Questions, open issues, and points of discussion in red

Discussion Paper:

The CMIP5/AR5 Model Data Quality Control

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

April 7th, 2010

Summary

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.

An overall block diagram of the CMIP5 data ingest and publication process in the ESG Federation (ESGF) together with tasks for data acceptance, documentation and quality control is provided in fig. 1.

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.

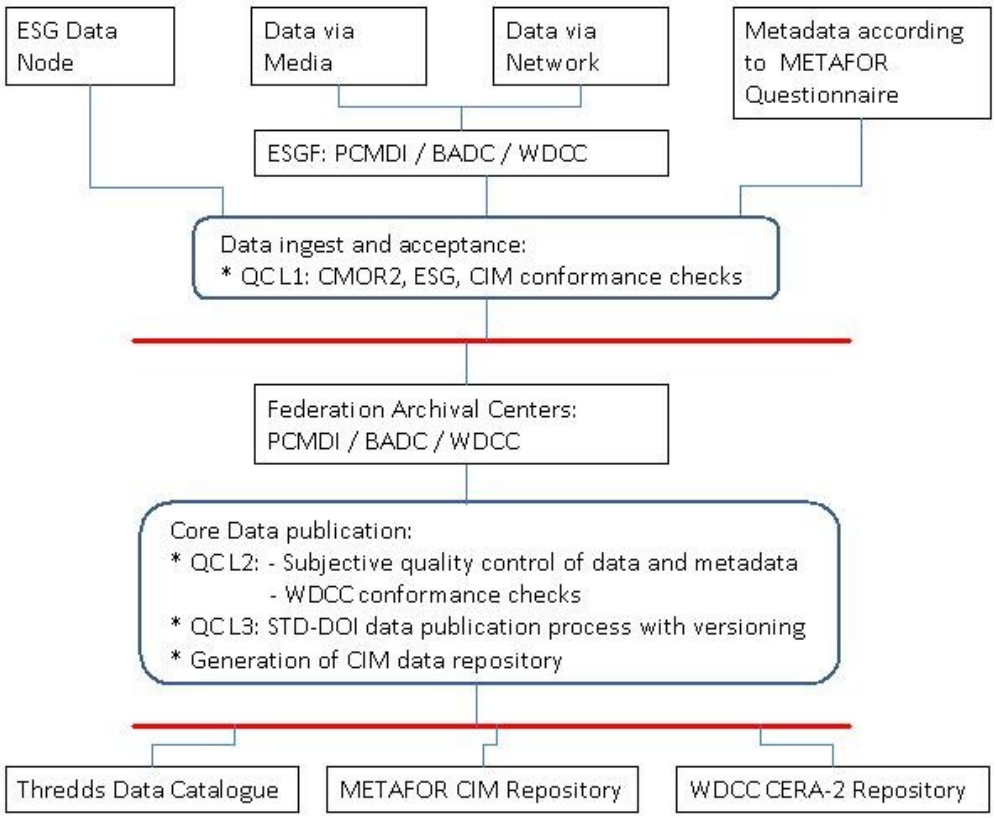

Figure 1: CMIP5/AR5 data ingest and publication process.

	QC Level 1: CMOR2, ESG, CIM Conformance	**QC Level 2:** WDCC Conformance and subjective controls	**QC Level3:** STD-DOI Publication
Data	preliminary; no user notification about changes; performed for all data	preliminary; no user notification about changes; performed for core data	published and persistent data with version and unique STD-DOI citation; performed for core data
Access	constrained to data author (modeling center)	constrained to CMIP5 members and IPCC WG I community	open for IPCC process (WG I – III) and research community
Access Control	PCMDI on the behalf of WMO/WGCM	PCMDI on the behalf of WMO/WGCM	IPCC-DDC on the behalf of TGICA
Citation	no citation reference	preliminary citation reference	final 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.

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 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.

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 three separate checks. The first two are performed within the ESG data node and the third within 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.

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 - *Flagged?*
- 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 (data flagged?). 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 suggested for the CMIP5/AR5 core data. These quality checks fulfill most of the testing properties for the STD-DOI data publication review process (fig. 3).

a) File consistency

1) a file exists for each variable for the prescribed time step(s) (e.g. 6hourly, daily, monthly)
2) in the end files will have the right number of records. The number is given in the metadata.
3) strictly regular time steps (ESG checker allows for time gaps)

b) Data base property (check of consistency between metadata and data files)

4) each entry in the data base has a counter part in the file system (and vice versa).

5) specifications in the meta data of the data base correspond exactly to the layout of the files

c) Physical properties of variables

6) 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.

7) 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.

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.

2.3 STD-DOI Data Publication Process (QC L3)

The results of the quality checks of level 1 and 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 L1 and 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, Hannover).

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

At the end of the STD-DOI publication process the data entity is accessible within the IPCC AR5 process (WG I – III) and within the wider research community. The STD-DOI data publication process is discussed in detail in a parallel document (Lautenschlager et al., 2010).

3. Implementation of Quality Control

If we consider the Quality Control in the overall CMIP5/AR5 data ingestion and publication process, 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. 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.

   At this point different strategies for data replication between the three core data archives are discussed. The strategies replication depend on technical boundary conditions (network bandwidth sufficient for initial replication or initial replication by shipping disks) and the individual role each of the three core data archives in the federation.

   - ***Initial replication among all Core Data Nodes****:* All requested data which are published in the ESG are available at each of the three core data nodes and all core nodes host the same data. During the course of quality control replication will probably be necessary multiple times, which would result in high network loads. The core nodes have to agree on how the QC L2 Conformance checks are shared within the ESGF. In this complete replication scenario modeling groups can easily incorporated into the quality assurance process because of the existing multiple copies. But data synchronization is an issue with respect to logistics and network bandwidth.

   - ***Initial distribution to a single Core Data Node****:* Specific parts of the requested data are send to one specific core node each, where the QC L2 Conformance checks are performed. After finalization of QC L2 the data are replicated among the Core Data Nodes. This replication model saves network bandwidth and can be more easily

performed by shipping discs. But the integration of modeling into the quality assurance process is more complicated as in the initial replication model because data are spread across the three Core Data Nodes and copies might not be identical before finalization of QC L2. Users have to know who is responsible for QC L2 for certain data and where the most reliable data are stored.

One 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) and that BADC and WDCC share the work for QC L2. Additionally BADC maintains the CIM data repository while WDCC performs QC L3 and scientific data publication. (Discussion with Karl T. March 30th, 2010 in Hamburg).

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.

4. If the data failed the QC L2 tests or open questions aroused 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 not archived.

5. Replicated data with QC Level 2 is passed to the STD-DOI publication process (including final checks for QC L3) at the WDC Climate and replicated to the other core data centers if yet not done. A target URL is created which contains beside other information the preliminary citation direction.

6. If the data failed to reach the QC L3 or open questions aroused from the 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. New data starts the submission and QC processes again at step 1.

7. Replicated data with QC Level 3 and the final approval by the author are assigned persistent identifiers (DOI / URN) and a fixed ESG data version. The preliminary citation direction is converted into the final citation direction of the STD-DOI and 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. The STD-DOI reference appears in the gateway at the granularity level of the STD-DOI publication (model simulation).

8. For replicated data changes or 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 not be deleted. In case of minor changes an erratum can be added to the STD-DOI metadata. For mayor 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".

In the above described QC workflow some components are still missing. And some communication channels haven't been established, so far.

3.1 Missing Components

- Portal extension to show QC Flag and therefore user availability of data.
- Completion of CIM / Questionnaire metadata by data descriptions from TDS and postgres database of ESG publisher.

3.2. Missing Communication

- QC Flag and extended CIM metadata (plus data descriptions from TDS and QC Flag) synchronization with METAFOR repository and ESG portal.
- Exchange of QC log files (levels 2 – 3), level 1 will be available at the specific ESG data node.
- Agreement within ESGF on sharing responsibilities (see discussion on replication)

REFERENCES

Lautenschlager, M., V. Balaji, B. Lawrence (2010): Proposal: Scholarly citations for CMIP5 model output, Draft March 11th, 2010.

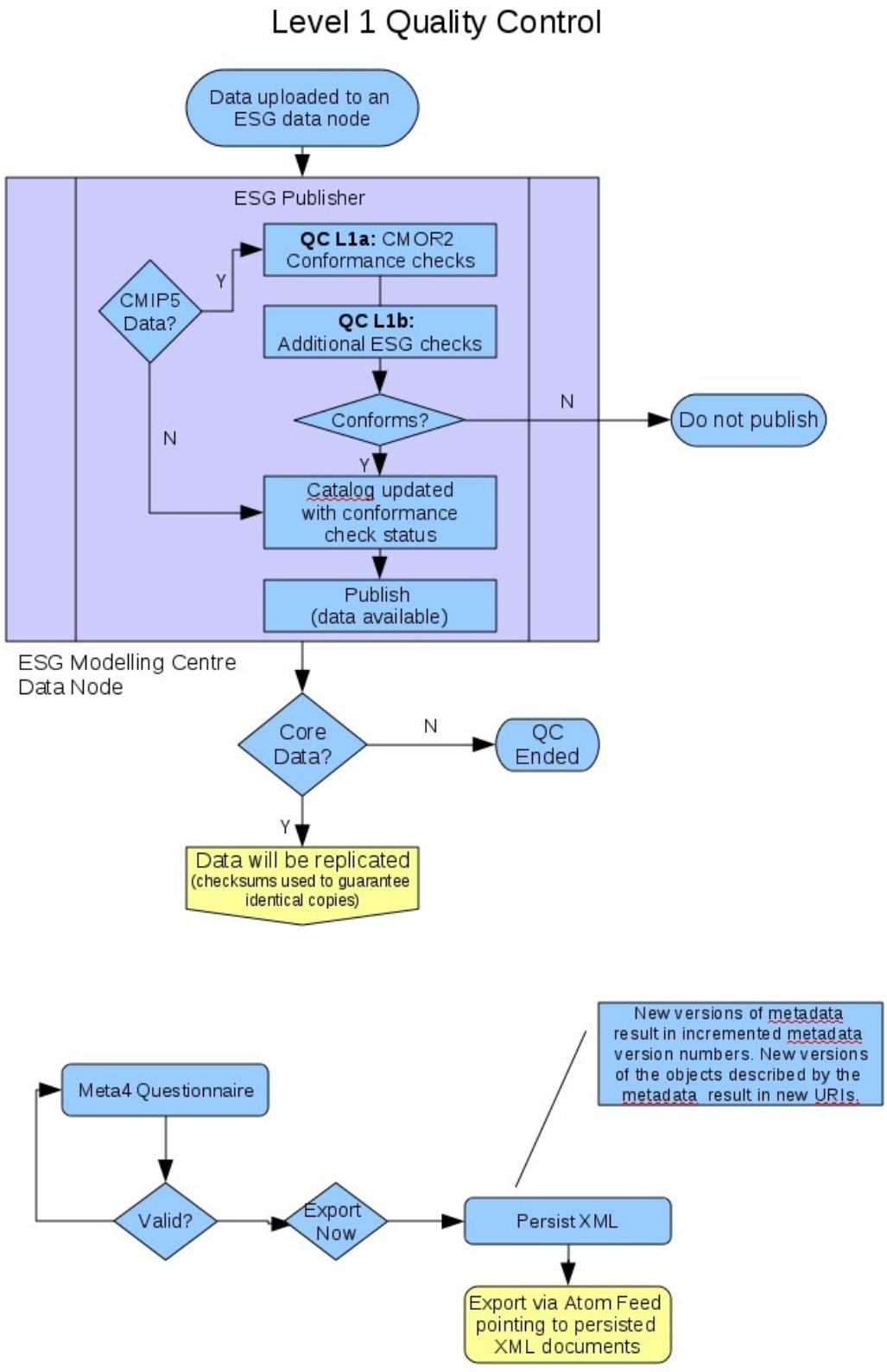

Figure 2: QC Level 1 ESG/CMOR2 and Metafor Conformance Checks for all data and metadata.

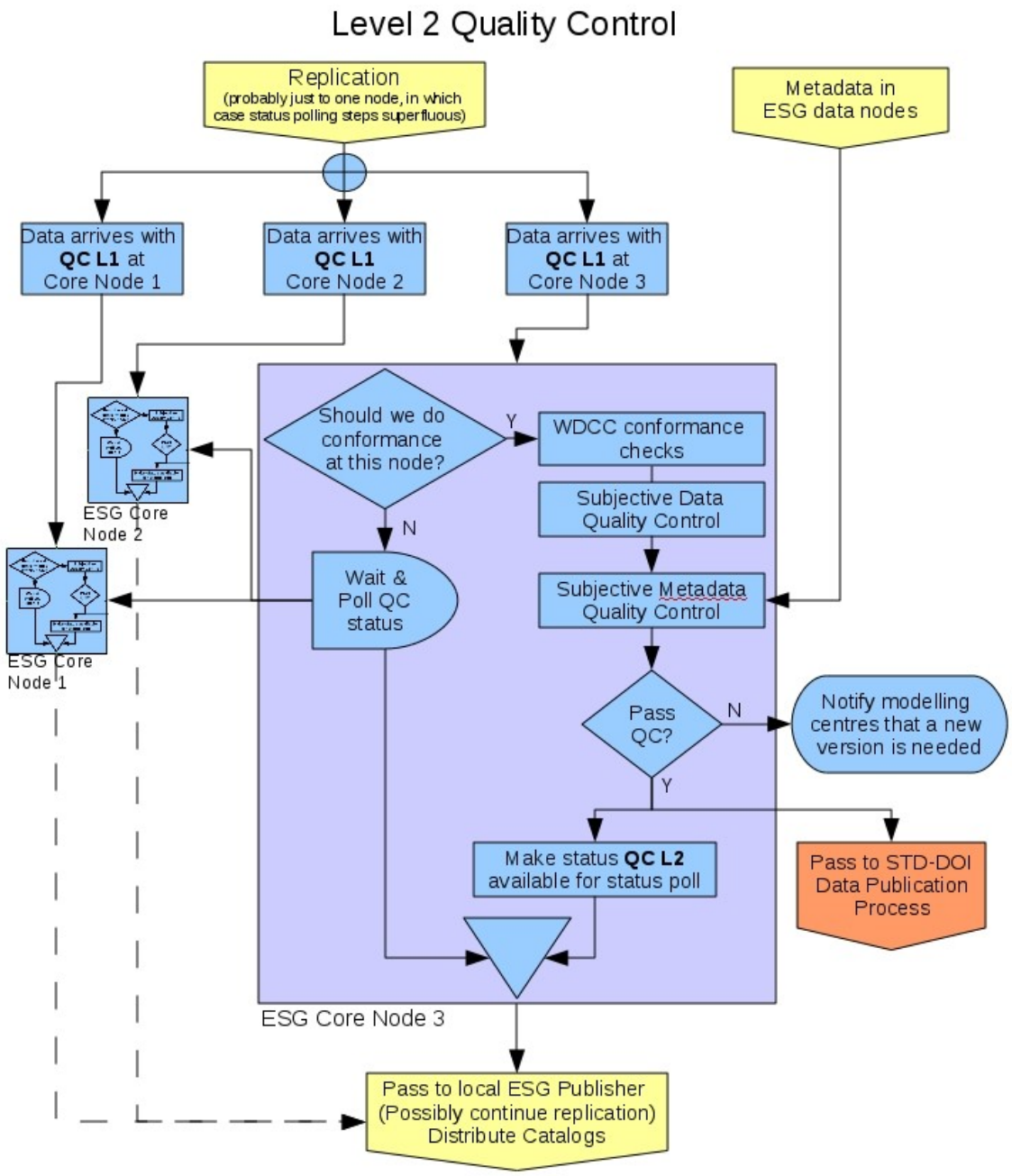

Figure 3: QC Level 2 - WDCC Conformance Checks for CMIP5 Core Data.

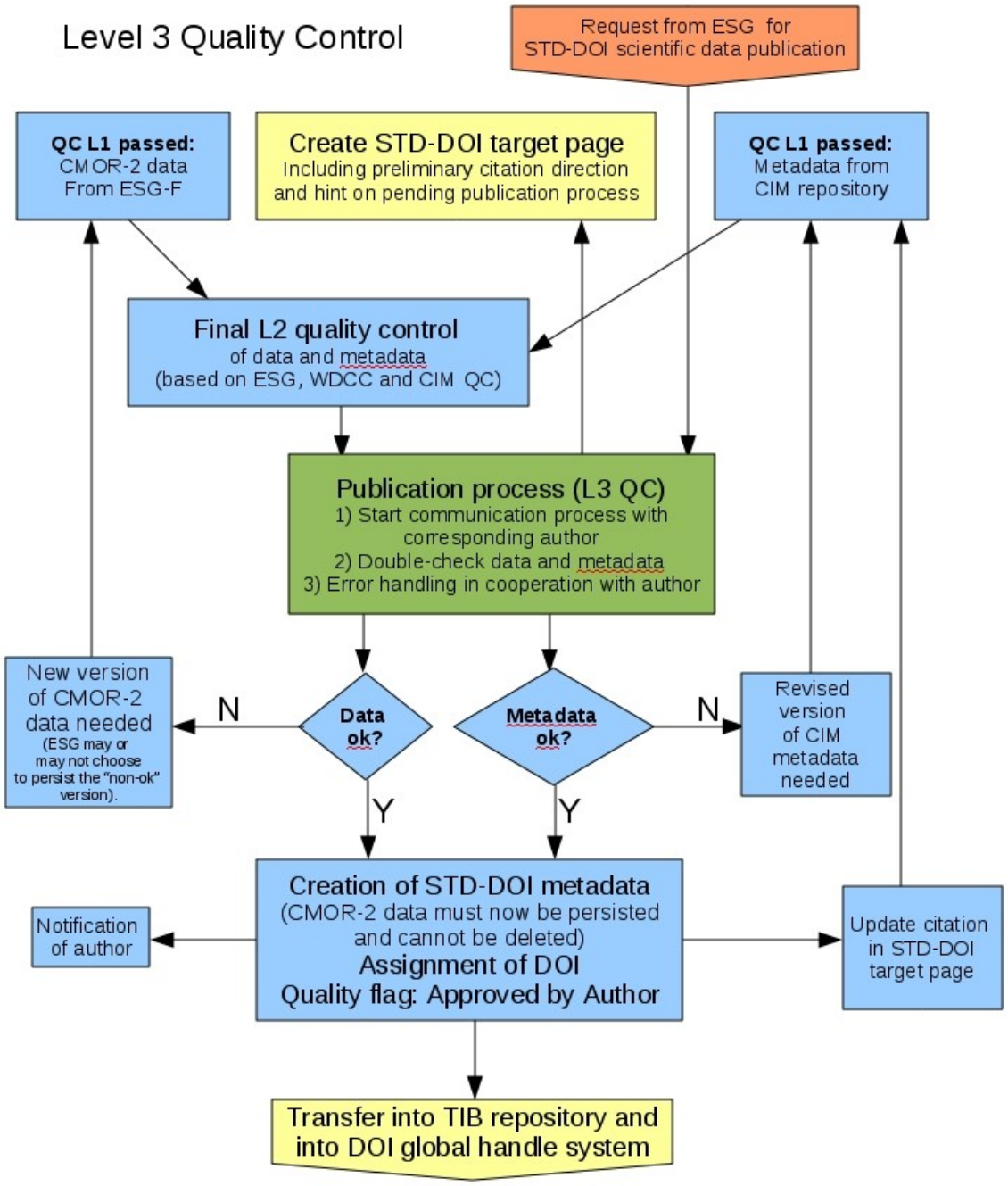

Figure 4: Final STD-DOI Data Publication Process for CMIP5 Core Data.