The CMIP5/AR5 Model Data Quality Control

Michael Lautenschlager, Bryan Lawrence, Martina Stockhause,
Frank Toussaint, Stephan Kindermann

August 20th, 2010

1. Introduction

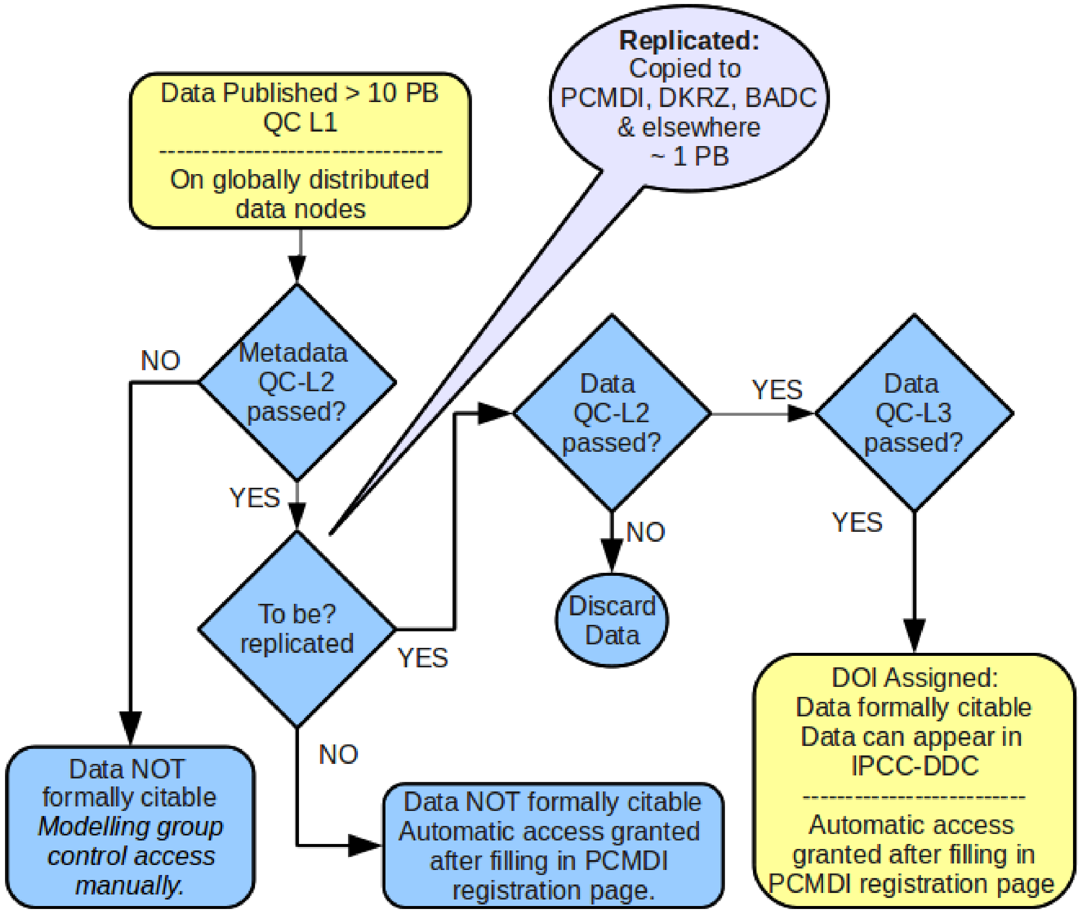
Model output archives such as the IPCC and CMIP archives enable scientists to write papers based on runs done by others and to perform their own scientific research. Beside the definition of a proper method to give credit to the modeling groups while using their data (agreed climate model data citation reference) 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.

The CMIP5 data space can be subdivided into three layers with respect to quality control and access constrains:

a) CMIP5 model data: All model data which are produced for CMIP5 and which are directly published by the modeling groups via ESG data nodes (estimated volume: 10 PB and more).

b) CMIP5 requested data: Climate model data which are requested by PCMDI for model inter-comparison (estimated volume: roughly 2 PB according to CMIP5 variables list, URL: http://cmip-pcmdi.llnl.gov/cmip5/docs/standard_output.pdf)

c) CMIP5 replicated data: IPCC-AR5 relevant data are replicated to the three core data archives (PCMDI, BADC, WDCC) for quality assurance, for data dessimination for the IPCC process and for long-term archiving (estimated volume: roughly 1 PB, variables list (subset of b) as part of the IPCC DDC and to be adjusted with TGICA)



(Informal citation still requested where formal citation not available)

Figure 1: CMIP5 quality control and related access constraints and citation reference

The CMIP5/AR5 data acceptance and publication is mainly related to three activities: ingest control of data and metadata (Quality Control Level 1 – QC L1), additional quality checks for CMIP5 core data and metadata (QC Level 2), and the final versioning and STD-DOI data publication (QC Level 3). The STD-DOI data publication process is discussed in a parallel document. Central part for data dissemination by the ESG Federation Archival Centers is quality control.

The different QC levels are related to an increase in data access ranging from individual modeling groups over IPCC WG I and CMIP5 members to an overall scientific data access with the IPCC AR5 assessment process.

2. Quality Control Levels

The CMIP5/AR5 quality control for core data is performed in three steps resulting each in a separate data quality level. These are described in greater detail in the following subsections and subsumed in tab. 1. For transparency to the users it is important to indicate clearly at the user interfaces the different QC levels of the data and what these levels mean in detail.

Before entering the structured ESGF QC work flow and disseminate massive amounts of data into an ESG CMIP5 data node a preparatory **spot checking** of CMIP5 data files (QC L0) will be offered to modeling groups. This QC L0 offers a test facility like in CMIP3/
IPCC-AR4 for modeling groups to infer their CMIP5 data processing environment before massive data production starts.

QC Level 1 is assigned to the data after passing the data checks described in 1 and 2 in section 2.1. Initial metadata checks of 3 in section 2.1 have to be passed, before starting the checks for QC Level 2.

	QC Level 1: CMOR2, ESG Conformance of Data and CIM Conformance of Metadata	**QC Level 2:** WDCC Conformance and subjective controls	**QC Level3:** STD-DOI Data Publication
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 CMIP5 requested metadata and replicated model data	published and persistent data with version and unique DOI as perrsistent 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
Access Control	PCMDI on the behalf of WMO/WGCM	PCMDI, BADC, WDCC/DKRZ core data archives on the behalf of WMO/WGCM	IPCC-DDC on the behalf of TGICA
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”)

Table 1: CMIP5/AR5 Quality Control Levels and Access Constraints.

2.1 Basic ESGF Conformance Checks (QC L1)

The basic ESGF conformance is checked for all CMIP5/AR5 data during data ingest. It consists of two separate parts: two checks on the data within the ESG data node and one check on the metadata inside the METAFOR questionnaire (fig. 2). The quality checks of level 1 are passed with the publication of data by the ESG publisher or with the saving of the metadata in the questionnaire, respectively. QC L1 is assigned after passing the data checks. CIM metadata from the METAFOR questionnaire may be incomplete. Completion is required for entering QC L2. Checks for QC L2 start after passing all QC L1 checks – for data and metadata.

1. CMOR2 Conformance Checks:

a) DRS/Filename:
   - name matches the profile:
     varid_tableid_modelid_exptid_rid[iid][pid][_startdate-enddate][_suffix][_clim].nc
   - file path matches DRS requirements

b) Global Attributes: check for validity of required global attributes
   - experiment_id, experiment, project_id match tables
   - parent_experiment_id is valid and different from experiment_id
   - forcing, frequency, realization, branch_time are valid
   - creation_date is valid and in right format

c) Axis:
   - checks for axes names validity
   - dimensions ordering
   - checks for required attributes, needed bounds, needed formula_terms, range validity, units (with udunits), type, direction stored, requested_values exist
   - singleton dimensions are defined via coordinate attribute, not actual dimension
   - singleton dimension value validity
   - validity of formula terms
   - for time axes, is in days since
   - for time axes, bounds seem ok
   - for lon/lat axes, checks the grid is an abstract rectangular grid

d) Variable:
   - name is valid
   - file contains only 1 variable
   - checks for optional/required attributes and validity (e.g. CF 'standard_name'), optional/required additional attributes, associated_files defined correctly, type, units (with udunits)

e) cross-checks:
   - variable name indicated by file is in the file
   - file name matches what file says for:
     table_id, model_id, exp_id, physics_version, initialization_method, realization, climatology, start and end times, frequency
   - versions of CMOR and CMOR tables are consistent; table date matches file table date

f) warnings in CMOR2 for:
   - size ≥ 2 GB
   - optional attributes are not defined
   - global attributes are neither required nor optional

2. ESG Conformance Checks:

- File is readable ('online' data)
- File format is recognized
- File is of size > 0 bytes
- Discovery data - especially DRS fields - are identifiable and have correct values. If any mandatory fields are missing or invalid, an error is raised and the data cannot be published.
- CF Standard names are valid. A warning is issued if the standard name is missing or unrecognized.
- Coordinate axes are recognizable (definition based on CF conventions) - particularly time. A calendar is defined.
- Time values are monotonic and do not overlap between files. If overlap is discovered, a warning is logged. This is checked when aggregations are generated. It is not considered an error if timepoints are missing.

3. CIM Conformance Checks (in parallel to NetCDF/CF data checks):

- Mandatory fields checked for completeness; technical validation of CIM-XML.

2.2 WDCC Conformance Checks and Subjective Quality Control (QC L2)

Based on the experience of the WDC Climate (WDCC) with IPCC AR4 data from regional climate model downscaling, the following data quality checks are currently planned for the CMIP5/AR5 core data. Additionally, consistency between data files and the CIM metadata repository is ensured. These quality checks fulfill most of the testing properties for the STD-DOI data publication review process (fig. 3).

a) File consistency

1) in the end files will have the right number of records. The number is given in the metadata.
2) strictly regular time steps (ESG checker allows for time gaps)

b) Metadata consistency (check of consistency between metadata of the standard_output table and the metadata of the file headers)

c) Physical properties of variables

3) minimum and maximum are checked against specified ranges (default for an invalid current exterm value of a global field:

mean:
mextr(t) = 1 / N * sum_{i=0}^{N=t / (delta t) - 1} extr_{i}
N:= index over all time steps 'delta t' up to the actual time t

Default (case for all variables in the standard_output table of K. Taylor):
mextr_{N-1} - extr_{N} > order_of_magnitude(mextr_{N-1} * 10^{5}

4) time series are calculated for:

- min
- max
- globally weighted mean (in case of no _FillValue)
- area weighted mean (in case of existing _FillValue; reasonable, e.g., for temperature of snow)
- standard deviation of the globally weighted mean.

A consideration of the CMIP5/AR5 related work and required time on DKRZ's infrastructure has been accomplished. Based on the observed times on a desktop PC, the times required on the HPC IBM Power6 were estimated, conservatively:

Desktop PC: 50 min per atomic dataset (6hourly interval storage)

IBM Power6 – 1 node (ca. 100 times the performance of a Desktop PC):
0.5 min per atomic dataset (6hourly interval storage),

500 days for all 1.5 Mio. atomic datasets.

The WDCC Conformance checks are completed by subjective quality controls of data and metadata. A logfile of the quality checks for level 2 is stored in the metadata repository for further use in QC L3 and in the data publication process.
After reaching QC L2 the data is accessible for non-commercial research and educational purposes.

2.3 STD-DOI Data Publication Process (QC L3)

The results of the quality checks of level 2 are directly used as testing criteria for the STD-DOI data publication review process of the WDCC (fig. 4). The most essential part in the data publication process is the communication with the data authors and their approval of metadata and model data. For STD-DOI data publication the data review process is finalized by:

1) Double checks of QC L2 based on log files; discussion and clarification with corresponding data author if necessary.

2) Creation of STD-DOI metadata and assignment of persistent identifiers (DOI / URN) for each experiment / simulation.

3) Data author approval to freeze the data entity in its present version; and update the quality flag to "approved by author".

4) Integration of STD-DOI metadata and persistent identifiers for the frozen version of the data entity into the TIBORDER library catalogue (German National Library of Science and Technology, a member of the DataCite[1] organization) and additional catalogues if requested.

5) Notification of corresponding data author and ESGF about the finalization of the data publication process. A short protocol of the quality checks for level 3 is stored in the metadata repository.

[1] http://www.datacite.org

At the end of the STD-DOI publication process the data entity is accessible for unrestricted use. The STD-DOI data publication process is discussed in detail in a parallel document (Lautenschlager et al., 2010). CMIP5 data are published as independent data entities with citation references for use in scientific literature and with DOIs for transparent data access.

3. Implementation of Quality Control

If we consider the Quality Control in the overall CMIP5/AR5 data ingestion and publication process (fig. 5), we recognize the following phases:

1. At all ESG data nodes the QC Level 1 checks (CMOR2 and ESG conformance) are carried out for all CMIP5/AR5 data from the modeling centers. Log files of the checks are available at the ESG data nodes. The ESG portal is notified and the QC Flag “automated conformance checks passed” will be visible at the user data access interfaces.

2. CMIP5 requested data with QC Level 1 are extracted by those core data centers, which are responsible for the QC L2 for these specific data entities.

   Due to existing network band width the initially data will probably be distributed to the Core Data Nodes by shipping disks. Specific parts of the requested data are send to one specific core node each, where the QC L2 Conformance checks are performed. After the finalization of the QC L2 checks the data are replicated among the Core Data Nodes. Therefore during the performance of the QC L2 checks the data at the Core Data Nodes might not be identical. Users have to know who is responsible for QC L2 for certain data and where the most reliable data are stored.

   The model of sharing responsibilities between the three Core Data Nodes could be that PCMDI has the lead in QC L0 and L1 (data ingestion part, excluding the metadata QC L1 checks) and that BADC and WDCC share the work for QC L2. Additionally, BADC is responsible for the QC L1 metadata checks within the METAFOR questionnaire and maintains the CIM data repository while WDCC performs QC L3 and scientific data publication including maintenance of the data-DOI services.

3. At the QC L2 Core Data Nodes the WDCC conformance checks for QC Level 2 are carried out. Log files are entered into the METAFOR repository. After finalization of QC L2 the ESG portal is notified and the QC Flag “subjective quality control passed” will be visible to the users. Access will be given for non-commercial and educational purposes.

4. If the data failed the QC L2 tests or open questions arose from the subjective quality checks, the modeling center is notified. Updates of data and/or metadata will be done by replacement or modification of the existing data. At this stage of the quality control process corrected data starts the QC process again at step 1 with a new version. Old versions of data are normally not archived.

5. Data with assigned QC L2 is replicated across the Core Data Nodes and documented in the CIM metadata repository at BADC.

6. Replicated data with QC Level 2 is passed to the STD-DOI publication process (including final checks for QC L3) at the WDC Climate. A target URL is created which contains beside other information the preliminary citation direction.

7. If the data failed to reach the QC L3 or open questions aroused from subjective quality checks, the modeling center is notified. Updates of data or metadata can be done by replacement or modification at this stage of the quality control process. Completely new data starts the submission and QC processes again at step 1.

8. Replicated data with QC Level 3 and the final approval by the author are assigned persistent identifiers (DOI / URN), which point to a distinct and fixed list of data file versions. The preliminary citation reference is converted into the final citation reference of the STD-DOI and initially published into the TIBORDER library catalogue. The ESG portal is notified and the QC Flag "approved by author" will be visible. Data are no longer matter of change. They are accessible without restrictions. The STD-DOI reference appears in the gateway at the granularity level of the STD-DOI publication (model simulation).

9. For changes in replicated data or data replacements after STD-DOI publication, the whole QC processes has to be carried out for these data again (steps 1 to 7) but the old data version is still available under the assigned DOI and will not be deleted. In case of minor changes an erratum can be added to the STD-DOI metadata. For major changes a new version has to be processed and a new DOI has to be assigned. The ESG portal is notified and the QC Flag for the outdated data version is set to "approved by author, but suspended" with corresponding rationale in the quality documentation.

3.1 Components

- STD_DOI URL page with compact information and provided links to information (CIM at BADC) and data (BADC and WDCC as DOI publication agent) services
- Information service (CIM):
  - Provide information about a DOI / a DRS experiment / a metafor simulation
  - Provide information about the metafor simulation to which TDS data belongs; thredds2cim tool for adding data objects to a simulation
- DOI Data service (DOI-WDCC):
  - Provide a list of TDS links for the replicated data belonging to a DOI
  - Provide a DOI for a CMIP5 file given its tracking_id or its DRS name as input

3.2. Communication

- QC results and the QC flag are send to the CIM repository from where selected information is harvested into the PCMDI portal. Doi2cim tool for adding and changing metadata in CIM after QC L2 and after QC L3 are reached.
- PCMDI portal gets QC flag information from CIM repository and sets the access constraints for the data
- Update of contact on DOI target page if changed during author approval

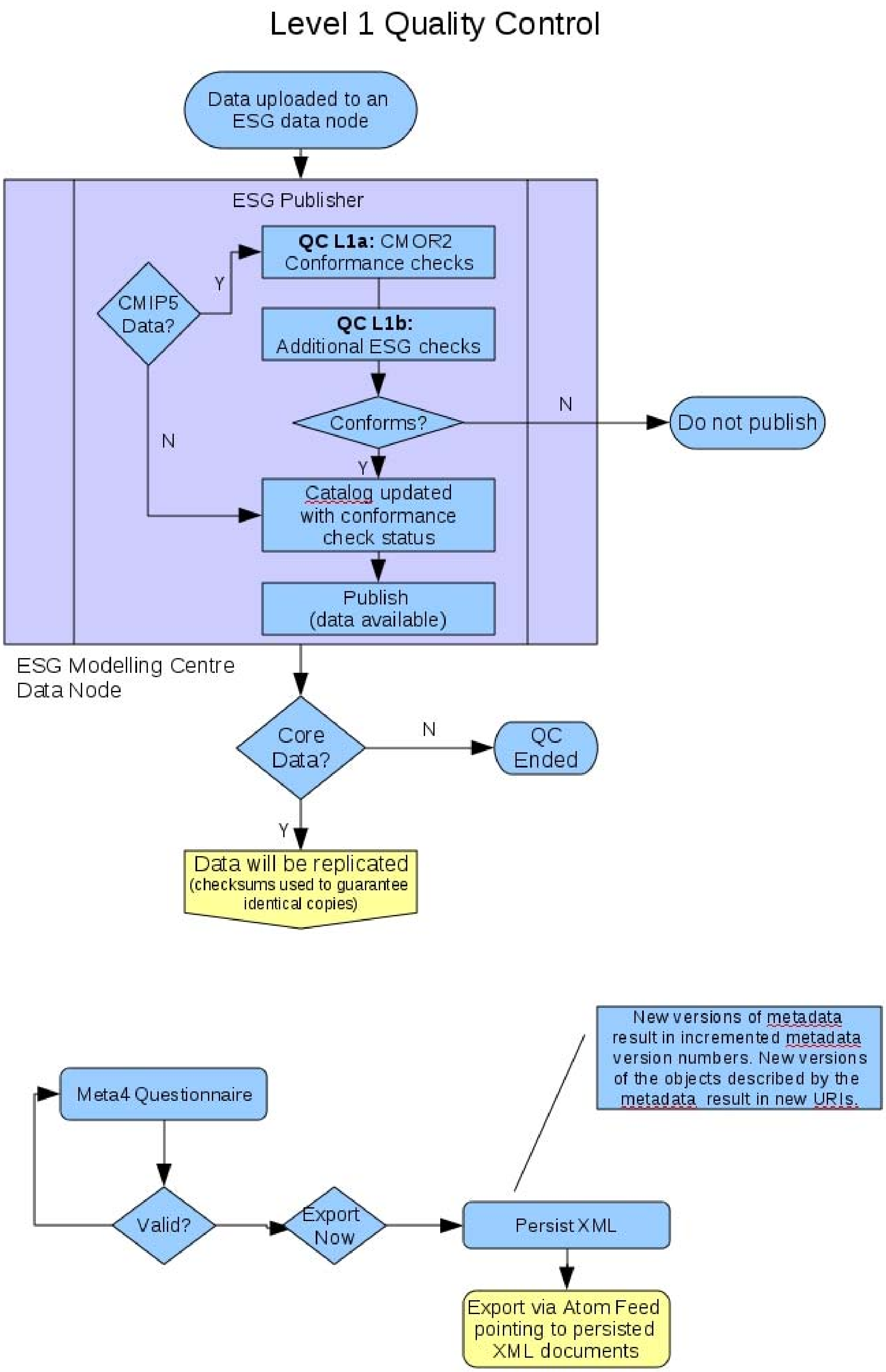

Level 1 Quality Control
Data uploaded to an ESG data node
ESG Publisher
QC L1a: CMOR2 Conformance checks
CMIP5 Data?
Y
QC L1b: Additional ESG checks
Conforms?
N
Do not publish
N
Y
Catalog updated with conformance check status
Publish (data available)
ESG Modelling Centre Data Node
Core Data?
N
QC Ended
Y
Data will be replicated (checksums used to guarantee identical copies)
New versions of metadata result in incremented metadata version numbers. New versions of the objects described by the metadata result in new URIs.
Meta4 Questionnaire
Valid?
Export Now
Persist XML
Export via Atom Feed pointing to persisted XML documents

Figure 2: QC Level 1 ESG/CMOR2 and Metafor Conformance Checks for all data and metadata.

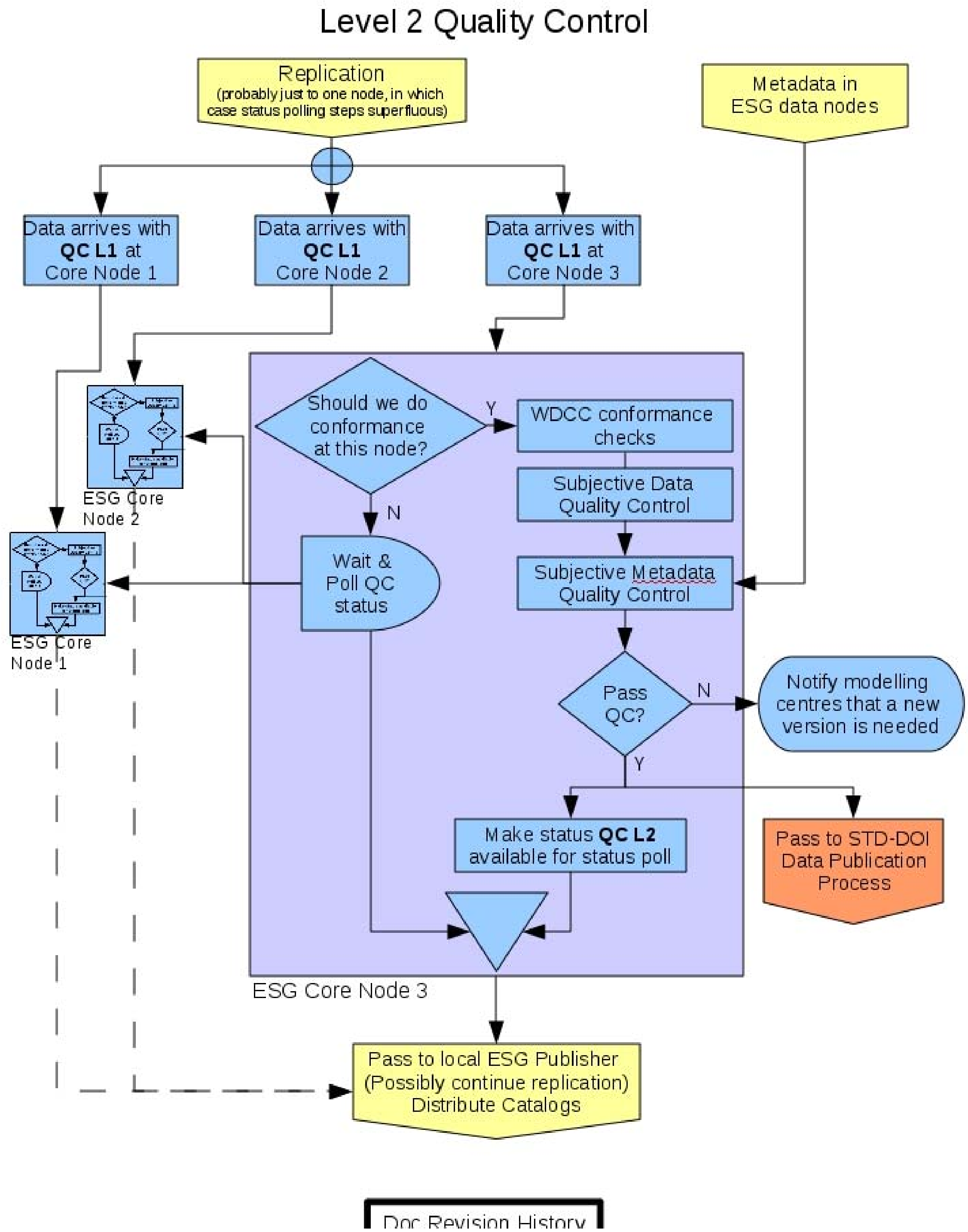

Level 2 Quality Control
Replication
(probably just to one node, in which case status polling steps superfluous)
Metadata in ESG data nodes
Data arrives with **QC L1** at Core Node 1
Data arrives with **QC L1** Core Node 2
Data arrives with **QC L1** at Core Node 3
Should we do conformance at this node?
Y
WDCC conformance checks
Subjective Data Quality Control
N
Subjective Metadata Quality Control
Wait & Poll QC status
ESG Core Node 2
ESG Core Node 1
Pass QC?
N
Notify modelling centres that a new version is needed
Y
Make status **QC L2** available for status poll
Pass to STD-DOI Data Publication Process
ESG Core Node 3
Pass to local ESG Publisher (Possibly continue replication) Distribute Catalogs
Doc Revision History

Figure 3: QC Level 2 - WDCC Conformance Checks for CMIP5 Core Data.

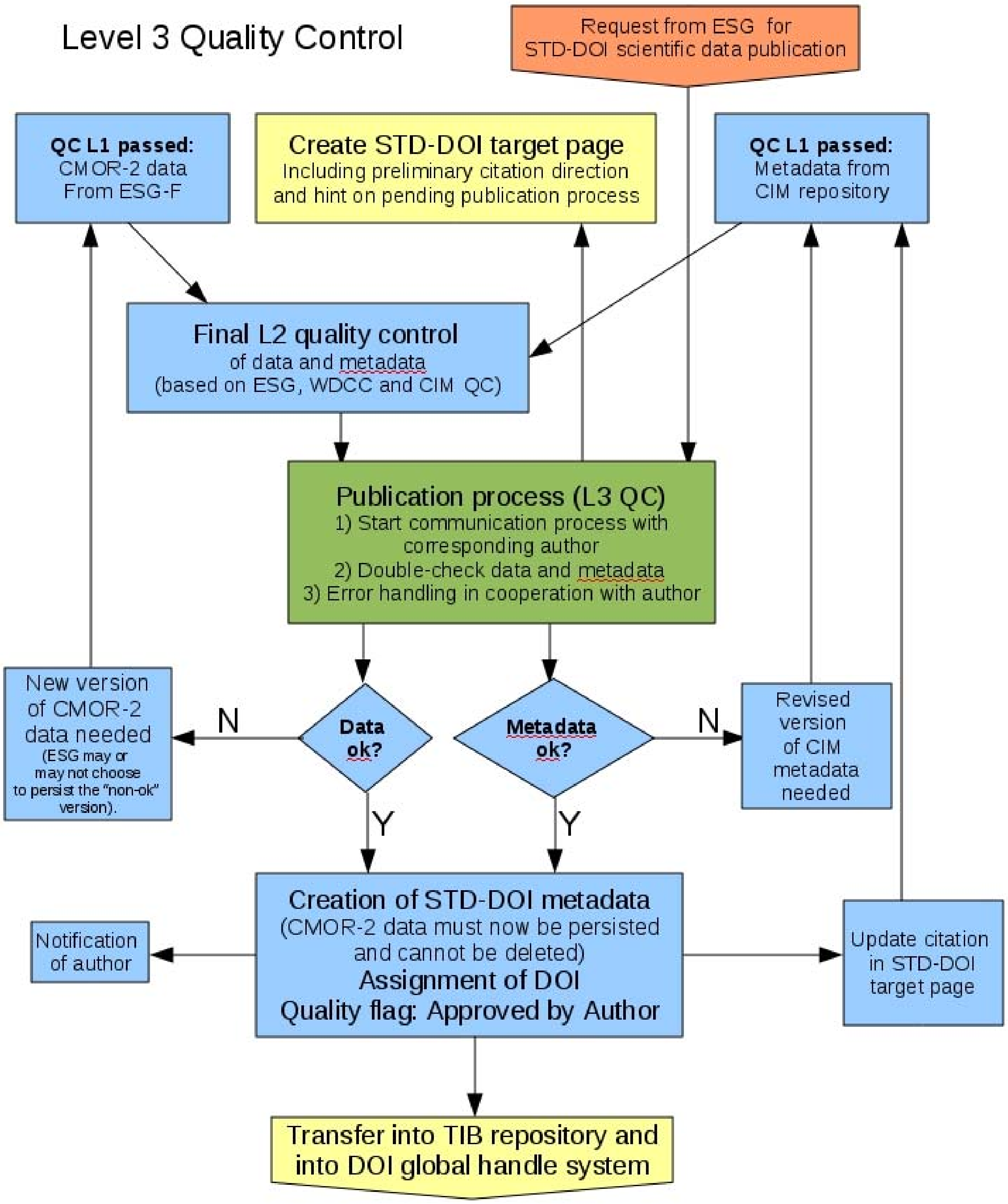

Level 3 Quality Control
Request from ESG for STD-DOI scientific data publication
QC L1 passed: CMOR-2 data From ESG-F
Create STD-DOI target page
Including preliminary citation direction and hint on pending publication process
QC L1 passed: Metadata from CIM repository
Final L2 quality control
of data and metadata
(based on ESG, WDCC and CIM QC)
Publication process (L3 QC)
1) Start communication process with corresponding author
2) Double-check data and metadata
3) Error handling in cooperation with author
New version of CMOR-2 data needed
(ESG may or may not choose to persist the "non-ok" version).
N
Data ok?
Metadata ok?
N
Revised version of CIM metadata needed
Y
Y
Notification of author
Creation of STD-DOI metadata
(CMOR-2 data must now be persisted and cannot be deleted)
Assignment of DOI
Quality flag: Approved by Author
Update citation in STD-DOI target page
Transfer into TIB repository and into DOI global handle system

Figure 4: Final STD-DOI Data Publication Process for CMIP5 Core Data.

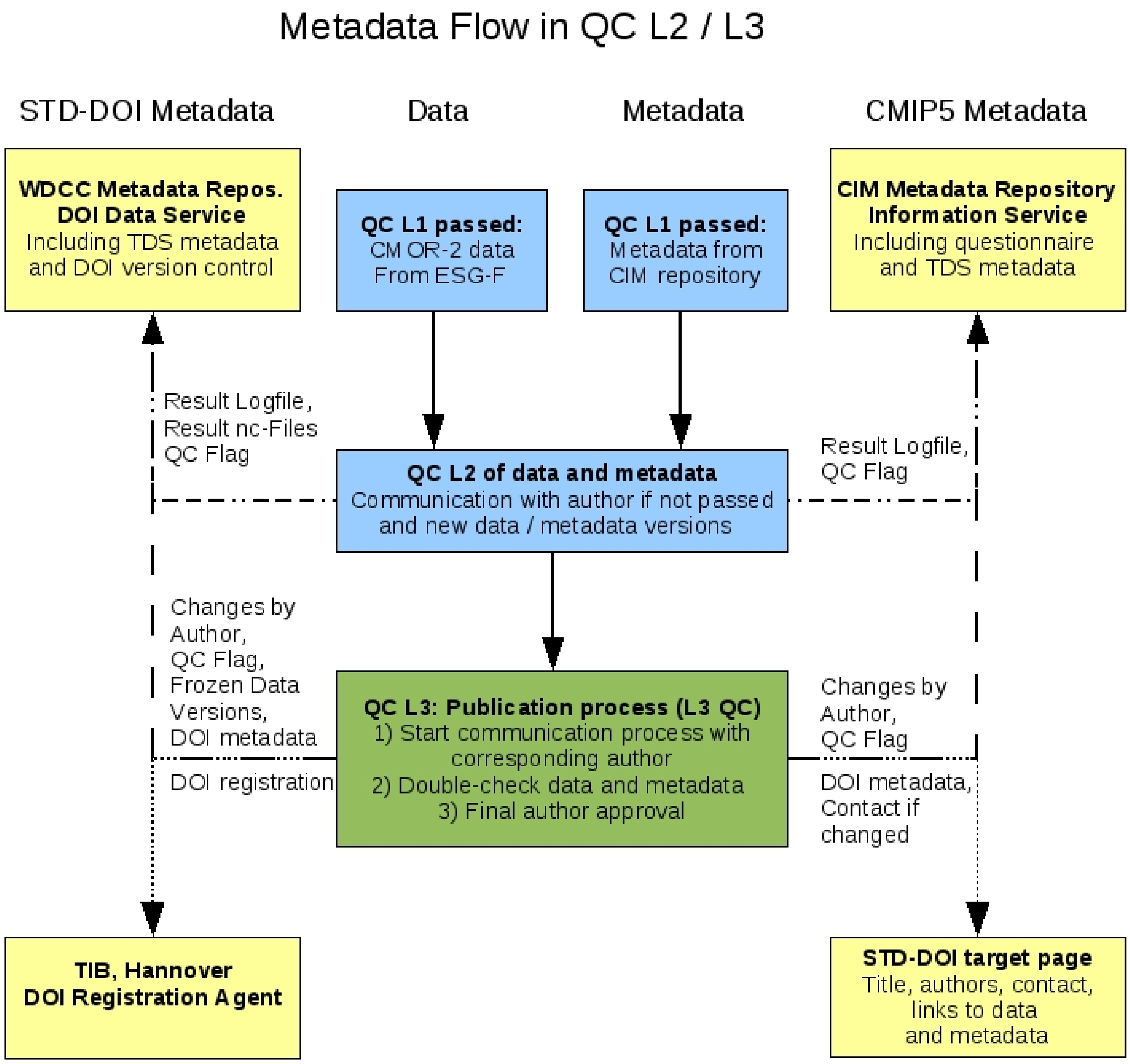


Figure 5: Metadata Exchange within CMIP5 during QC L2/L3 Checks