Questions, open issues, and points of discussion in red

Discussion Paper:

The CMIP5/AR5 Model Data Quality Control

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

March 15th, 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.

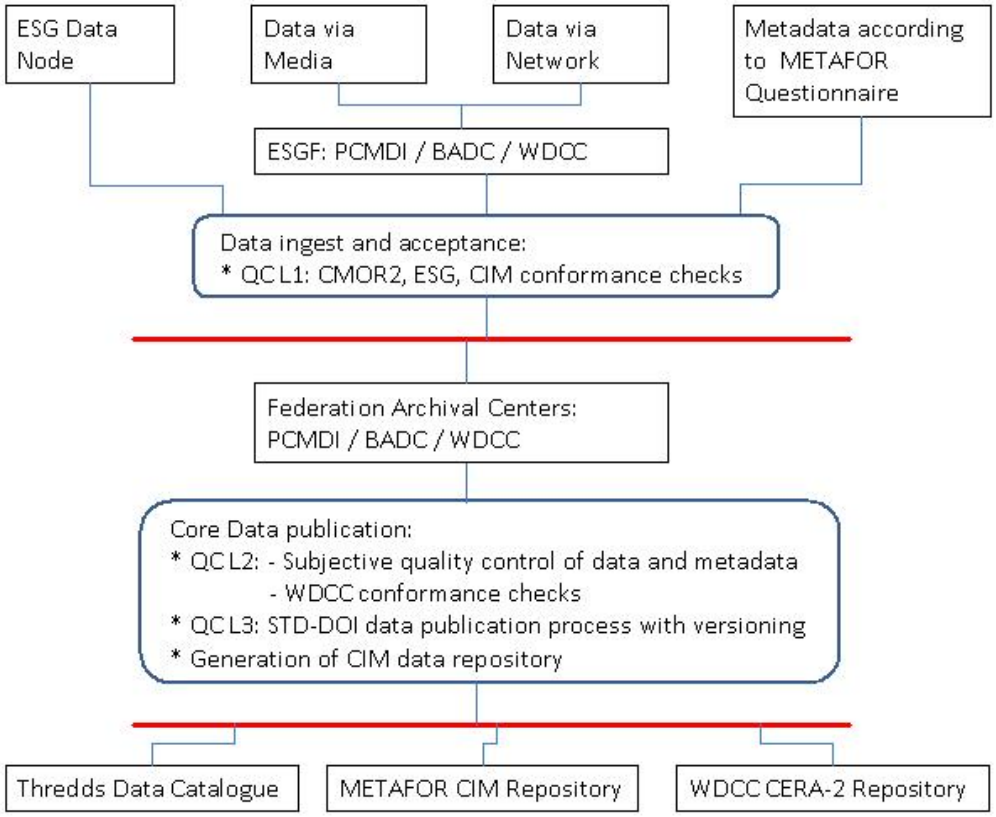

Figure 1: CMIP5/AR5 data ingest and publication process.

	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
Citation	no citation reference	preliminary citation reference	final citation reference
Quality Flag	"automated conformance checks passed"	"subjective quality control passed"	"approved by author"

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.

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):

1. CMOR2 Conformance Checks: ? *Axis, Bounds and CF standard names checked against CMOR2 tables? Logfiles ?*
2. ESG Conformance Checks: *? time formats, merge of files to datasets (time, space), parameter mapping?, setting of use metadata out of data headers? Logfiles ?*
3. CIM Conformance 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, the following data quality checks are currently suggested for the CMIP5/AR5 core data, which 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) 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
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
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.

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.

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). 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.

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 entered into the

METAFOR repository? The ESG portal is notified (QC Flag = “automated conformance checks passed”).

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

3. At these Core Nodes the WDCC Conformance Checks for QC Level 2 are carried out. Log files are entered into the METAFOR repository. The ESG portal is notified (QC Flag = “subjective quality control passed”). (or all QC L2 checks are shared between PCMDI, BADC and WDCCperformed at WDCC?)

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 or metadata can be done by replacement or modification at this stage of the quality control process. Corrected Data starts the QC process again at step 1. Old versions of data are not archived.

5. Core Data with QC Level 2 is passed to the STD-DOI publication process (checks for QC L3) at the WDC Climate and replicated to the other core data centers. 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. Core 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 (QC Flag = “approved by author”). Data are no longer matter of change.

8. For Core 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). 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 assigned. The ESG portal is notified (QC Flag = “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 1 - 3).
- Synchronization of new data (for data QC < L3); portal visibility of change?
- Notification of new data version and replication of new data version (for data with QC = L3)
- Agreement within ESGF Archival Center on which center is to perform the QC L2 checks (or all performed at WDCC?)

REFERENCES

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

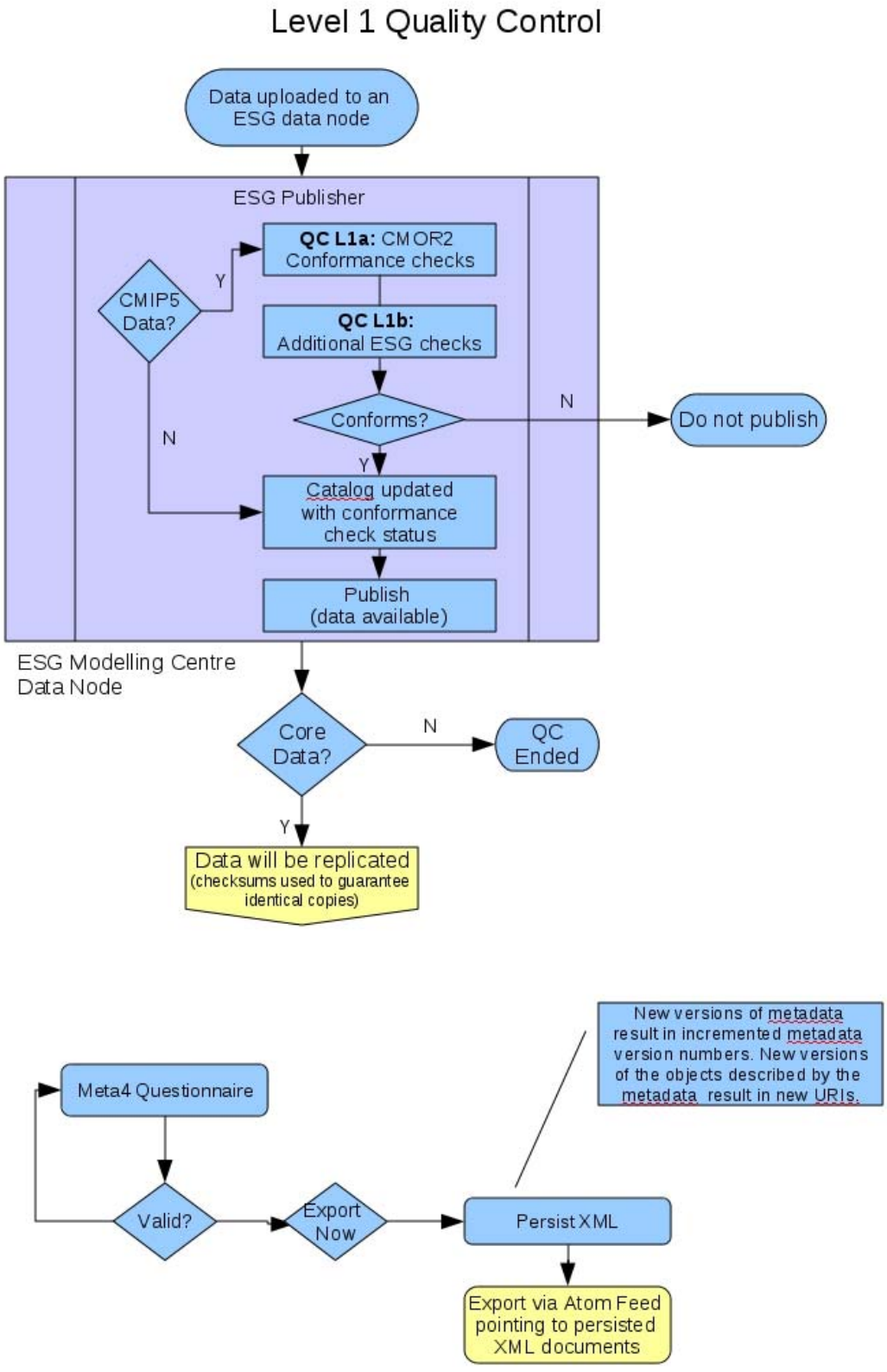

Figure 2: QC Level 1 ESG/CMOR2 and Metafor Conformance Checks for all data and metadata.

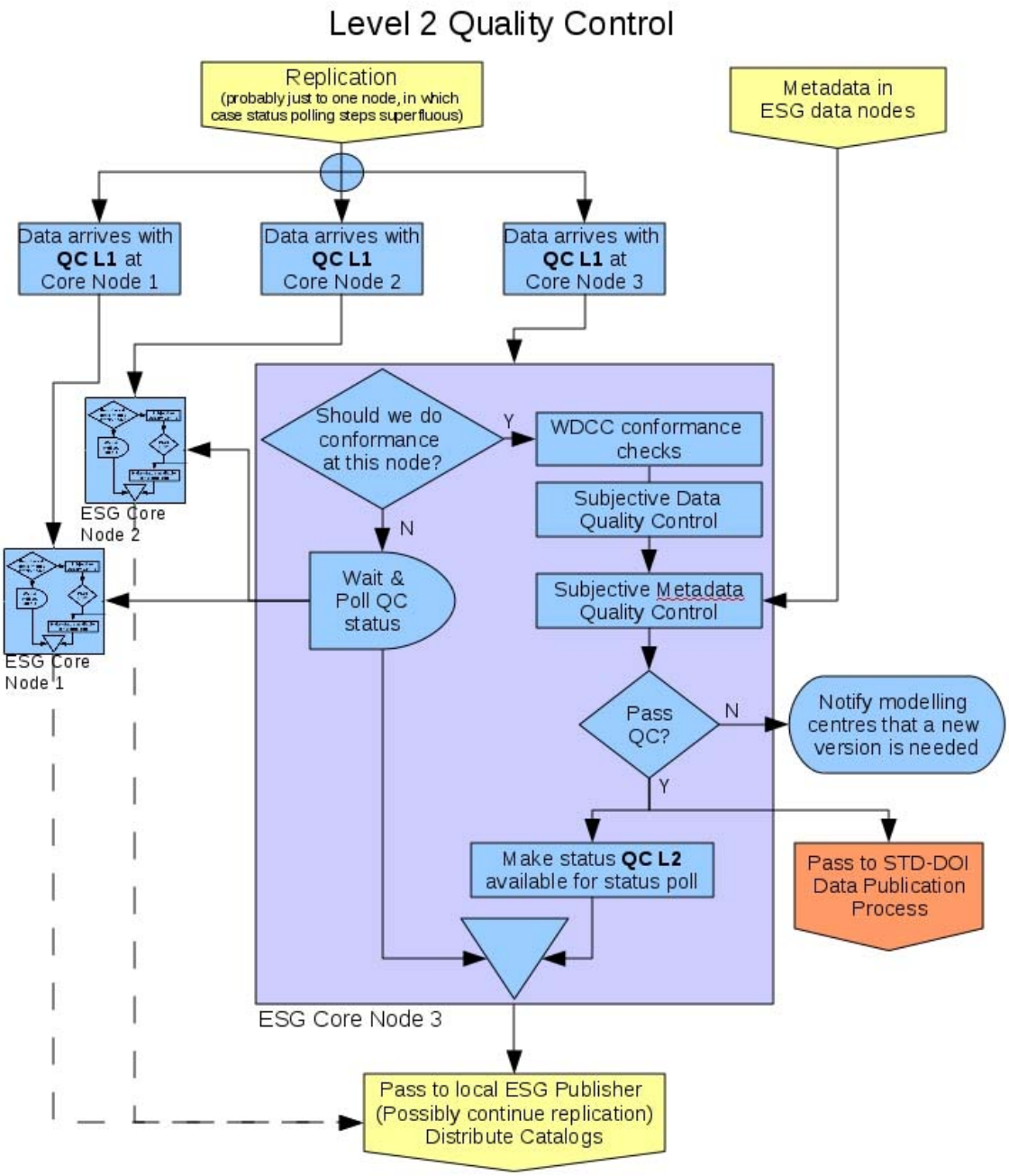

Figure 3: QC Level 2 - WDCC Conformance Checks for CMIP5 Core Data.

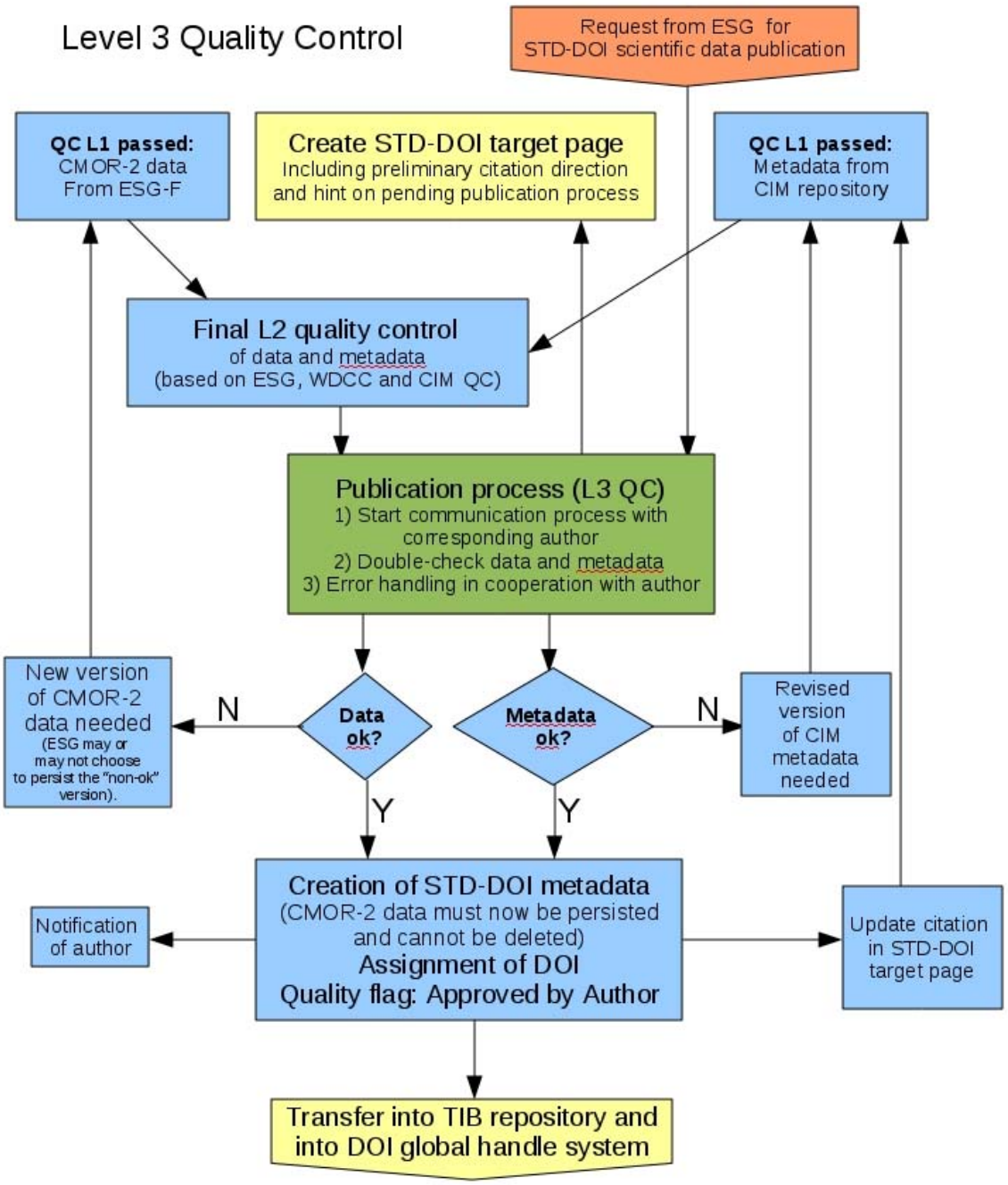

Figure 4: Final STD-DOI Data Publication Process for CMIP5 Core Data.