



- Two services for CMIP6 / AR6 provided by WDC Climate at DKRZ -

M. Stockhause and M. Lautenschlager German Climate Computing Center (DKRZ)









Data Citation Service

Concept/Workflow

The Data Citation Service is an integral part of the CMIP6 Infrastructure. Its concept is approved by the WGCM Infrastructure Panel.

CMIP6 Integration

The Data Citation Service is integrated into the ESGF CoG portal and the furtherInfoURL page.

Implementation

The implementation of the CMIP6 Citation Service consists of an Oracle database schema with a GUI, a landing page and a Restful API.

Scholix Integration

Scholix develops services to improve the integration of data citations into the scholarly publication environment.

Data Citation

IPCC DDC





Data Citation Service: Concept (1/5)

The Data Citation Service is an integral part of the CMIP6 Infrastructure approved by the WGCM Infrastructure Panel (WIP). An early data citation possibility was added for CMIP6.

Early Data Citation and IPCC DDC Data Citation:

CMIP6 data should be cited in scholarly publications using DataCite DOIs. It is recommended to cite data long-term archived in the IPCC DDC when available. The evolving CMIP6 data published and disseminated by the ESGF will be assigned early citations.

• Early Citation: Data gets citable with its early citation reference as soon as the first datasets

(prefix: 10.22033) are published in the ESGF and the mandatory citation information is available,

i.e. author list, title, and contact person. Citation references include the latest

dataset version or download date:

creators (publication year): title. version: YYYYMMDD. ESGF. doi.

• **DDC Citation:** Data will be long-term archived in the IPCC DDC at a snapshot date defined by the

(prefix: 10.1594) IPCC. This data subset of the CMIP6 data will be registered a DDC DOI. Citations

of long-term archived data are static and thus do not include versions:

creators (publication year): title. WDCC at DKRZ. doi.

Terms of Use:

• CMIP6 Data: https://pcmdi.llnl.gov/CMIP6/TermsOfUse

Data Citation Guidelines: http://bit.ly/2gBCuqM





Citation: Concept

Citation: Implementation

Citation: CMIP6
Integration

Citation: Scholix Integration







Data Citation Service: Concept (2/5)

The Data Citation Service is an integral part of the CMIP6 Infrastructure approved by the WGCM Infrastructure Panel (WIP). An early data citation possibility was added for CMIP6.

Citation Granularities:

Experiences from CMIP5 showed that the data citation of individual simulations contributed by an institute with a model is too fine-granular to be suitable for use in reference lists for some scholarly publications (i.e. number of data and article references are unbalanced). Therefore a second coarser data citation granularity was added for CMIP6 data citations.

The two provided data citation granularities for CMIP6 data are:

Coarse: Data of a model contributed by an institution to a CMIP6 MIP

• Fine: Data of a model contributed by an institution to a CMIP6 experiment

The CMIP6 data citation service is also provided for the CMIP6 forcing data of the input4MIPs project.

References:

- "CMIP6 Data Citation and Long-Term Archival" WIP white paper, http://bit.ly/23rsG8l.
- CMIP6 Citation Service: http://cmip6cite.wdc-climate.de
- WIP: https://www.earthsystemcog.org/projects/wip/
- input4MIPs: http://goo.gl/r8up31







Citation: Concept

Citation: Implementation

Citation: CMIP6
Integration

Citation: Scholix Integration









Data Citation Service: Concept (3/5)

The Data Citation Service is an integral part of the CMIP6 Infrastructure approved by the WGCM Infrastructure Panel (WIP). An early data citation possibility was added for CMIP6.

Data Citation Workflow (1) – CMIP6 Phase:

Precondition: CMIP6 CV: Registration of institution_id and source_id as CMIP6 Controlled Vocabulary (CV)

- 1. Early Citation / CMIP6 Phase:
 - a. CMIP6 Citation GUI: Provide mandatory citation information on authors, title and a contact person for the coarse model/MIP granularity via the CMIP6 Citation GUI.
 - b. ESGF: Prepare and ESGF-publish the first datasets
 - → Data Citation information is available in the ESGF CoG portal
 - → A DataCite DOI on the early citation will be registered, which remains stable while the CMIP6 data evolves. The CMIP6 data is citable.
 - c. CMIP6 Citation GUI: Maintain or add to the provided data citation information on coarse and fine citation granularities.

References:

CMIP6 CV: https://github.com/WCRP-CMIP/CMIP6 CVs

CMIP6 Citation GUI: http://cera-www.dkrz.de/citeXA

CMIP6 Citation: http://cmip6cite.wdc-climate.de







Citation: Concept

Citation: Implementation

Citation: CMIP6
Integration

Citation: Scholix Integration









Data Citation Service: Concept (4/5)

The Data Citation Service is an integral part of the CMIP6 Infrastructure approved by the WGCM Infrastructure Panel (WIP). An early data citation possibility was added for CMIP6.

Data Citation Workflow (2) – IPCC DDC Phase:

2. IPCC DDC Phase:

- a. CMIP6 Citation GUI: The snapshot date for closing the IPCC DDC Reference Data long-term Archive will be announced. Afterwards the data citation information can be finalized in the Citation GUI.
 - → Data Citation Information is no longer changeable.
- b. Long-Term Archival in the DDC: Data is moved from the CMIP6 data node repositories into the IPCC DDC long-term archive hosted at DKRZ. Available documenting ancillary metadata, e.g. ESDOC metadata, is added to the data and quality assurance is applied.
 - → The DDC mints DataCite DOIs on this snapshot and subset of CMIP6 data with references to the CMIP6 early citation DOI.
 - → The displayed data citation recommendation for users will be changed to using and citing the stable long-term archived IPCC DDC data.

References:

CMIP6 Citation GUI: http://cera-www.dkrz.de/citeXA

Data Citation Guidelines: http://bit.ly/2gBCuqM







Citation: Concept

Citation: Implementation

Citation: CMIP6
Integration

Citation: Scholix Integration









Data Citation Service: Concept (5/5)

The Data Citation Service is an integral part of the CMIP6 Infrastructure approved by the WGCM Infrastructure Panel (WIP). An early data citation possibility was added for CMIP6.

Future Plans:

1. User-defined citable data collections:

The current concept focusses on the credit idea. Thus for an identification of the exact datasets underlying a scholarly publication, the article has to be read. An extension of the data citation concept by a user-defined data collection citation is planned for CMIP7. The user then should cite the source CMIP6 data and the used data collection in a scholarly publication. CMIP policies are required for

- Authorship of data collections over several data citation entities
- Long-term availability of the data Integration in CMIP replication strategy

2. Federated data citation:

Different data nodes within the ESGF plan to or already register DataCite DOIs. For CMIP7 it is likely that the central data citation approach from CMIP5/AR5 and CMIP6/AR6 will be changed into a federated approach. One possibility is an ESGF membership of DataCite. An integration of the citation information gathering into the data preparation process would be required.



Citation: Concept

Citation: Implementation

Citation: CMIP6
Integration

Citation: Scholix Integration









Data Citation Service: Implementation (1/3)

The implementation of the CMIP6 Citation Service consists of an Oracle database schema with a Graphical User Interface, a landing page for displaying and a Restful API for machine-access of data citation information.

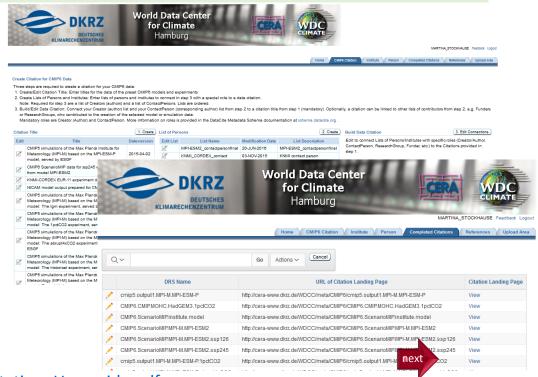
Graphical User Interface (Oracle APEX):

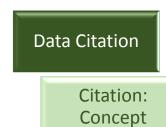
Every citation manager of the contributing modeling centers is provided with an account to maintain the data citation information. Author lists, titles and a contact person are mandatory. Information on contributors like funders or research groups or references to articles are optional.

Precondition to use the data citation service is the registration of institution_id and source_id (model acronym) for CMIP6.

GUI: http://cera-www.dkrz.de/citeXA

Tutorial: http://cera-www.dkrz.de/docs/pdf/CMIP6 Citation Userguide.pdf







Citation: CMIP6
Integration

Citation: Scholix Integration







Data Citation Service: Implementation (2/3)

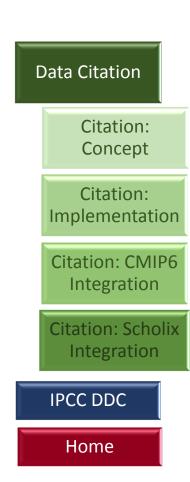
The implementation of the CMIP6 Citation Service consists of an Oracle database schema with a Graphical User Interface, a landing page for displaying and a Restful API for machine-access of data citation information.

Landing Page

The landing page displays the available data citation information ordered under different tabs.

The URL of this page gets registered as target page for the DataCite DOI, after the first data is ESGF-published and the mandatory data citation information is provided by the citation manager.









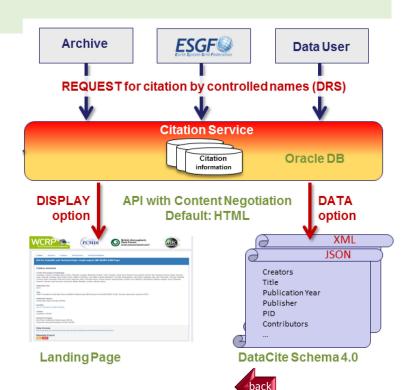
Data Citation Service: Implementation (3/3)

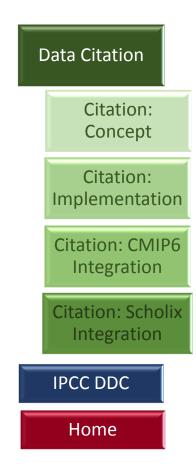
The implementation of the CMIP6 Citation Service consists of an Oracle database schema with a Graphical User Interface, a landing page for displaying and a Restful API for machine-

access of data citation information.

Restful API

For machine-access of the CMIP6 data citation information a Restful API is available. It provides information in DataCite 4.0 XML and JSON formats using content negotiation. E.g. the ESGF CoG portal display option for citation information is based on the JSON format (see next section on "Citation: CMIP6 Integration").





DataCite 4.0 Schema:

http://doi.org/10.5438/0012





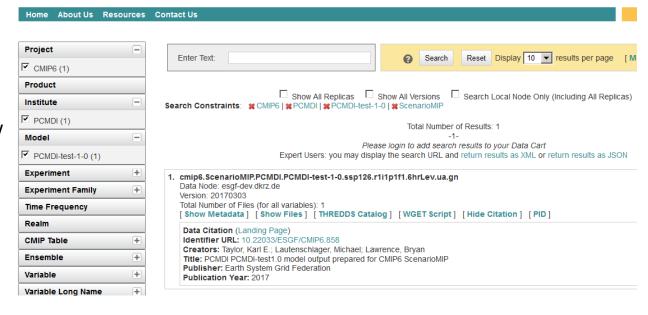


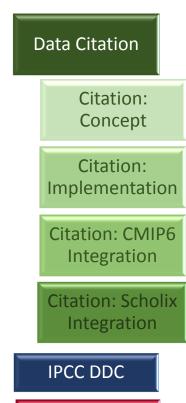
Data Citation Service: CMIP6 Integration (1/2)

Data Citation information is displayed in the ESGF CoG portal and integrated into the furtherInfoURL information page.

ESGF CoG Portal

Data users searching for and downloading CMIP6 data can access the connected data citation information via the 'Show Citation' link. Core citation information is displayed using the JSON format of the Restful API (see last section on "Citation: Implementation"). The provided URL gives access to the landing page displaying the complete data citation information.













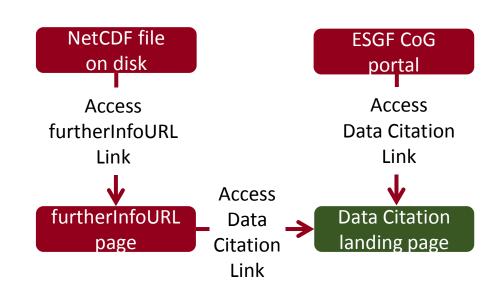
Data Citation Service: CMIP6 Integration (2/2)

Data Citation information is displayed in the ESGF CoG portal and integrated into the furtherInfoURL information page.

FurtherInfoURL Page

The furtherInfoURL is part of the global attributes of every NetCDF file. It points to a page summarizing information on ancillary metadata, such as ES-DOC, Errata, and Data Citation.

Thus users having already downloaded a file, can access the connected data citation information via the furtherInfoURL.













Data Citation Service: Scholix Integration

Scholix (Scholarly Link eXchange) develops services to improve the integration of data citations in the scholarly publication environment.

Scholix

The Scholix initiative is a high level interoperability framework for exchanging information about the links between scholarly literature and data. It aims to build an open information ecosystem to understand systematically what data underpins literature and what literature references data. Within the RDA/WDS Scholarly Link Exchange (Scholix) WG the services are currently developed.

The IPCC DDC at DKRZ supports Scholix. The CMIP6 Citation Service plans to be an early adopter of the Scholix services to provide statistics on CMIP6 data usage in literature for the CMIP6 project and the contributing modeling centers.

Scholix: http://www.scholix.org

RDA/WDS WG: https://www.rd-alliance.org/groups/rdawds-scholarly-link-exchange-scholix-wg

Four Key Use Cases



This is **Natalie**. She works for a publisher, managing their online platform. She wants to present links to relevant data to her users that are reading an article online.

Live linking



This is **Paula**. She's a researcher interested in scholarly publication and wants to be able to crawl the web of article/data links to find out if data citation is effective.

Exploration



This is **Mike**. He works for a data center, and he wants to be able to get an overview of all articles and data sets referencing/citing data sets managed by his data center.

Overview



This is **Jeff**. He works for a data center, and he wants to be notified as soon as one of the data sets managed by his data center is referenced or cited in the literature.

Notification

(source: WDS, 2016)

Data Citation

Citation: Concept

Citation: Implementation

Citation: CMIP6
Integration

Citation: Scholix Integration

IPCC DDC





IPCC Data Distribution Centre (IPCC DDC)

Role

The IPCC DDC facilitates wide availability of climate data and scenarios to enable research and sharing of information across the three IPCC working groups.

CMIP6 Data Pool

In support of and in collaboration with the IPCC Working Groups a CMIP6 Data Pool is set up at the IPCC DDC at DKRZ facilitating script-based data analyses.

Reference Archive

The IPCC DDC at DKRZ hosts the Reference Data Archive for the climate model output since the Second Assessment Report (1995) in support of IPCC's integrity.

IPCC Integration

The IPCC DDC is overseen by the Task Group on Scenarios for Climate and Impact Assessment (TGCIA). Its mandate is currently under review of an IPCC Adhoc Task Force.



IPCC DDC





IPCC DDC: Role and Mandate

The IPCC DDC facilitates wide availability of climate data and scenarios to enable research and sharing of information across the three IPCC working groups.

IPCC Data Distribution Centre (IPCC DDC):

The IPCC DDC is jointly managed by the British Atmospheric Data Centre (BADC) in the UK, the Center for International Earth Science Information Network (CIESIN) in the USA, and the World Data Center Climate (WDCC) at DKRZ in Germany. All three data centers are WDS (World Data System) members.

IPCC DDC hosted by WDCC/DKRZ:

The IPCC DDC hosted by WDCC at DKRZ is maintaining

- the Reference Data Archive for the climate model output since the Second Assessment Report (SAR). This is the data underlying the first parts of each AR. The long-term availability of the IPCC data contributes to IPCC's integrity.
- The CMIP6 Data Pool is set up as new DDC service for AR6 in cooperation with the three Technical Support Units (TSUs) of the Working Groups (WG) in order to improve the support for the IPCC WGs.

References:

P IPCC DDC: http://ipcc-data.org

IPCC DDC at DKRZ: http://ipcc.wdc-climate.de

Mandate: http://ipcc.ch/pdf/activity/tgica-mandate.pdf

WDS: http://icsu-wds.org/

Data Citation

IPCC DDC

IPCC DDC:

IPCC DDC: Reference Archive

IPCC DDC: CMIP6 Data Pool

IPCC DDC: Integration







IPCC DDC: Reference Data Archive (1/3)

The IPCC DDC at DKRZ hosts the Reference Data Archive for the climate model output since the Second Assessment Report (1995) in support of IPCC's integrity.

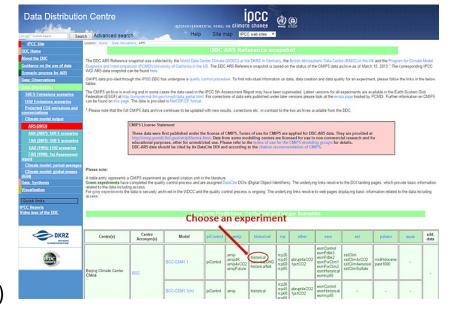
IPCC DDC Reference Data Archive:

After the end of each CMIP phase, a subset of the data is transferred into the long-term archive of the IPCC DDC at DKRZ and data-related information is added.

The DDC data is quality assured prior to long-term archival and a DataCite DOI citation reference is registered for the data. Data is stored with a second copy and curated on the long-term to ensure long-term availability and usability.

For **IPCC AR5** two data collections were archived with the snapshot date 15.03.2013:

- IPCC Working Group I snapshot (collected by the WG I)
- DDC Reference snapshot (disseminated by ESGF)









IPCC DDC: Reference Archive

IPCC DDC: CMIP6 Data Pool

IPCC DDC: Integration

Home

References:

- Reference Data Archive: http://ipcc-data.org/sim/gcm_monthly/
- IPCC DDC at DKRZ: http://ipcc.wdc-climate.de







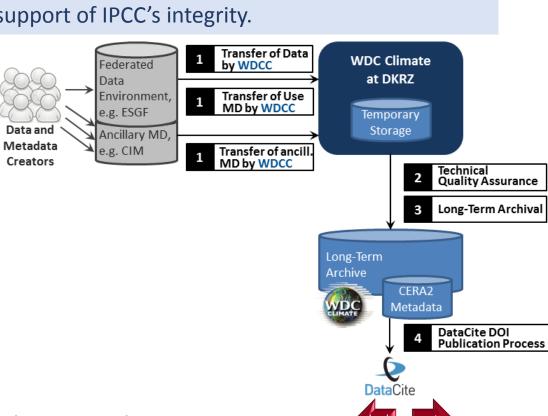
IPCC DDC: Reference Data Archive (2/3)

The IPCC DDC at DKRZ hosts the Reference Data Archive for the climate model output since the Second Assessment Report (1995) in support of IPCC's integrity.

Long-Term Archival Workflow:

On the IPCC DDC snapshot date the longterm archival workflow starts:

- Transfer of data with synda tool and ancillary metadata, e.g. experiment description and data citation details
- 2. Quality Assurance of data and metadata
- Long-Term data archival with storage of second data copy
- DDC data citation established by DataCite DOI registration





IPCC DDC

IPCC DDC:

IPCC DDC: Reference Archive

IPCC DDC: CMIP6 Data Pool

IPCC DDC: Integration

Home

References:

- Reference Data Archive: http://ipcc-data.org/sim/gcm_monthly/
- IPCC DDC at DKRZ: http://ipcc.wdc-climate.de







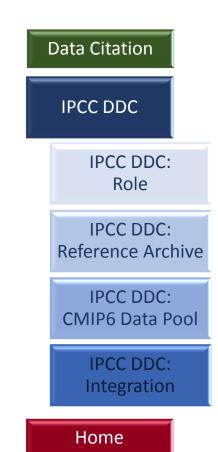
IPCC DDC: Reference Data Archive (3/3)

The IPCC DDC at DKRZ hosts the Reference Data Archive for the climate model output since the Second Assessment Report (1995) in support of IPCC's integrity.

Future Challenges for the Reference Data Archive:

- 1. Increasing data volume
- 2. Heterogeneity of data quality standards at the ESGF data nodes
- 3. Increasing number of repositories for ancillary metadata
- 4. Embedding DDC data in research environment, e.g. relations to data, researchers, scholarly publications etc.
- Broadening data user community needing different services and levels of assistance

Relations to Data CMIP Relations Relations **Data Centers with different** Citation Scientific **Quality and Curation Standards Publications Publications** on Different Granularities Researchers Scientific **Data Products** Forcing Data Credit and Institutions **Publications** Other Data Software Data Creator Raw Data Versions On Model On Simulation **Norkflows** Researcher Research Group On Data Hosting **IPCC Data** Collection Institution Project **Distribution Centre** Funder Data Analyst On Software / Workflow Editor ipcc 🐞 🙃 Publications Documentation **CMIP Relations** Software model output data and related and Workflows Relations **CMIP Data** Errata / in ESGF Annotation Model and Software Workflow Simulation Controlled **Provenance** Documentation Vocabulary Information Service **Data Citation** Quality available on file level. A connection to the data reference on simulation or to bridge the granularity gap. Lineage / CMIP (Inter-)Relations Institutions On Files On Collections? **Provenance** (source: EGU 2016-11315)



References:

- Reference Data Archive: http://ipcc-data.org/sim/gcm_monthly/
- IPCC DDC at DKRZ: http://ipcc.wdc-climate.de





IPCC DDC: CMIP6 Data Pool

In support of and in collaboration with the IPCC Working Groups a CMIP6 Data Pool is set up at the IPCC DDC at DKRZ facilitating script-based data analyses.

CMIP6 Data Pool – Motivation/Status:

For AR5 IPCC WG I set up a repository with a user-defined subset of the CMIP5 data. The main reasons were the complicated ESGF solutions for authentication/authorization and script-based data access and the non-transparent dataset versioning concept.

The IPCC DDC hosted at DKRZ will maintain a similar repository of CMIP6 data, the CMIP6 data pool, to support the IPCC AR6 WGs. This idea was first presented and discussed with the IPCC AR6 co-chairs at the IPCC Expert Meeting on the Future of TGICA in 01/2016 in Geneva and further developed at the Future Earth Workshop in 09/2016 in Stockholm.

For the CMIP6 Data Pool as DDC service a coordination on the technical level is currently established between the IPCC DDC at DKRZ and the IPCC AR6 WG TSUs. A coordination on the management level is about to be added involving the IPCC WG co-chairs.

References:

- CMIP6 Data Pool: https://redmine.dkrz.de/projects/dkrz cdp/wiki
- IPCC EM 01/2016: http://ipcc.ch/pdf/supporting-material/EMR TGICA Future.pdf



IPCC DDC: Integration







IPCC DDC: IPCC Integration (1/2)

The IPCC DDC is overseen by the Task Group on Scenarios for Climate and Impact Assessment (TGCIA). Its mandate is currently under review of an IPCC Adhoc Task Force.

Task Group on Scenarios for Climate and Impact Assessment (TGCIA):

TGICA was established in 1996. The IPCC Data Distribution Centre (IPCC DDC) is overseen by TGICA. TGICA comprises members with expertise and experience drawn from a cross section of research communities representing all three IPCC Working Groups. The DDC Managers are ex-officio members of TGICA.

Adhoc Task Force on TGICA (ATF-TGICA):

An Adhoc Task Force on TGICA was established at the 44th Session of the IPCC in October 2016 to design a strategic plan and revised mandate and terms of reference for the transformation of TGICA functions to serve the needs of IPCC during and beyond the Sixth Assessment. The Task Force will conclude its work in time for its proposals to be considered at the 46th Session of the IPCC in September 2017.

References:

IPCC TGICA: http://ipcc.ch/activities/tgica.shtml







IPCC DDC:

IPCC DDC: Reference Archive

IPCC DDC: CMIP6 Data Pool

IPCC DDC: Integration







IPCC DDC: IPCC Integration (2/2)

The technical coordination on processes and CMIP6 Data Pool content is directly performed by the IPCC DDC and the Technical Support Units (TSUs) of the three IPCC WGs.

Technical Coordination between IPCC DDC and IPCC WG TSUs:

Practical and technical issues for the new DDC service of the CMIP6 Data Pool require a direct coordination between the DDC hosted at DKRZ and WG TSUs, e.g. on

- CMIP6 Data Pool content,
- CMIP6 Data Pool services and support, and
- Transfer of CMIP6 Data Pool content into the AR6 Reference Data Archive.

The CMIP6 Data Pool is part of the CMIP6 replication strategy, which is in discussion in the WIP. Apart from IPCC requirements the CMIP6 data pool has to meet both institutional (e.g. available resources) and national demands (support of the national climate scientists). The DDC at DKRZ relies on institutional resources and national funding from the Bundesministerium für Bildung und Forschung (BMBF). Long-term funding is currently not guaranteed.

Data Citation IPCC DDC **IPCC DDC:** Role IPCC DDC: Reference Archive IPCC DDC: **CMIP6 Data Pool IPCC DDC:** Integration Home



Funding:

The IPCC DDC AR6 activities at DKRZ are funded by the Bundesministerium für Bildung und Forschung (BMBF; Federal Ministry of Education and Research) under the grant number: 01LP1605A.

