BADC / WDCC
Data acceptance and publication:
* Ingest control of data and metadata
(CMOR-2 and METAFOR CIM compliant)
* STD-DOI data publication process
(quality control, citation, persistent identifier, versioning)
* Generation of CIM data repository
Thredds Data Catalogue
METAFOR CIM Repository
WDCC CERA-2 Repository

The CMIP5/AR5 data acceptance and publication is mainly related to three activities: ingest control of data and metadata, generation of the CIM data repository and the STD-DOI data publication process. The STD-DOI data publication process is discussed in a parallel document. Central part for data dissemination by the ESG Core Data Federation is quality control. Ideally data quality control for ESG data publication and for STD-DOI data publication is identical.

Quality Control Features

Based on WDCC's experience with IPCC AR4 data, the following data quality checks are currently suggested for the CMIP5/AR5 file archive. A detailed consideration of the CMIP5/AR5 related work and required time on DKRZ's infrastructure has been started. These tests are in addition to the CMIP5 data ingest control which will mainly focus on CMOR-2 and METAFOR CIM compliance insurance.

STD-DOI file testing properties are extracted from the existing regional climate model data review process:

a) File consistency ("WDCC Conformance Checks" in implementation flow chart)
1) a file exists for each variable for the prescribed time step(s) (e.g. 6hourly, daily, monthly)
2) files are not empty and in the end will have the right number of records.
3) the layout of each file is consistent to the model design (gridding, filling values)
4) strictly regular time steps
5) time bounds are consistent to the time interval specified in the file name
6) no overlap of consecutive time bounds

b) Data base property ("Additional Quality Checks" in implementation flow chart)
7) each entry in the data base has a counter part in the file system (and vice versa).
8) specifications in the meta data of the data base correspond exactly to the layout of the files

c) Physical properties of variables ("Additional Quality Checks" in implementation flow chart)
9) minimum and maximum are checked against specified ranges (default for each grid cell: the magnitude of the current weighted global mean plus twice the standard deviation is smaller than a prescribed threshold (10 to the power of 5), where current weighted global mean is the value from the beginning to the current time step.
10) time series are calculated for:

- min
- max
- globally weighted mean
- area weighted mean (reasonable, e.g., for temperature of snow)
- global arithmetic mean
- standard deviation of the globally weighted mean.

11) global frequency distributions for each file for prescribed periods (every three decades)

Implementation of Quality Control

The two flow diagrams from Bryan (with slight modifications)

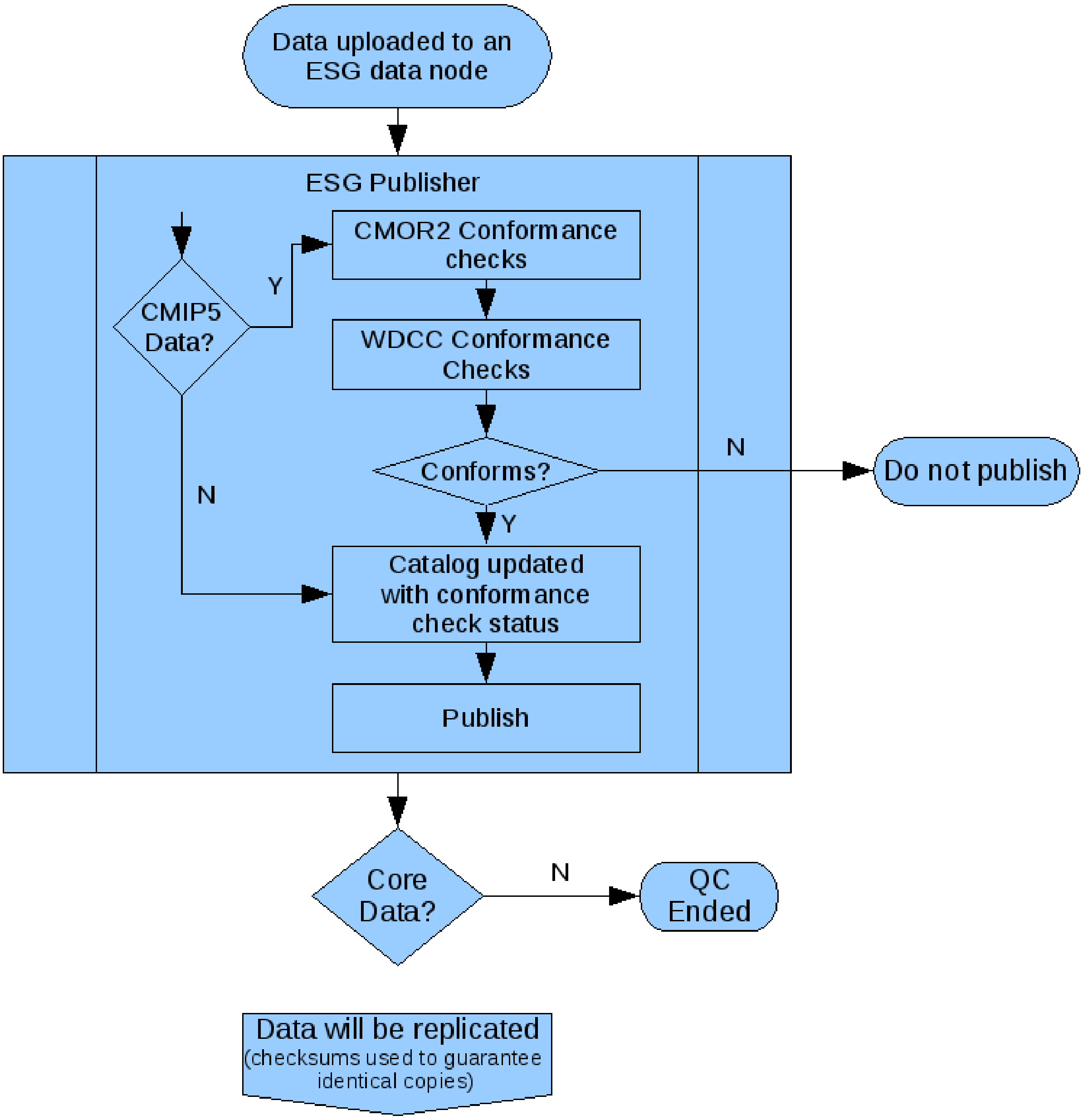
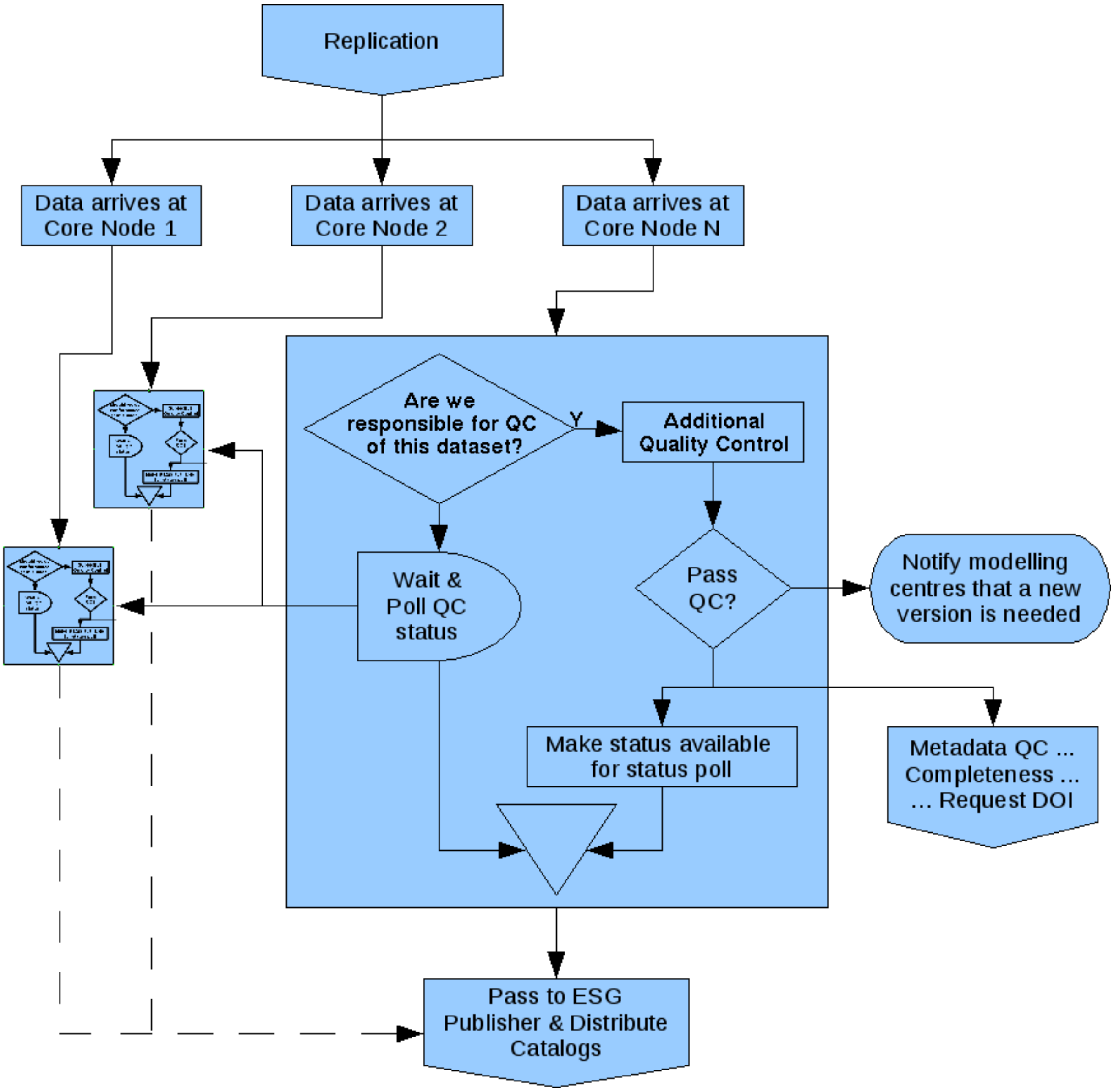


Diagram 1



Replication
Data arrives at Core Node 1
Data arrives at Core Node 2
Data arrives at Core Node N
Are we responsible for QC of this dataset?
Y
Additional Quality Control
Wait & Poll QC status
Pass QC?
Notify modelling centres that a new version is needed
Make status available for status poll
Metadata QC ... Completeness ... ... Request DOI
Pass to ESG Publisher & Distribute Catalogs

Diagram 2