CMIP5 Data Management

CAS2K13

08. - 12. September 2013, Annecy

Michael Lautenschlager (DKRZ)

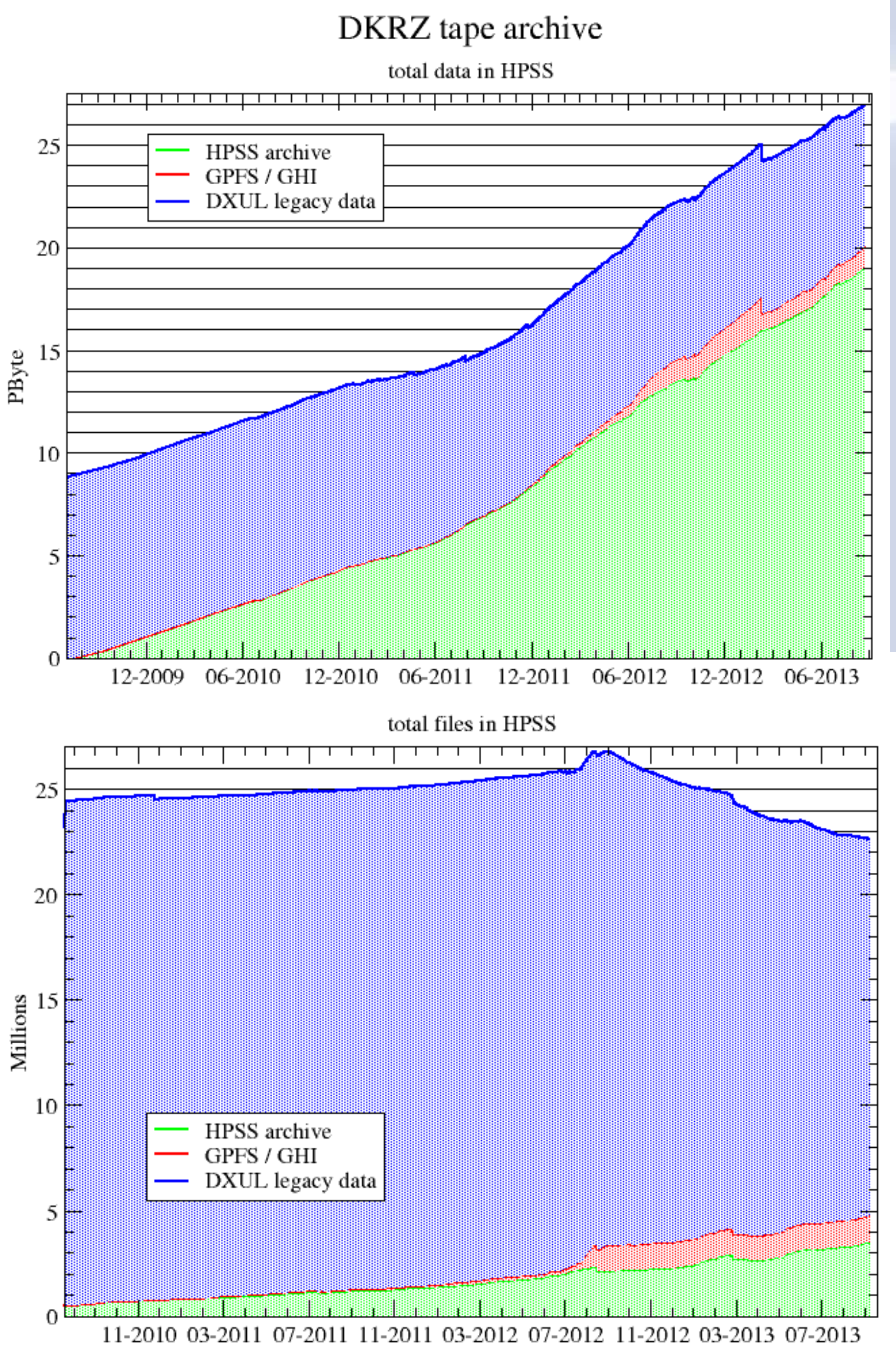
With Contributions from ESGF CMIP5 Core Data Centres
PCMDI, BADC and DKRZ

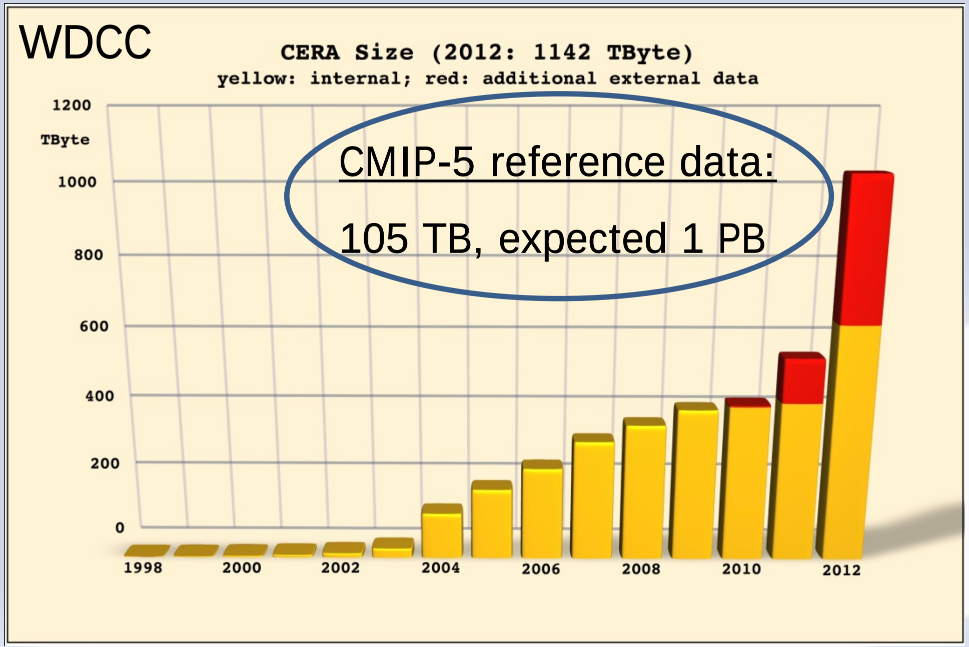
DKRZ

Status DKRZ Data Archive

HLRE-2 archive concept from 2009:

Annual growth rate with 6 PB/year is less than expected and total number of files is small compared to HLRE-1





CMIP-5 reference data:

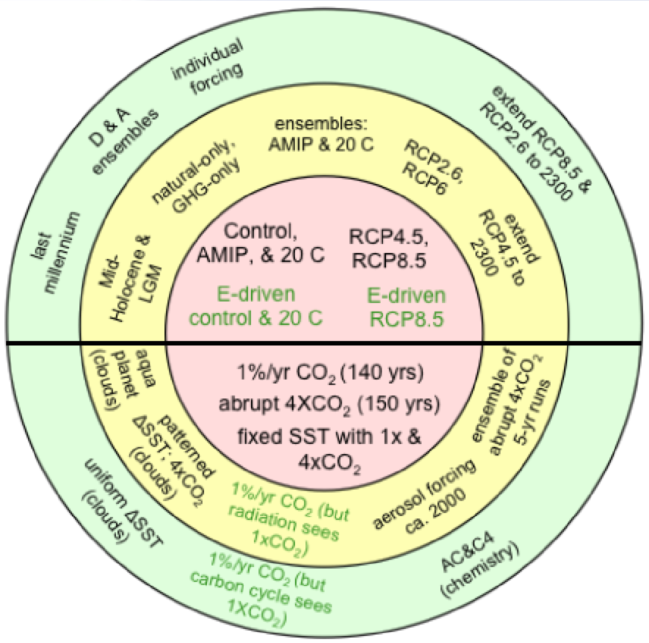
105 TB, expected 1 PB

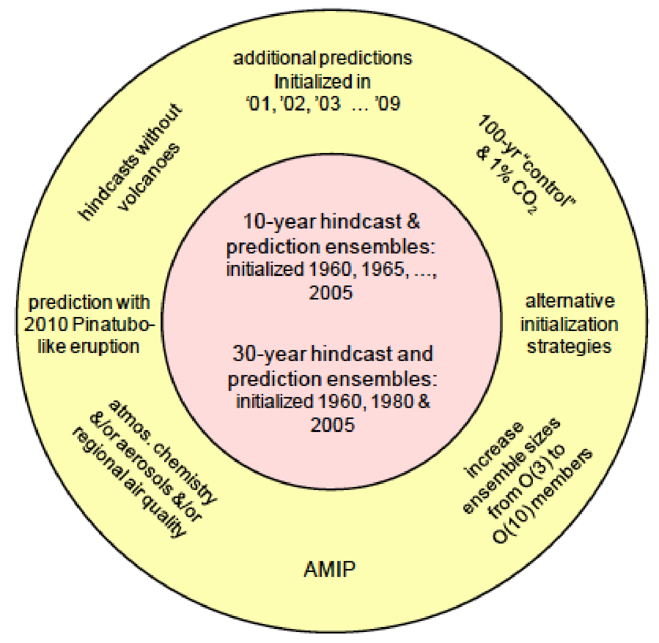
enschlager, DKRZ)

CMIP5 Protocol +Timeline

Taylor et al (2009), "A Summary of the CMIP5 Experiment Design“

Centennial Experiments

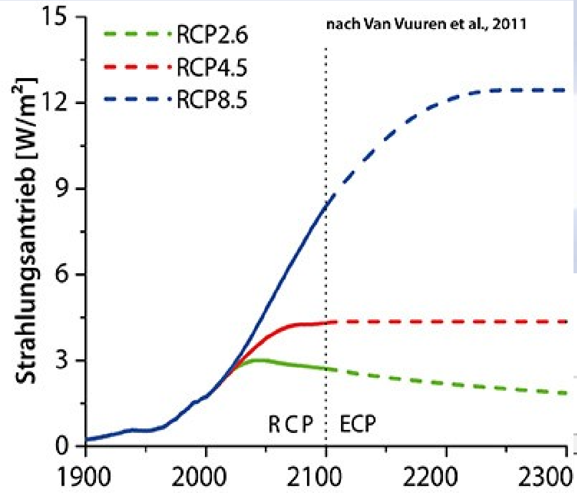



additional predictions Initialized in ‘01, ’02, ’03 … ’09
hindcasts without volcanoes
100-yr “control” & 1% CO_2
10-year hindcast & prediction ensembles: initialized 1960, 1965, …, 2005
prediction with 2010 Pinatubo-like eruption
alternative initialization strategies
30-year hindcast and prediction ensembles: initialized 1960, 1980 & 2005
atmos. chemistry &/or aerosols &/or regional air quality
increase ensemble sizes from O(3) to O(10) members
AMIP

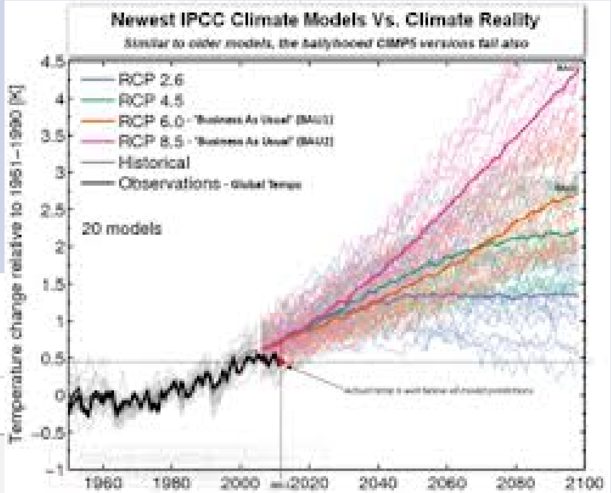
Decadal Experiments

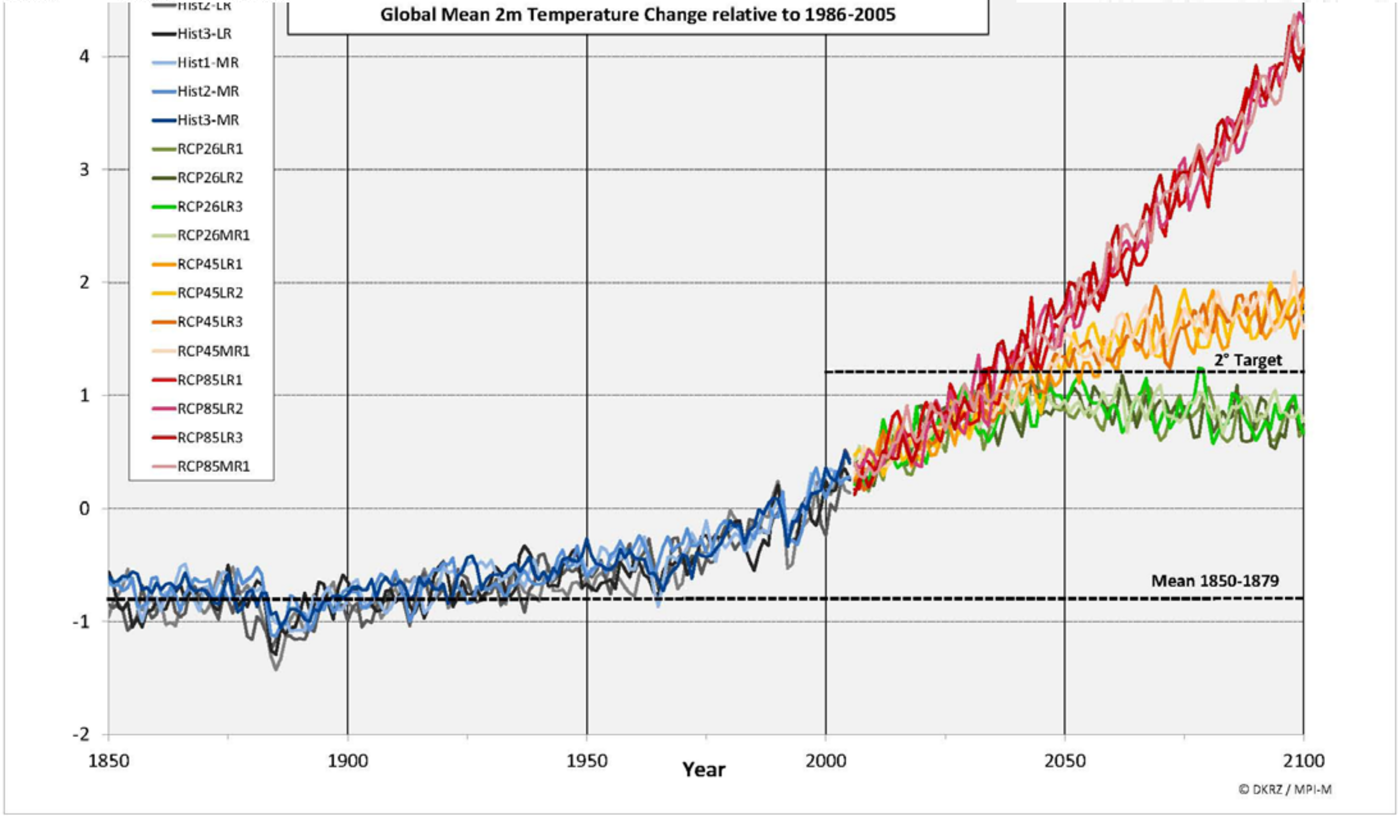
Timeline:

- 2007 – 2009: CMIP5 definition with Taylor et al (2009) as result
- 2010 – 2011: Climate model calculations and archive design
- 2011 – 2013: CMIP5 archive build up (presentation at CAS2K11)



CMIP5 Results





CMIP5 Release WS MPI-M/DKRZ (Feb. 2012)

Data Amounts CMIP3/CMIP5

- CMIP3 / IPCC-AR4 (Report 2007)
  - Participation: 17 modelling centres with 25 models
  - In total 36 TB model data central at PCMDI and ca. ½ TB in IPCC DDC at WDCC/DKRZ as reference data

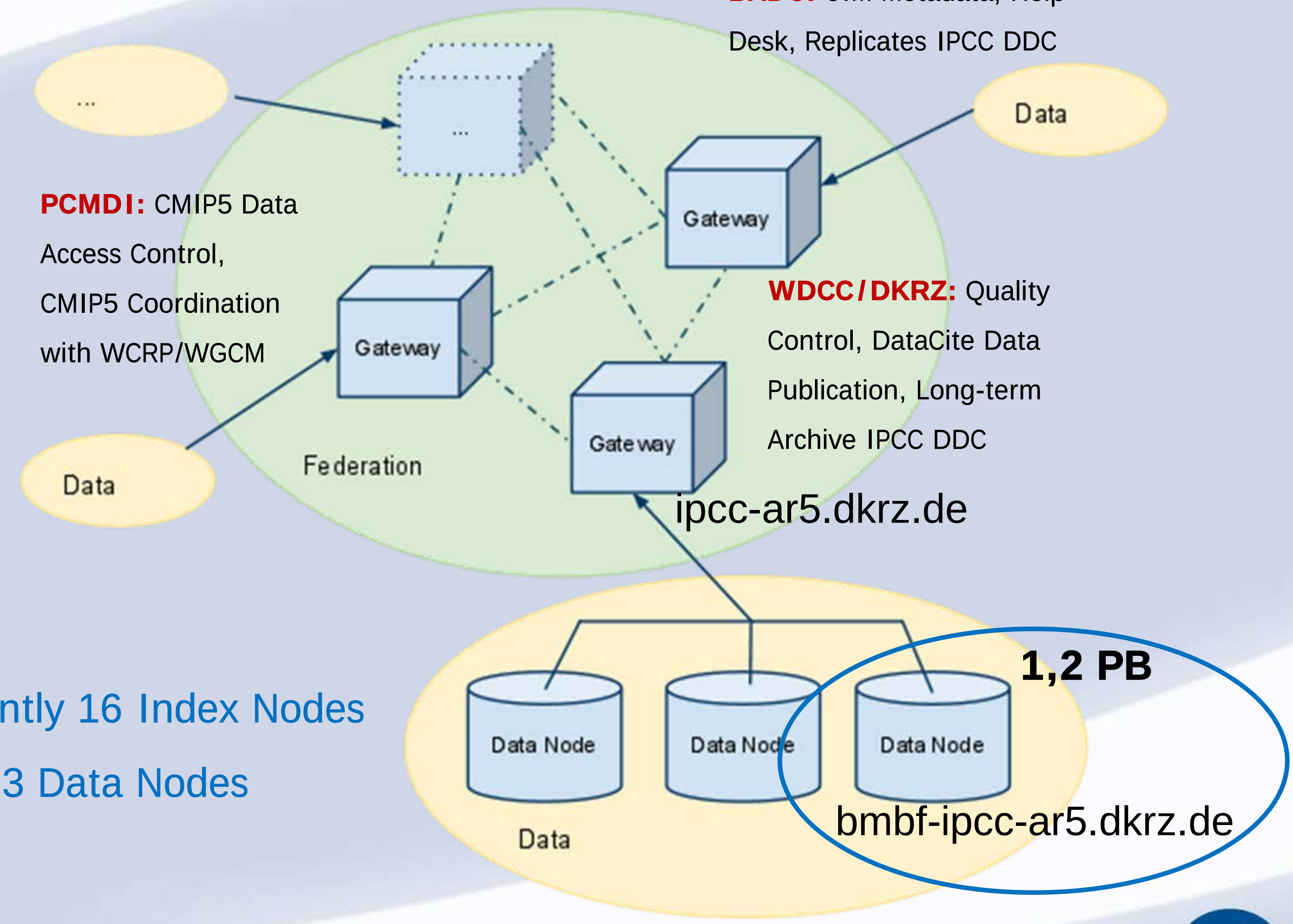
- CMIP5 / IPCC-AR5 (Report 2013/2014)
  - Participation: 29 modelling groups with 61 models
  - Produced data volume: ca. 10 PB with 640 TB from MPI-ESM
  - CMIP5 requested data volume: ca. 2 PB (in CMIP5 data federation)
  - Data volume for IPCC DDC: ca. 1 PB (complete quality assurance process) with 60 TB from MPI-ESM

- Status CMIP5 data archive (June 2013):
  - 1.8 PB for 59000 data sets stored in 4.3 Mio Files in 23 data nodes
  - CMIP5 data is about 50 times CMIP3

Usage Requirements for CMIP5

- Results from CMIP5 (Coupled Model Intercomparison Project No. 5) are for
  - Model intercomparisons with respect to climate model improvement and consolidation of the climate system knowledge
  - Usage as common data basis for scientific publications as basis for the IPCC Assessment Report No. 5 (IPCC-AR5)
- New in IPCC-AR5: all three working groups should use the same model data base
- Resulting interdisciplinary applications (*IAV – Impact, Adaptation/Mitigation, Vulnerability*) imposes high requirements to data quality and documentation
- This has implications for treatment and provision of climate data in the IPCC DDC (IPCC Data Distribution Centre) compared to AR4
- This means accomplishment of quality control and data documentation in connection or just after the climate model runs in order to remove data errors and inconsistencies prior to the (interdisciplinary) usage.

CMIP5 Data Federation (P2P)



BADC: CIM Metadata, Help-Desk, Replicates IPCC DDC

PCMDI: CMIP5 Data Access Control, CMIP5 Coordination with WCRP/WGCM

WDCC/DKRZ: Quality Control, DataCite Data Publication, Long-term Archive IPCC DDC

ipcc-ar5.dkrz.de

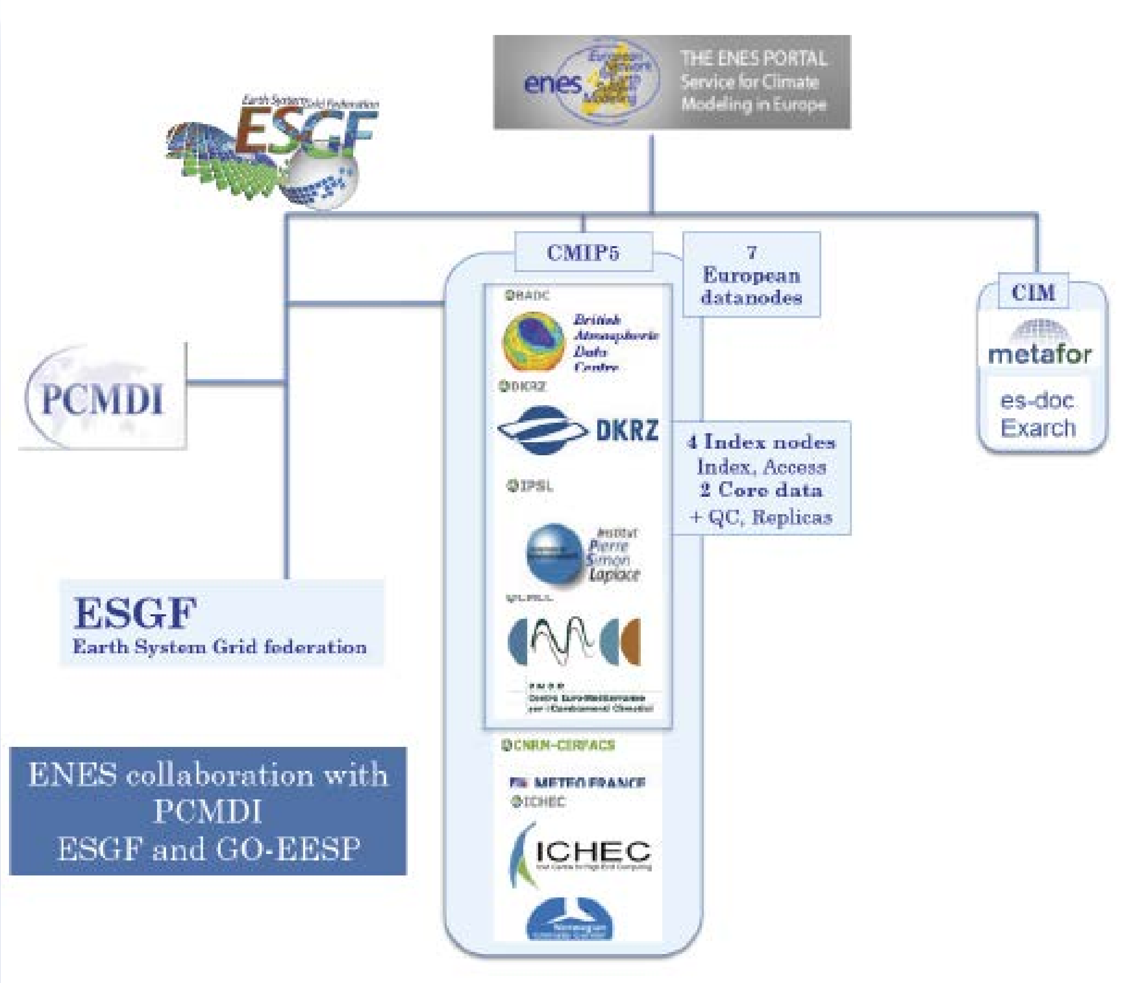
1,2 PB

bmbf-ipcc-ar5.dkrz.de

Currently 16 Index Nodes and 23 Data Nodes

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK
FOR EARTH SYSTEM MODELLING

European Contribution to ESGF-CMIP5



THE ENES PORTAL
Service for Climate
Modeling in Europe

CMIP5

7
European
datanodes

4 Index nodes
Index, Access
2 Core data
+ QC, Replicas

CIM

metafor

es-doc
Exarch

ESGF
Earth System Grid federation

ENES collaboration with
PCMDI
ESGF and GO-EESP

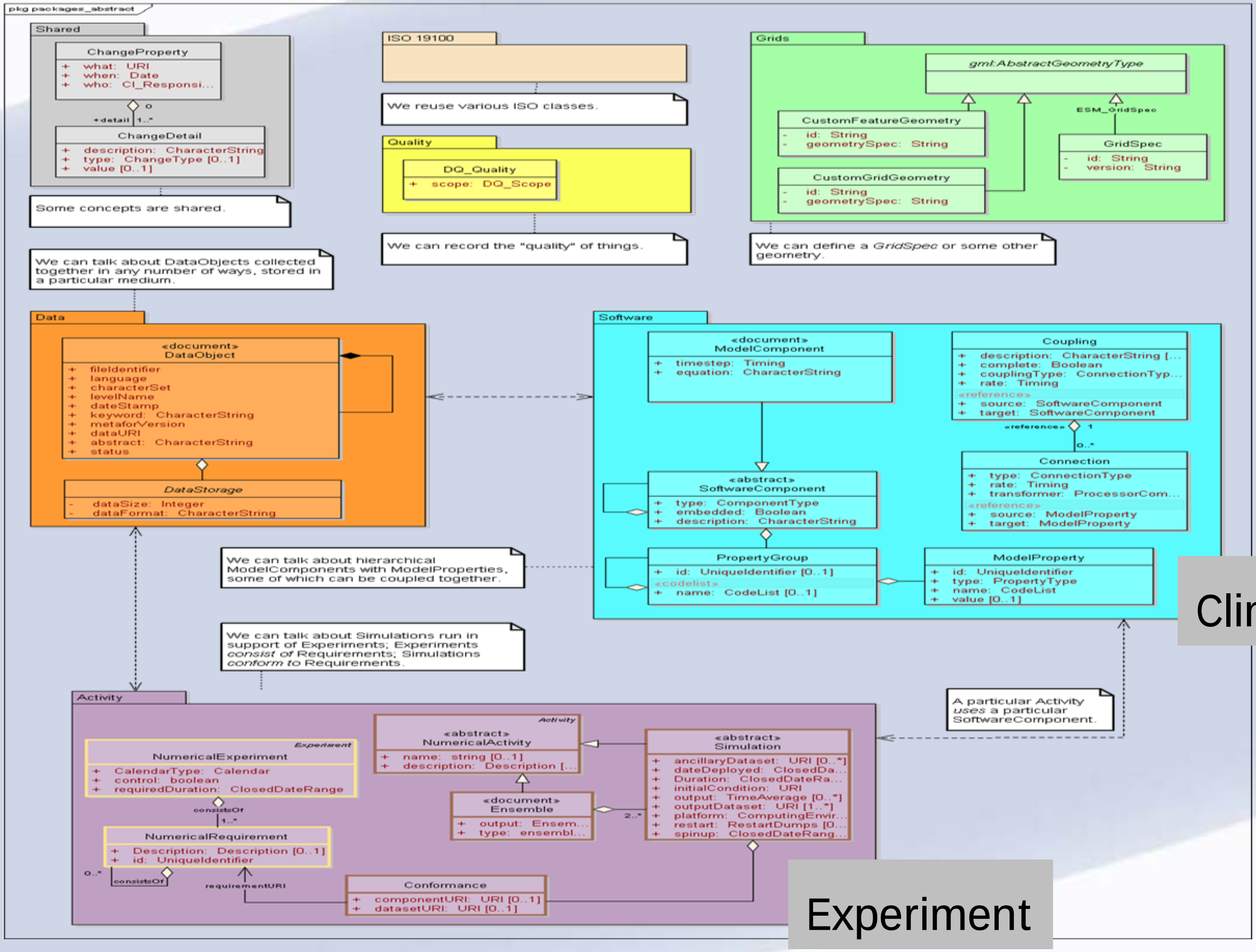
enes
Earth System Grid Federation
ESGF
PCMDI
BADC
British Atmospheric Data Centre
DKRZ
DKRZ
IPSL
Institut Pierre Simon Laplace
CMCC
CNRM-CERFACS
METEO FRANCE
ICHEC
ICHEC

The FP7 project IS-ENES contributes to ESGF-CMIP5 with 7 European data nodes and 4 index nodes

CMIP5 Data Federation

- 3 central management components have been planned for interdisciplinary data re-use
  - Highly structured data files in self-descriptive data format NetCDF/CF with use-metadata
  - New: searchable model and experiment descriptions (CIM metadata from EU-Project METAFOR)
  - New: 3 layer quality assurance concept for data and metadata
    - QC-L1: ESGF publisher conformance checks
    - QC-L2: Data consistency checks
    - QC-L3: Double- and cross-checks of data and metadata and DataCite data publication

CIM Metadata



Some concepts are shared.

We reuse various ISO classes.

We can record the "quality" of things.

We can define a *GridSpec* or some other geometry.

We can talk about DataObjects collected together in any number of ways, stored in a particular medium.

We can talk about hierarchical ModelComponents with ModelProperties, some of which can be coupled together.

We can talk about Simulations run in support of Experiments; Experiments *consist of* Requirements; Simulations *conform to* Requirements.

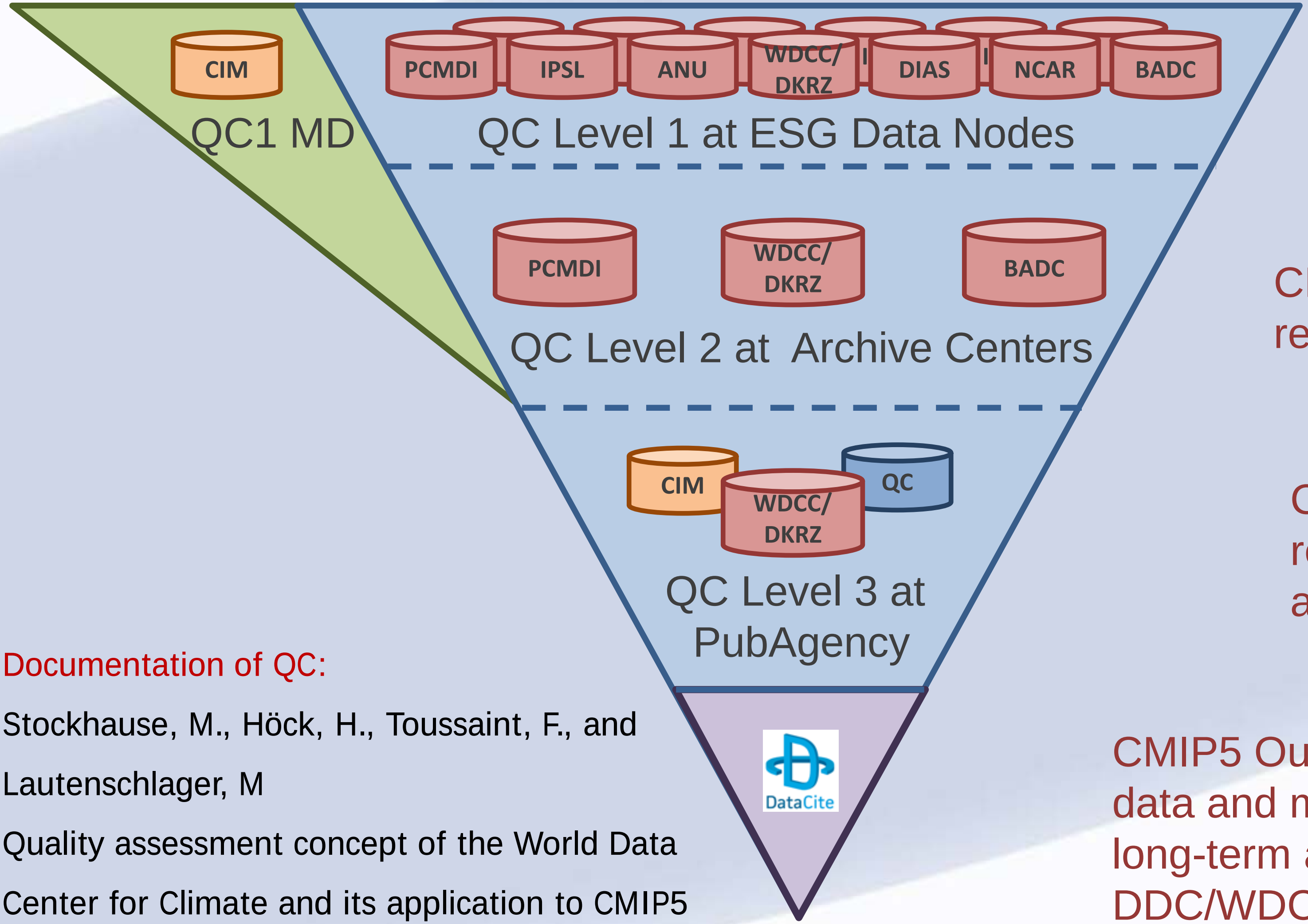
A particular Activity *uses* a particular SoftwareComponent.

Climate Model

Experiment

- Development: FP7 METAFOR
- New: Documentation of data creation process in close connection with climate data
- Improvement: Searchable model and experiment description

3-Layer Quality Assurance Concept



All CMIP5
requested data
(CMOR-2)

CMIP5 Output1 /
replicated Data

CMIP5 Output1 /
replicated data
and metadata

CMIP5 Output1 / replicated
data and metadata in the
long-term archive (IPCC
DDC/WDCC)

Documentation of QC:

Stockhause, M., Höck, H., Toussaint, F., and Lautenschlager, M

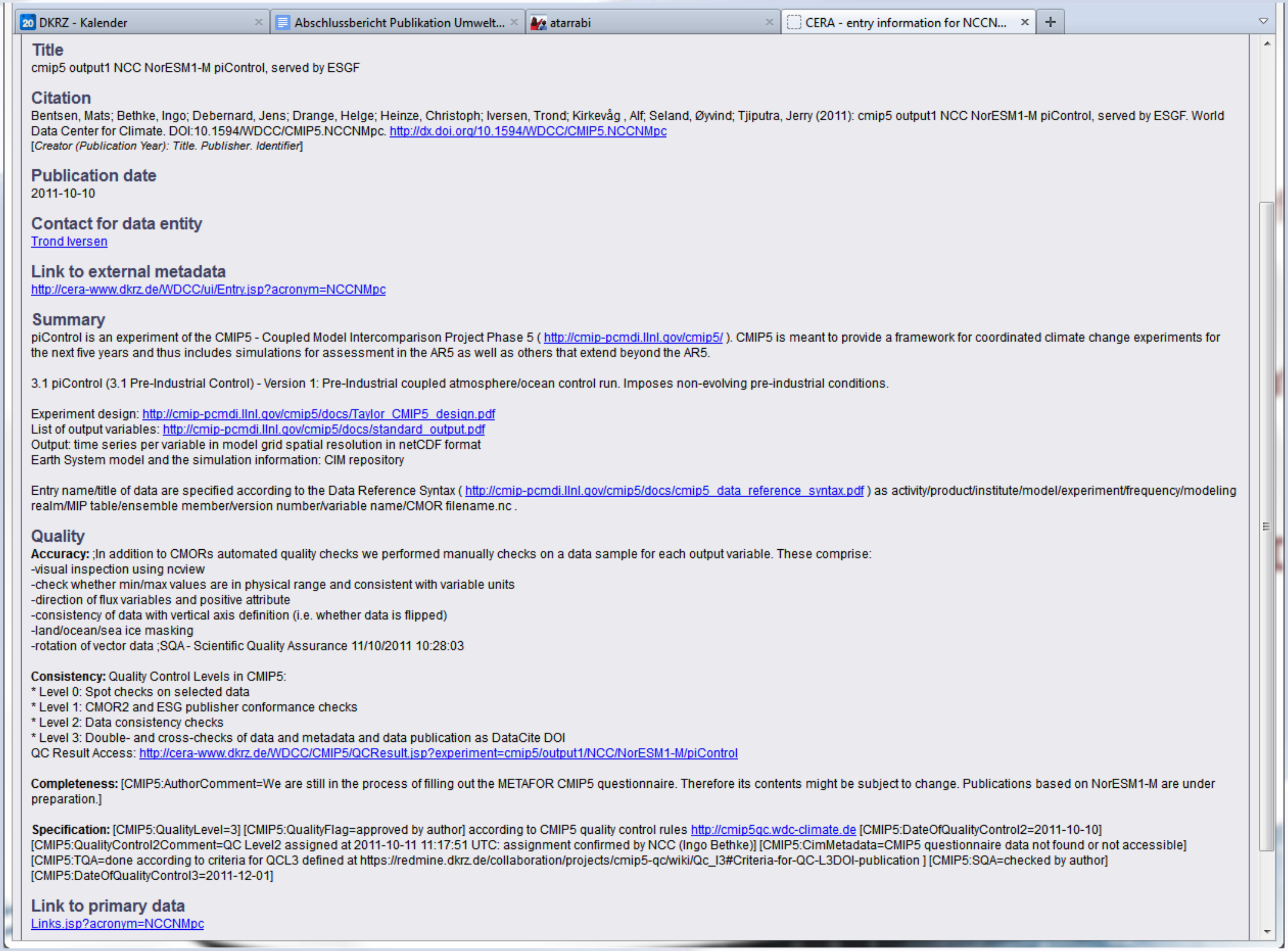
Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data. Geosci. Model Dev., 5, 1023-1032 , 2012

DOI : 10.5194/gmd-5-1023-2012

Finalisation of Quality Assessment

- After final control of data and metadata (CIM und CF) CMIP5 data are transferred from the ESGF archive (*most recent version*) into the reference data archive (*snapshot around March 2013*)
  - Quality status: „approved by author“
  - Data are marked as irrevocable
  - Long-term archiving in WDC Climate of DKRZ
- Final step is the DataCite data publication and integration of associated citation reference into library catalogues
  - Data entity (here one climate model experiment) receives a citation reference for direct usage in scientific publications and a DOI (Digital Object Identifier) for the transparent data access
  - Citation reference contains data author and title as well as WDC Climate as DataCite DOI publisher and the DOI
  - Resolution of the DOI leads to a „Landing Page“, which address is stored in the central data base of the DOI Handle Server at DataCite

DOI Landing Page


DKRZ - Kalender
Abschlussbericht Publikation Umwelt...
atarrabi
CERA - entry information for NCCN...

Title

cmip5 output1 NCC NorESM1-M piControl, served by ESGF

Citation

Bentsen, Mats; Bethke, Ingo; Debernard, Jens; Drange, Helge; Heinze, Christoph; Iversen, Trond; Kirkevåg , Alf; Seland, Øyvind; Tjiputra, Jerry (2011): cmip5 output1 NCC NorESM1-M piControl, served by ESGF. World Data Center for Climate. DOI:10.1594/WDCC/CMIP5.NCCNMpc. http://dx.doi.org/10.1594/WDCC/CMIP5.NCCNMpc
[Creator (Publication Year): Title. Publisher. Identifier]

Publication date

2011-10-10

Contact for data entity

Trond Iversen

Link to external metadata

http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=NCCNMpc

Summary

piControl is an experiment of the CMIP5 - Coupled Model Intercomparison Project Phase 5 (http://cmip-pcmdi.llnl.gov/cmip5/). CMIP5 is meant to provide a framework for coordinated climate change experiments for the next five years and thus includes simulations for assessment in the AR5 as well as others that extend beyond the AR5.

3.1 piControl (3.1 Pre-Industrial Control) - Version 1: Pre-Industrial coupled atmosphere/ocean control run. Imposes non-evolving pre-industrial conditions.

Experiment design: http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf
List of output variables: http://cmip-pcmdi.llnl.gov/cmip5/docs/standard_output.pdf
Output: time series per variable in model grid spatial resolution in netCDF format
Earth System model and the simulation information: CIM repository

Entry name/title of data are specified according to the Data Reference Syntax (http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf) as activity/product/institute/model/experiment/frequency/modeling realm/MIP table/ensemble member/version number/variable name/CMOR filename.nc .

Quality

Accuracy: ;In addition to CMORs automated quality checks we performed manually checks on a data sample for each output variable. These comprise:
-visual inspection using ncview
-check whether min/max values are in physical range and consistent with variable units
-direction of flux variables and positive attribute
-consistency of data with vertical axis definition (i.e. whether data is flipped)
-land/ocean/sea ice masking
-rotation of vector data ;SQA - Scientific Quality Assurance 11/10/2011 10:28:03

Consistency: Quality Control Levels in CMIP5:
* Level 0: Spot checks on selected data
* Level 1: CMOR2 and ESG publisher conformance checks
* Level 2: Data consistency checks
* Level 3: Double- and cross-checks of data and metadata and data publication as DataCite DOI
QC Result Access: http://cera-www.dkrz.de/WDCC/CMIP5/QCResult.jsp?experiment=cmip5/output1/NCC/NorESM1-M/piControl

Completeness: [CMIP5:AuthorComment=We are still in the process of filling out the METAFOR CMIP5 questionnaire. Therefore its contents might be subject to change. Publications based on NorESM1-M are under preparation.]

Specification: [CMIP5:QualityLevel=3] [CMIP5:QualityFlag=approved by author] according to CMIP5 quality control rules http://cmip5qc.wdc-climate.de [CMIP5:DateOfQualityControl2=2011-10-10] [CMIP5:QualityControl2Comment=QC Level2 assigned at 2011-10-11 11:17:51 UTC: assignment confirmed by NCC (Ingo Bethke)] [CMIP5:CimMetadata=CMIP5 questionnaire data not found or not accessible] [CMIP5:TQA=done according to criteria for QCL3 defined at https://redmine.dkrz.de/collaboration/projects/cmip5-qc/wiki/Qc_l3#Criteria-for-QC-L3DOI-publication] [CMIP5:SQA=checked by author] [CMIP5:DateOfQualityControl3=2011-12-01]

Link to primary data

Links.jsp?acronym=NCCNMpc

Citation Reference

Contact Person

Metadata

Summary

Information

on Quality

Assurance

Direct Access

to Climate Data

Status QC for CMIP5

- QC Status CMIP5 (8. August 2013)
  - Quality Control 1: 1142 Experiments
  - Quality Control 2: 830 Experiments (finalised 403)
  - Quality Control 3: 174 Experiments
  - DataCite DOI: 116 Experiments (WDCC / IPCC-DDC)
    - RCPs, AMIP, Historical

CMIP5 data management achievements

- CMIP5 federation with 3 core data nodes (PCMDI, DKRZ, BADC), 16 index nodes and 23 data nodes operates an distributed archive of nearly 2 PB of climate model data which is an increase by a factor of 50 compared to the last CMIP in 2007.
- A searchable data catalogue is available across the federation.
- A description of climate models and experiments has been established.
- A three layer quality assurance process has been established which ends in a DataCite data publication for finalised reference data.
- Long-term archiving of reference data in the WDCC/DKRZ and integration in the ICSU WDS (World Data System) and the WIS (WMO Information System)
- Approved terms of use are available with open access for non-commercial use and 2/3 of the archive is available without any restrictions.

Future

- ESGF started to analyse the CMIP5 experiences in order to improve the ESGF data infrastructure:
  - Managing large data archives is not only a technical problem.
  - The establishment of a stable distributed ESGF infrastructure requires stable commitments and funding
- ESGF has requests from alternative modelling efforts and related observations to be included in ESGF in order to have all these data more easily inter-comparable.
- Federated data infrastructures like ESGF or Data Clouds seem the way to go for the next generation of climate data archives
  - CMIP3 to CMIP5: 36 TB to 1.8 PB, which means factor 50 increase
  - CMIP5 to CMIP6: 1.8 PB * 50 = 90 PB for one these MIPS
  - If a few or several of these MIPs are considered then ……
- Requested improvements
  - Usability of ESGF data access interface
  - Automated data replication between ESGF data nodes
  - More powerful, more stable and scalable wide area data networks (service level agreements)

Status QC for CMIP5

- QC Status CMIP5 (8. August 2013)
  - Quality Control 1: 1142 Experiments
  - Quality Control 2: 830 Experiments (finalised 403)
  - Quality Control 3: 174 Experiments
  - DataCite DOI: 116 Experiments (WDCC / IPCC-DDC)
    - RCPs, AMIP, Historical from NCC and MPI-M

Information on QC Status:
http://cera-www.dkrz.de/WDCC/CMIP5/QCResult.jsp

DataCite DOI Interface at WDCC (CMIP5 Reference Data Archive)

- Show me all data entities in the WDCC, which have an DataCite DOI
  - URL: http://cera-www.dkrz.de/WDCC/ui/FindDoiPublications.jsp

- Show me for a given CMIP5 NetCDF/CF dataset, whether it belongs to a DataCite published data entity
  - URL: http://cera-www.dkrz.de/WDCC/CMIP5/Citation.jsp

CMIP5 data management deficiencies

- ESGF data access interface did not match user requirements in all cases.
  - There was a major data interface release change after the first year (transition to ESGF Index nodes)
  - Data access felt to be slow, complicated and unstable
  - Often ESGF authentication and authorisation issues for new users (learning curve).
  - Data processing facilities were missing for example data reduction operations at data nodes.
  - Typical data processing workflows were not supported like multi-model multi-ensemble averages in IPCC WG-I.
- CMIP5 QC process was slower than expected because of exception handling
  - CMIP5 data publication and data update publication in ESGF is done in an uncoordinated and decentralized way whereas the CMIP5 QC activity was designed as a strongly coordinated process.
  - Scientific workflow only partly matched the data management concept. QC and scientific data evaluation started in parallel. Data are replaced due to the scientific process without notification of QC and replication.
  - A decentralized approach for CMIP5 QC was not feasible (resource limitations, coordination complexity, missing automatic software support, data node operational issues) thus the QC process is strongly interwoven with the replication process (at central data nodes)
  - Inconsistencies in CMIP5 data node operation hindered replication and qc
    - Inconsistencies between published data and data on disk (e.g. by replacing old versions “under the hood”)
    - Different data management policies at file system level at data centres