Discussion Paper:

Scholarly citations for CMIP5 model output

Michael Lautenschlager, V. Balaji, Bryan Lawrence

December 20th, 2009

1. Background

Model output archives such as the IPCC and CMIP archives enable scientists to write papers based on runs done by others. We propose a method by which credit can properly be assigned for the teams that perform model runs. The intent is to create a publication of climate model data entities related to a defined citation for model output. The granularity of climate model data entities which is suitable for scientific literature seems to be on the level of experiments. This is the level of granularity which has been successfully implemented by the World Data Center for Climate (www.wdc-climate.de) of DKRZ (German Climate computing Centre).

A consortium of scientific long-term data archives (WDC Climate, WDC-Mare, WDC-RSAT and GFZ-Potsdam) and the German National Library of Science and Technology (TIB) developed the concept of 'Publication and Citation of Scientific Primary Data' (STD-DOI, URL: www.std-doi.de). The implementation is available at the TIB and first scientific data publications can be obtained from the TIB library catalogue together with classical scientific publications. More information in addition to the web page can be obtained from Brase (2004) and Klump et al. (2006).

Publication Process

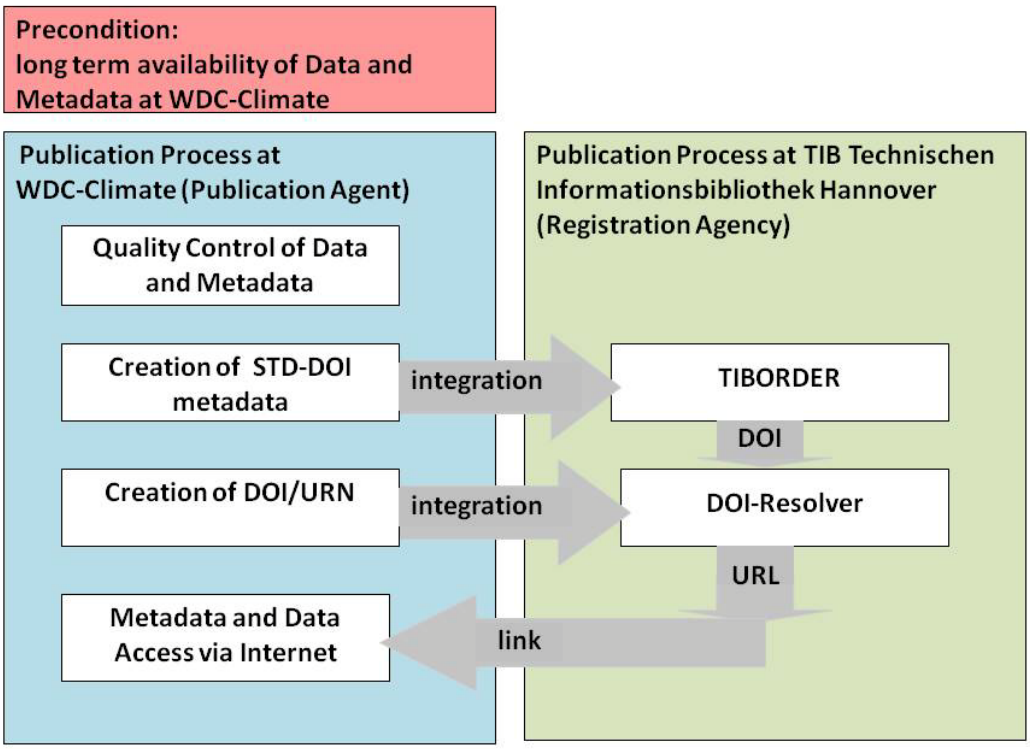
The STD-DOI service requires several organizational and technical pre-conditions to make primary scientific data citable as publications:

- Quality control of the primary data set by the author and by the data publishing agency,
- Quality control of the descriptive metadata set by the author and by the data publishing agency
- Long-term availability of the published data in online repositories
- Search function for data publications in library catalogues (e.g. TIBORDER)
- Access to the primary data with assignment of a persistent identifier and resolver system (DOI resolver)
- Published data entities are no longer matter of change (versioning)

DOI assignment is integrated in a publication process and requires final quality control of data and metadata. The main steps for scientific data publication at the WDC Climate are:

- Quality control of the scientific primary data and descriptive metadata by the authors and Publication Agencies (here WDC Climate)
- Creation of STD-DOI metadata (metadata for publication of electronic media)
- Creation of persistent identifiers (DOI/URN)
- STD-DOI metadata integration into library catalogue (TIBORDER)
- Link to primary data and metadata at the WDCC archive level by integration of the persistent identifiers in the resolver systems (DOI resolver / URN) by the Registration Agency (here TIB).
- Primary data and metadata access via internet and graphical web-browser interface of WDCC.

Flow chart of STD-DOI publication process at WDC Climate:



The STD-DOI metadata integrate the standard DOI metadata and metadata according to ISO 690-2 which are suitable for electronic publication. The detailed definition of the STD-DOI metadata profile can be obtained from http://www.icdp-online.org/contenido/std-

doi/upload/pdf/STD_metadata_kernel_v3.pdf and the corresponding XSD definition is located at http://www.tib.uni-hannover.de/doi/std-doi.xsd.

Data Quality Assurance

The quality assurance procedure contains metadata and climate data. For climate data semantic and syntactic quality assurance is distinguished. Semantic and metadata quality assurance are performed in cooperation with the data provider and will be documented in the metadata as part of the provenance information.

The syntactic quality assurance is performed by the corresponding data archive. Technical Quality Control (TQA) of data are developed and implemented at WDCC for the publication of primary data together with persistent identifiers DOI and URN. WDCC's TQA ensures consistency between data and metadata within the WDCC's database tables (field-based data access). The check list and corresponding procedures are used for simulation data and are presently expanded to meteorological measurements:

1) Number of data sets is correct and not equal 0
   2) Size of every data set is not equal 0
   3) The data sets and corresponding metadata are all accessible via internet
   (this does not exclude a personal account and the data may restricted for free use in research only)
   4) The data size is controlled and correct
   5) The time description (metadata) and existence of data are consistent
   The data is complete and no duplicates exist corresponding to the metadata description
   a) start- stop date of metadata and header information are consistent
   b) Number of accessible data units is consistent with metadata description
   c) The continuous time steps (metadata) for the accessible data units are correct
   This means no vacancies or links exist in the metadata description of the accessible data units
   6) The format is correct
   7) Variable description and data are consistent.

Quality assurance of the content of climate model data is mainly in the responsibility of the data authors (SQA – Semantic Quality Assurance). Quality assurance is documented in the metadata and the assigned quality flag in the database system is "approved by author".

Citation of Data

The size of the data sets used in a scientific publication often prohibits their publication as data tables and, as a result, data used as the basis of a publication are rarely published anymore. The lack of access to scientific data is an obstacle to interdisciplinary and international research.

Persistent identifiers together with their bibliographical information provide the opportunity to find and to cite primary data in scientific publications.

A citation of a data set follows the classical citation rules in scientific literature, e.g. author(s), publication year, data set name, persistent identifier.

Examples of data citations

Nozawa, Toru (2004): IPCC-DDC_CCSRNIES_SRES_B2: 211 YEARS MONTHLY MEANS, National Institute for Environmental Studies and Center for Climate System Research Japan, *WDCC*. doi:10.1594/WDCC/CCSRNIES_SRES_B2

Kamm,H; Machon, L; Donner, S (2004): Gas Chromatography (KTB Field Lab), *GFZ Potsdam*. doi:10.1594/GFZ/ICDP/KTB/ktb-geoch-gaschr-p

Stein, R.; Fahl, K. (2003): Distribution of grain size and clay minerals in surface sediments of the Kara Sea, *PANGAEA*, doi:10.1594/PANGAEA.119754.

Application in the literature

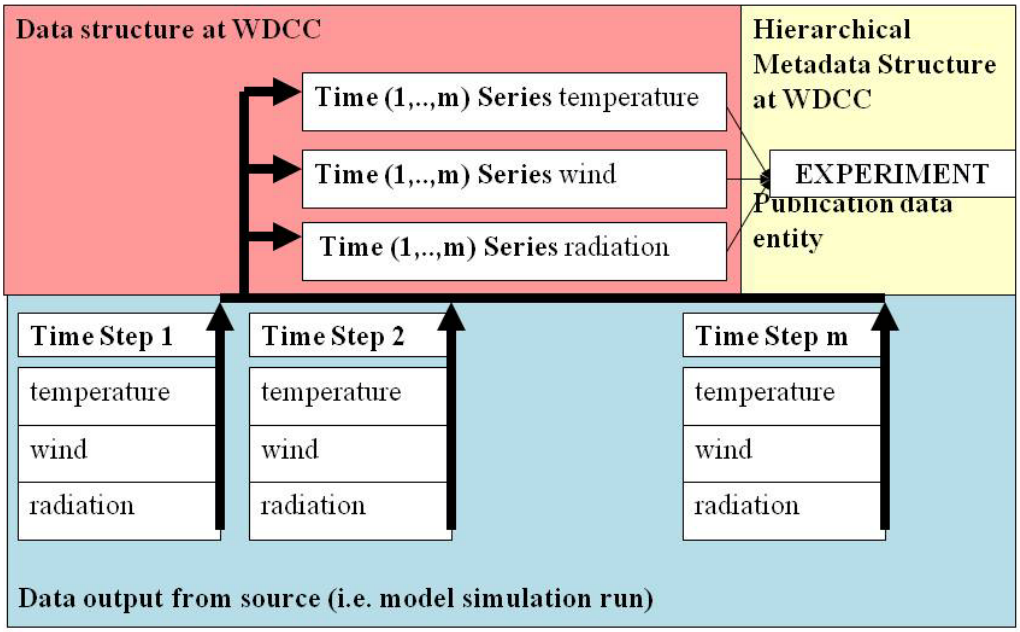
Lorenz, S.J., Kasang, D., Lohmann, G. (2005): Globaler Wasserkreislauf und Klimaänderungen - eine Wechselbeziehung, **In:** *Warnsignal Klima: Genug Wasser für alle?* Lozán, Graßl, Hupfer, Menzel, Schönwiese (Eds.), pp. 153-158. Wissenschaftliche Auswertungen, Hamburg, Germany.

This article uses and cites:
Stendel, M., T. Smith, E. Roeckner, U. Cubasch (2004): ECHAM4_OPYC_SRES_A2: 110 years coupled A2 run 6H values. *World Data Center for Climate*. doi:10.1594/WDCC/EH4_OPYC_SRES_A2.

Granularity

The complete list of data entities which are published in the TIB library catalogue via STD-DOI can be obtained from TIBORDER (http://tiborder.gbv.de/DB=2.63/LNG=EN/) choosing WDCC as search item. The result is a list of 114 data entities which ran through the WDCC publication process and which are integrated as WDCC data publications in the TIB library catalogue together with scientific literature entries. The total amount of STD-DOI data publications counts more than 1800 entities from different scientific disciplines.

WDCC defines independent data entities for climate model data at the level of individual model experiments.



The WDCC time series (data base tables) correspond with the atomic data sets in the CMIP5/AR5 data environment.

Essentials

The STD-DOI publication process (Details: http://www.mad.zmaw.de/projects-at-md/publication-and-citation/) includes assignment of persistent identifiers (DOI and URN) and of citation directive together with data entity integration in the TIB library catalogue for interdisciplinary data usage. The STD-DOI data publication process is part of the long-term archive implementation of WDCC and DKRZ. Important aspects of WDCC's scientific data publication service are in summary:

- The identification of independent data entities which are suitable for publication at the complexity level of scientific literature reference lists,
- The execution of an elaborated review process for metadata and climate data (quality control),
- The assignment of additional metadata for electronic publication (ISO 690-2) and of persistent identifiers (DOI / URN) and
- The integration of publication metadata and persistent identifiers into the TIBORDER library catalogue (German National Library of Science and Technology, Hannover) so that primary data entities are searchable and citable together with scientific literature.

- Quality characteristic is presently “approved by author”, could be “peer reviewed” with ESSD (Earth System Science Data Journal).
- Published data entities cannot be modified any longer.
- Data are freely available via Internet.

The STD-DOI data publication at WDCC is ready to use and it is suggested to use it for assignment of persistent identifiers and citation directives to CMIP5/AR5 data climate model data entities. The STD-DOI publication process requires suitable metadata even at the scientific level. These metadata are closely related to the METAFOR metadata questionnaire and the resulting CIM repository. The envisaged CMIP5 data and metadata quality checks can be directly connected with the ESG data publication process. DOI assignment is closely related to versioning. Each new version requires a new persistent identifier and citation directive in order to identify the data entity precisely within the literature.

2. Proposal summary

The proposal consists of a series of steps that will be jointly undertaken by the modeling centers submitting data to the archive, and the data centers responsible for data distribution: the IPCC archive federation at BADC, the WDCC of DKRZ, and PCMDI for CMIP5.

- WDCC’s STD-DOI scientific data publication will be used to assign persistent identifiers (DOI/URN) and citation directions to CMIP5/AR5 data entities.

- A DOI will be registered by the WDCC for each simulation and version in the replicated data archive. Once registered, the data entity is a static and persistent referenced. Versioning is reflected by different DOIs. The relation between DOI and the data archive location (URL) is maintained in the STD-DOI metadata repository. Though its DOI target can be moved.

- The DOI will point to a page to be hosted by the WDCC and linked to the CIM scientific data repository at British Atmospheric Data Centre (and to be transitioned to the IPCC-DDC). The page will contain information relevant for citation (title and description of simulation, author list, institution, funding, etc.), detailed information on the data themselves (use metadata) and transparent data access. (See any experiment in the CERA database http://cera-www.dkrz.de or http://www.mad.zmaw.de/service-support/consortium-model-runs/ipcc-experiments/ for IPCC AR4 for a current example of what a citation might look like.)

- After resolving the DOI the information on this page will be automatically generated from the WDCC repository together with a link to the “CIM metadata repository” generated from the CMIP5 questionnaire that modeling centers will use to document their submissions to CMIP5.

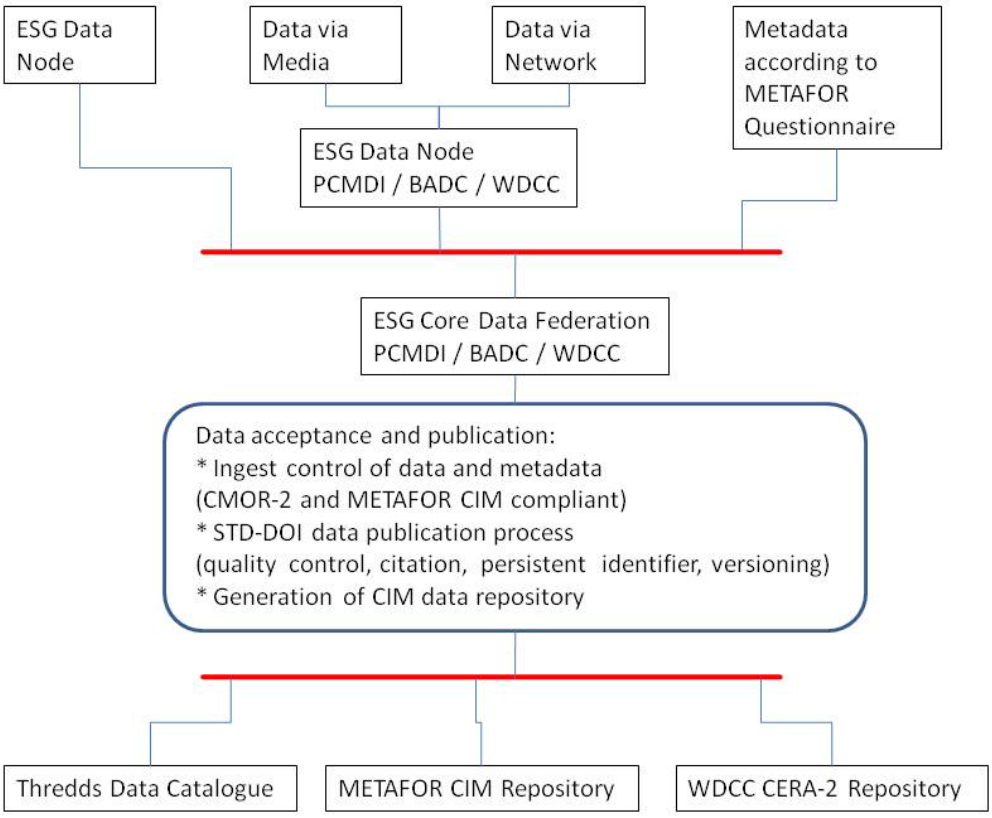
- A journal such as Earth System Science Data: ESSD (http://www.earth-system-science-

data.net/) will accept entries of this nature as submissions. ESSD's peer-review process will be used to monitor and validate the simulation description. The ESSD external review process requires more effort and time as the STD-DOI scientific data publication process and should considered as add-on to the underlying STD-DOI data publication. The ESSD data publication will be in the responsibility of the individual modeling groups and is not part of the CMIP5/AR5 data publication process.

The CMIP5/AR5 scientific data publication process which is based on the STD-DOI data publication service at WDCC is suggested to be integrated into overall CMIP5/AR5 data management.

3. Implementation of the CMIP5/AR5 model data publication process

A flow diagram of the CMIP5 data ingest and publication together with tasks for data acceptance and publication is provided in the following diagram:



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 suggested STD-DOI scientific data publication should be closely related to ESD data publication

and versioning. ESG data publication and STD-DOI scientific data publication is synchronous in the CMIP5/AR5 data archive federation and each ESG data version will receive an individual persistent identifier and a specific citation direction. This synchronized publication procedure yields some benefits.

- The CMIP5/AR5 data federation provides quality proven data.
- Each model data version is connected to an agreed citation direction.
- Each model data version can be identified and accessed by an individual persistent identifier (DOI/URN).
- The STD-DOI scientific data publication allows for model data search together with scientific journal publications in library catalogues (TIBORDER).
- The STD-DOI scientific data publications allows for data access directly by DOI and the IDF global handle server independent from the actual archive location.

But the STD-DOI data publication is connected to a few requirements:

- The more elaborate quality assurance of the STD-DOI data publication must be finished before ESG data publication in order to ensure versioning, referencing and citation.
- Data are fixed and no longer matter of change after the STD-DOI publication and persistent identifier assignment. Small errors can be addressed as “Erratum”, larger modifications result in a new version of the data entity with a new persistent identifier (DOI/URN).
- The existing STD-DOI scientific data publication process requires metadata data entry and data integration in the WDCC data repository CERA-2.
- Quality assurance must be designed to work with 2 or 3 three times of the amount of replicated CMIP5/AR5 data within one year. Here the experience enters that model data normally have to send more than once because of errors.

STD-DOI Quality Assurance

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
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) Phsical 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:

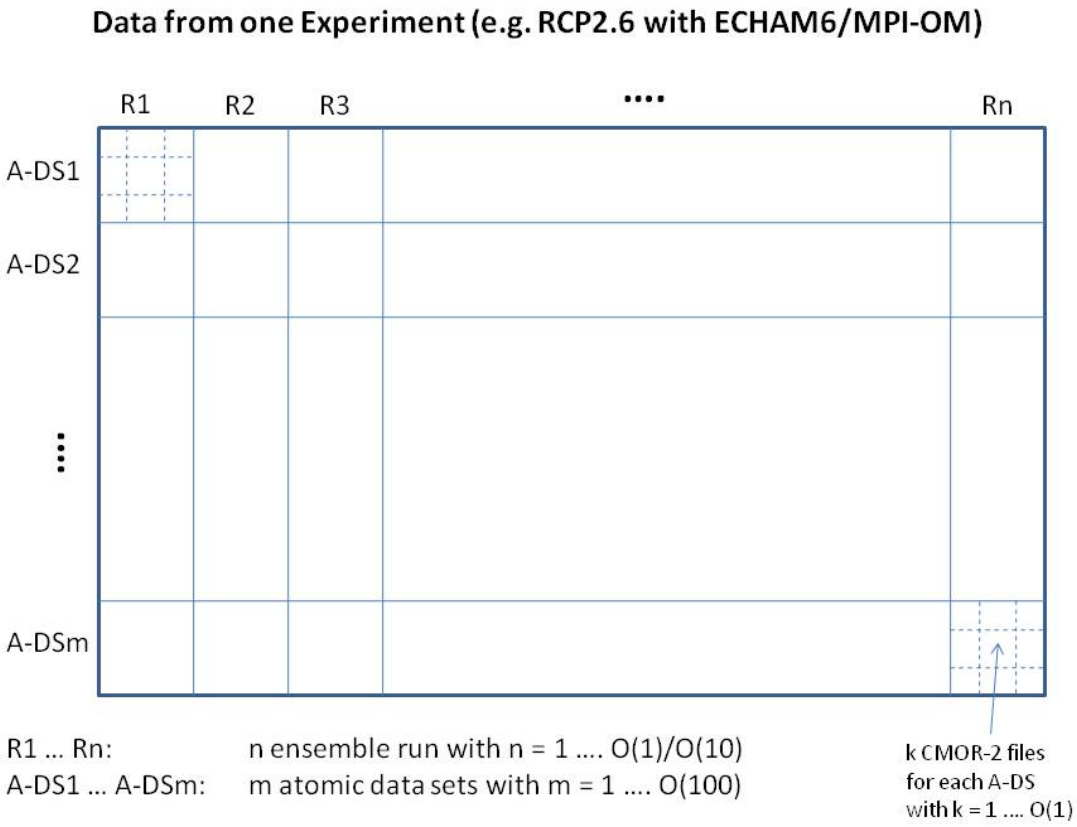
- min
- max
- globally weighted mean
- area weighted mean (reasonable for instance of temperature of snow)
- global arithmetic mean
- standard deviation of the globally weighted mean.

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

Granularity

An additional aspect in the CMIP5/AR5 data publication is the specification of the granularity of independent data entities which is suitable for scientific journal publications. The granularity must be fine enough for direct data access and it must be coarse enough not to overburden reference lists in scientific journals. Data citations and citations of scientific articles must be balanced in the reference lists in order to achieve acceptance of scientific journal publishers and editors as well as of readers.

The following figure provides an illustration of the CMIP5 data production matrix for a specific experiment calculated with a specific model.



The actually implemented STD-DOI scientific data publication service at WDCC/DKRZ bases on the individual model experiment as definition for the suitable independent data entity. For ensemble experiments with small number of runs the existing WDCC data publication procedure assigns one DOI/URN per run. This works for ensemble sizes of 2-5. But this does not take into account climate model ensemble experiments with larger numbers of realizations (or runs).

Therefore the suggestion for CMIP5/AR5 is to assign persistent identifiers at two levels, at the level of individual runs (R1, R2, …, Rn) and one additional, summary DOI for all ensemble realizations. That gives the freedom to reference individual runs and/or to cite the entire experiment in dependence of the requirements of the specific scientific publication.

REFERENCES

Brase, J. (2004): Using Digital Library Techniques - Registration of Scientific Primary Data. Lecture Notes in Computer Science 3232, 488-494.

Guilyardi, E., 2006: El Niño: mean state-seasonal cycle interactions in a multi-model ensemble.
Climate Dynamics, **26 (4)**, 329–348.

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., Schindler, U., Sens, I. and Wächter, J. (2006): Data publication in the Open Access Initiative. Data Science Journal 5, 79-83. doi:10.2481/dsj.5.79.