

IS-ENES3 Milestone M7.4

Complete ENES-CDI long term archival for CMIP6

Reporting period: 01/01/2022 – 31/03/2023

Authors:

Andrea Lammert (DKRZ), Fabian Wachsmann (DKRZ), Stephan Kindermann (DKRZ)

ABSTRACT

This report describes the work on long term archival of CMIP6 data collections to ensure their accessibility in the future and beyond the lifetime of the original data published as part of ESGF. As part of the ENES Climate Data Infrastructure DKRZ acts as a data archival center by transferring data to a tape backend and by exposing metadata as part of the World Data Center of Climate (WDCC). The CMIP6 subset used by the IPCC authors is long-term preserved in the IPCC Data Distribution Center (DDC). Additional CMIP6 data collections were archived and the archival process is still in progress as the new tape backend deployment at DKRZ during 2022 prevented a completion in 2022. During 2023 the archival of the ~5 PByte CMIP6 data subset which was collected, curated and made available as part of the DKRZ CMIP data pool will be completed (without IS-ENES funding). DKRZ has been long-term preserving the CMIP data underpinning the IPCC Assessments for 25 years, starting with the SAR. The long-term preservation in the DDC ensures the long-term availability and reusability of the CMIP data and provides a reference of past political decisions, especially after the end of the CMIP phase.

Dissemination Level	
PU	Public



Revision Table			
Version	Date	Name	Comments
1	22/02/2020	Stephan Kindermann	initial version created for collection of contributions
Final version	29/03/2023	Stephan Kindermann, Andrea Lammert	final version, updated with full list of IPCC CMIP6 collections

Table of contents

1. Introduction	3
2. CMIP6 input data subset of the DDC AR6 Reference Data Archive	4
3. Long term archival on tape	8
4. Integration into WDCC	8
Conclusion and Future work	8

1. Introduction

Data preservation not only refers to the long-term storage of data (archiving), but also includes ensuring the preservation and maintenance of data, as well as its context, understandability, interpretability, authenticity and integrity. All this applies in particular to the resulting datasets from CMIP6.

As part of the IPCC DDC Reference Data Archive for AR6, the CMIP6 data subset used by the IPCC authors and provided by the Technical Support Unit of Working Group I (WGI TSU) was transferred into DKRZ's long-term archive. DKRZ implemented a part of the IPCC FAIR Guidelines, enhancing the transparency of the IPCC results. Key aspects of the IPCC FAIR Guidelines (Pirani et al., 2022, <https://doi.org/10.5281/zenodo.6504469>) are the documentation of the figure creation process, citability of all digital objects used by the authors or created by the authors and the interlinking between the different IPCC products. Thus DKRZ added references to the AR6 WGI and their chapters as well as reference to the figure datasets long-term preserved at a DDC Partner. The general concept is outlined in Stockhause et al. (2019, doi:<http://doi.org/10.5334/dsj-2019-020>).

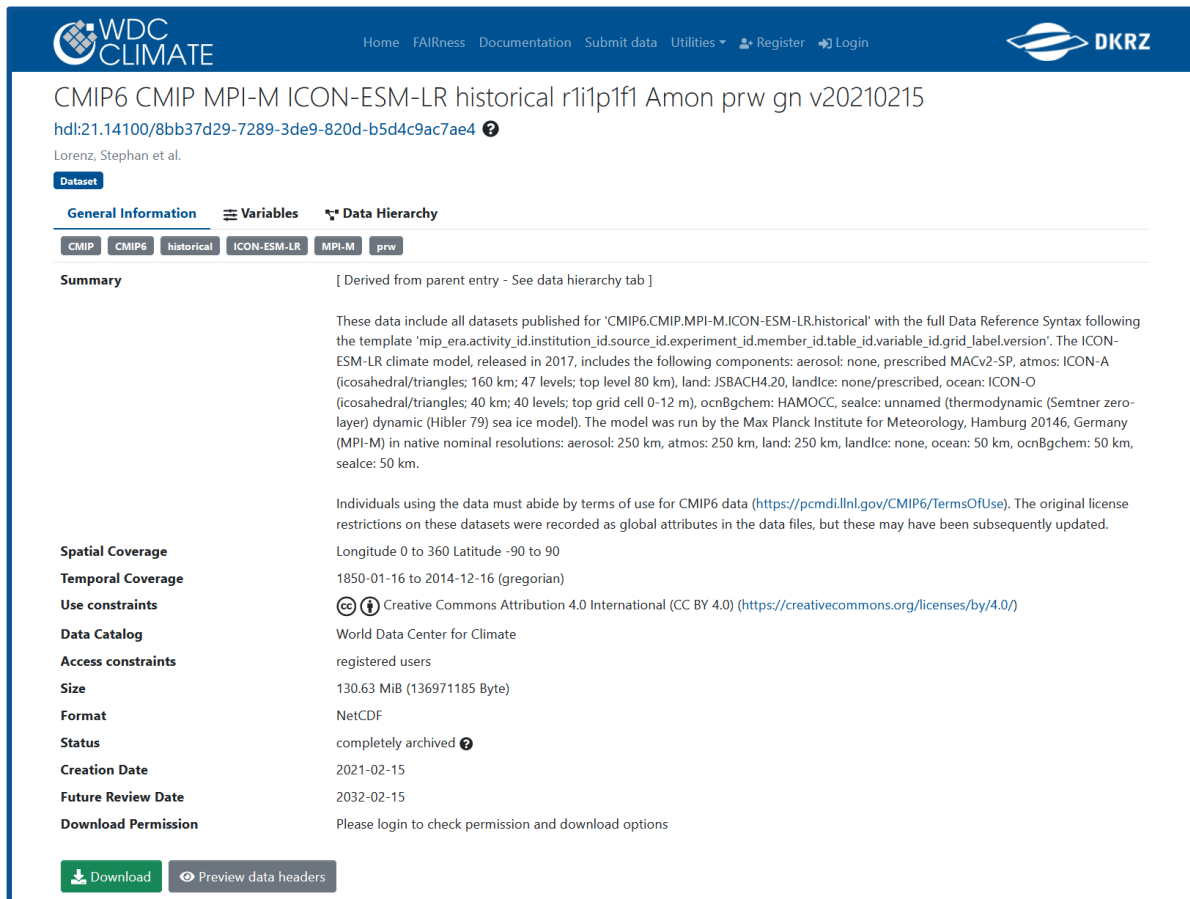
In addition to preserving the subset of the CMIP6 datasets used in the IPCC AR6, DKRZ has started to archive CMIP6 data on DKRZs long term tape storage and will continue this activity during 2023 (without further IS-ENES funding).

Integration of CMIP6 into the WDCC will ensure the usability and accessibility on the long term and for the wider research community. As mentioned before it was not possible to fully complete this task during the runtime of IS-ENES3 yet the operational data archival pipeline which was established during IS-ENES3 will ensure its completion during 2023.

2. CMIP6 input data subset of the DDC AR6 Reference Data Archive

The long-term preservation of the CMIP6 input datasets used in the AR6 builds the DDC AR6 Reference Data Archive together with intermediate datasets with a high reuse potential created by the authors and identified by the WGI TSU¹. In addition, DKRZ provided the Virtual Workspaces for the IPCC authors as collaboration platform with IS-ENES3 funding.

Technically, the WGI TSU provided CMIP6 dataset lists per chapter with information on the dataset and usage within the AR6 WGI. The dataset information within the list was used for data replication and metadata enrichment. Further metadata was gathered from the ESGF index and the CMIP6 Citation Service. The information provided by the Citation Service closed the information gap, which was present in AR5/CMIP5 and thus made the archival process much more straightforward. The Technical Quality Assurance (TQA) process of AR5 was rewritten and applied. It is planned to publish the DDC AR6 subset of CMIP6 through an ESGF Data Node.



WDC CLIMATE | Home | FAIRness | Documentation | Submit data | Utilities | Register | Login | DKRZ

CMIP6 CMIP MPI-M ICON-ESM-LR historical r1i1p1f1 Amon prw gn v20210215

hdl:21.14100/8bb37d29-7289-3de9-820d-b5d4c9ac7ae4

Lorenz, Stephan et al.

Dataset

General Information | Variables | Data Hierarchy

CMIP | CMIP6 | historical | ICON-ESM-LR | MPI-M | prw

Summary [Derived from parent entry - See data hierarchy tab]

These data include all datasets published for 'CMIP6.CMIP.MPI-M.ICON-ESM-LR.historical' with the full Data Reference Syntax following the template 'mip_era.activity_id.institution_id.source_id.experiment_id.member_id.table_id.variable_id.grid_label.version'. The ICON-ESM-LR climate model, released in 2017, includes the following components: aerosol: none, prescribed MACv2-SP, atmos: ICON-A (icosahedral/triangles; 160 km; 47 levels; top level 80 km), land: JSBACH4.20, landIce: none/prescribed, ocean: ICON-O (icosahedral/triangles; 40 km; 40 levels; top grid cell 0-12 m), ocnBgchem: HAMOCC, sealce: unnamed (thermodynamic (Semtner zero-layer) dynamic (Hibler 79) sea ice model). The model was run by the Max Planck Institute for Meteorology, Hamburg 20146, Germany (MPI-M) in native nominal resolutions: aerosol: 250 km, atmos: 250 km, land: 250 km, landIce: none, ocean: 50 km, ocnBgchem: 50 km, sealce: 50 km.

Individuals using the data must abide by terms of use for CMIP6 data (<https://pcmdi.llnl.gov/CMIP6/TermsOfUse>). The original license restrictions on these datasets were recorded as global attributes in the data files, but these may have been subsequently updated.

Spatial Coverage	Longitude 0 to 360 Latitude -90 to 90
Temporal Coverage	1850-01-16 to 2014-12-16 (gregorian)
Use constraints	Creative Commons Attribution 4.0 International (CC BY 4.0) (https://creativecommons.org/licenses/by/4.0/)
Data Catalog	World Data Center for Climate
Access constraints	registered users
Size	130.63 MiB (136971185 Byte)
Format	NetCDF
Status	completely archived
Creation Date	2021-02-15
Future Review Date	2032-02-15
Download Permission	Please login to check permission and download options

[Download](#) | [Preview data headers](#)

Figure 1: Landing page for a CMIP dataset as part of the DDC AR6 Reference Data Archive.

¹ Intermediate datasets in the DDC AR6 Reference Data Archive are accessible at: https://www.wdc-climate.de/ui/q?hierarchy_steps_ss=IPCC-DDC_AR6_Supplements

The data archival of the intermediate datasets is completed, the TQA and the DOI registration have to be finalized. The archival of the CMIP6 input datasets is nearly completed with the exception of a few single datasets (example given in Fig. 1). The list of all data collections² based on the CMIP6 Activity, Institution_id, and Source_ID is given in Table 1.

CMIP6 Activity	Institution_ID	Source_ID
CMIP6 AerChemMIP	BCC BCC-ESM1 CNRM-CERFACS CNRM-ESM2-1 EC-Earth-Consortium EC-Earth3-AerChem HAMMOZ-Consortium MPI-ESM-1-2-HAM MIROC MIROC6 MOHC UKESM1-0-LL	MRI MRI-ESM2-0 NASA-GISS GISS-E2-1-G NCAR CESM2-WACCM NCC NorESM2-LM NERC UKESM1-0-LL NIMS-KMA UKESM1-0-LL NOAA-GFDL GFDL-ESM4
CMIP6 CDRMIP	CCCma CanESM5 CNRM-CERFACS CNRM-ESM2-1 CSIRO ACCESS-ESM1-5 MIROC MIROC-ES2L	MOHC UKESM1-0-LL NCAR CESM2 NCC NorESM2-LM
CMIP6 CMIP	AS-RCEC TaiESM1 AWI AWI-CM-1-1-MR AWI AWI-ESM-1-1-LR BCC BCC-CSM2-MR BCC BCC-ESM1 CAMS CAMS-CSM1-0 CAS CAS-ESM2-0 CAS FGOALS-f3-L CAS FGOALS-g3 CCCma CanESM5 CCCma CanESM5-CanOE CCCR-IITM IITM-ESM CMCC CMCC-CM2-HR4 CMCC CMCC-CM2-SR5 CMCC CMCC-ESM2 CNRM-CERFACS CNRM-CM6-1 CNRM-CERFACS CNRM-CM6-1-HR CNRM-CERFACS CNRM-ESM2-1 CSIRO ACCESS-ESM1-5 CSIRO-ARCCSS ACCESS-CM2 EC-Earth-Consortium EC-Earth3 EC-Earth-Consortium EC-Earth3-AerChem EC-Earth-Consortium EC-Earth3-CC EC-Earth-Consortium EC-Earth3-LR EC-Earth-Consortium EC-Earth3P-VHR	IPSL IPSL-CM5A2-INCA IPSL IPSL-CM6A-LR IPSL IPSL-CM6A-LR-INCA KIOST KIOST-ESM MIROC MIROC-ES2H MIROC MIROC-ES2L MIROC MIROC6 MOHC HadGEM3-GC31-LL MOHC HadGEM3-GC31-MM MOHC UKESM1-0-LL MPI-M ICON-ESM-LR MPI-M MPI-ESM1-2-HR MPI-M MPI-ESM1-2-LR MRI MRI-ESM2-0 NASA-GISS GISS-E2-1-G NASA-GISS GISS-E2-1-G-CC NASA-GISS GISS-E2-1-H NASA-GISS GISS-E2-2-G NCAR CESM2 NCAR CESM2-FV2 NCAR CESM2-WACCM NCAR CESM2-WACCM-FV2 NCC NorCPM1 NCC NorESM1-F NCC NorESM2-LM NCC NorESM2-MM

² List of all archived CMIP6 input datasets in the IPCC DDC AR6 Reference Data Archive is available at https://www.wdc-climate.de/ui/q?hierarchy_steps_ss=IPCC-AR6_CMIP6

	<p>EC-Earth-Consortium EC-Earth3-Veg EC-Earth-Consortium EC-Earth3-Veg-LR E3SM-Project E3SM-1-0 E3SM-Project E3SM-1-1 E3SM-Project E3SM-1-1-ECA FIO-QLNM FIO-ESM-2-0 HAMMOZ-Consortium MPI-ESM-1-2-HAM INM INM-CM4-8 INM INM-CM5-0</p>	<p>NIMS-KMA KACE-1-0-G NIMS-KMA UKESM1-0-LL NOAA-GFDL GFDL-AM4 NOAA-GFDL GFDL-CM4 NOAA-GFDL GFDL-ESM4 NUIST NESM3 SNU SAM0-UNICON THU CIESM UA MCM-UA-1-0</p>
CMIP6 C4MIP	<p>BCC BCC-CSM2-MR CCCma CanESM5 CNRM-CERFACS CNRM-ESM2-1 CSIRO ACCESS-ESM1-5 E3SM-Project E3SM-1-1 E3SM-Project E3SM-1-1-ECA IPSL IPSL-CM6A-LR</p>	<p>MIROC MIROC-ES2L MOHC UKESM1-0-LL MPI-M MPI-ESM1-2-LR MRI MRI-ESM2-0 NCAR CESM2 NCC NorESM2-LM NOAA-GFDL GFDL-ESM4</p>
CMIP6 DAMIP	<p>BCC BCC-CSM2-MR CAS FGOALS-g3 CCCma CanESM5 CNRM-CERFACS CNRM-CM6-1 CSIRO ACCESS-ESM1-5 EC-Earth-Consortium EC-Earth3 IPSL IPSL-CM6A-LR MIROC MIROC-ES2L MIROC MIROC6</p>	<p>MOHC HadGEM3-GC31-LL MOHC UKESM1-0-LL MPI-M MPI-ESM1-2-LR MRI MRI-ESM2-0 NASA-GISS GISS-E2-1-G NCAR CESM2 NCC NorESM2-LM NOAA-GFDL GFDL-CM4 NOAA-GFDL GFDL-ESM4</p>
CMIP6 GMMIP	<p>BCC BCC-CSM2-MR CAMS CAMS-CSM1-0 CAS FGOALS-f3-L CCCma CanESM5 CNRM-CERFACS CNRM-CM6-1 CNRM-CERFACS CNRM-ESM2-1</p>	<p>IPSL IPSL-CM6A-LR MIROC MIROC6 MRI MRI-ESM2-0 NCAR CESM2 NOAA-GFDL GFDL-CM4</p>
CMIP6 HighResMIP	<p>AWI AWI-CM-1-1-HR AWI AWI-CM-1-1-LR BCC BCC-CSM2-HR CMCC CMCC-CM2-HR4 CMCC CMCC-CM2-VHR4 CNRM-CERFACS CNRM-CM6-1 CNRM-CERFACS CNRM-CM6-1-HR EC-Earth-Consortium EC-Earth3P EC-Earth-Consortium EC-Earth3P-HR ECMWF ECMWF-IFS-HR ECMWF ECMWF-IFS-LR ECMWF ECMWF-IFS-MR</p>	<p>INM INM-CM5-H MOHC HadGEM3-GC31-HH MOHC HadGEM3-GC31-HM MOHC HadGEM3-GC31-LL MOHC HadGEM3-GC31-MM MPI-M MPI-ESM1-2-HR MPI-M MPI-ESM1-2-XR NCAR CESM1-CAM5-SE-HR NCAR CESM1-CAM5-SE-LR NERC HadGEM3-GC31-HH NERC HadGEM3-GC31-HM NOAA-GFDL GFDL-CM4C192</p>
CMIP6 LS3MIP	<p>CNRM-CERFACS CNRM-CM6-1 CNRM-CERFACS CNRM-ESM2-1 IPSL IPSL-CM6A-LR MIROC MIROC6</p>	<p>MPI-M MPI-ESM1-2-LR NASA-GISS GISS-E2-1-G NCAR CESM2</p>

CMIP6 LUMIP	BCC BCC-CSM2-MR MOHC HadGEM3-GC31-LL MOHC UKESM1-0-LL	
CMIP6 OMIP	CAS FGOALS-f3-L NCAR CESM2	
CMIP6 PMIP	AWI AWI-ESM-1-1-LR CAS FGOALS-f3-L CAS FGOALS-g3 CNRM-CERFACS CNRM-CM6-1 CSIRO ACCESS-ESM1-5 EC-Earth-Consortium EC-Earth3-LR INM INM-CM4-8 IPSL IPSL-CM6A-LR MIROC MIROC-ES2L	MPI-M MPI-ESM1-2-LR MRI MRI-ESM2-0 NASA-GISS GISS-E2-1-G NCAR CESM2 NCC NorESM1-F NCC NorESM2-LM NERC HadGEM3-GC31-LL NUIST NESM3
CMIP6 ScenarioMIP	AS-RCEC TaiESM1 AWI AWI-CM-1-1-MR BCC BCC-CSM2-MR CAMS CAMS-CSM1-0 CAS CAS-ESM2-0 CAS FGOALS-f3-L CAS FGOALS-g3 CCCma CanESM5 CCCma CanESM5-CanOE CCCR-IITM IITM-ESM CMCC CMCC-CM2-SR5 CMCC CMCC-ESM2 CNRM-CERFACS CNRM-CM6-1 CNRM-CERFACS CNRM-CM6-1-HR CNRM-CERFACS CNRM-ESM2-1 CSIRO ACCESS-ESM1-5 CSIRO-ARCCSS ACCESS-CM2 DKRZ MPI-ESM1-2-HR DWD MPI-ESM1-2-HR EC-Earth-Consortium EC-Earth3 EC-Earth-Consortium EC-Earth3-AerChem EC-Earth-Consortium EC-Earth3-CC EC-Earth-Consortium EC-Earth3-Veg EC-Earth-Consortium EC-Earth3-Veg-LR E3SM-Project E3SM-1-1 FIO-QLNM FIO-ESM-2-0	HAMMOZ-Consortium MPI-ESM1-2-HAM INM INM-CM4-8 INM INM-CM5-0 IPSL IPSL-CM5A2-INCA IPSL IPSL-CM6A-LR KIOST KIOST-ESM MIROC MIROC-ES2L MIROC MIROC6 MOHC HadGEM3-GC31-LL MOHC HadGEM3-GC31-MM MOHC UKESM1-0-LL MPI-M MPI-ESM1-2-LR MRI MRI-ESM2-0 NASA-GISS GISS-E2-1-G NCAR CESM2 NCAR CESM2-WACCM NCC NorESM2-LM NCC NorESM2-MM NIMS-KMA KACE-1-0-G NIMS-KMA UKESM1-0-LL NOAA-GFDL GFDL-CM4 NOAA-GFDL GFDL-ESM4 NUIST NESM3 THU CIESM UA MCM-UA-1-0

Table 1: List of all data collections of the IPCC-DDC subset, based on the combination of CMIP6 Activity, Institution_ID, and Source_ID stored, published and available at the WDCC.

3. Long term archival on tape

In 2020 and 2021, the DKRZ backed-up its primary published CMIP6 data set of 1.5PB in the archive system hpss. This data set is freely accessible by all DKRZ users. It mainly contains earth system model data produced by AWI's and MPI-M's in-house models AWI-CM and MPI-ESM1-2 in various configurations.

After a migration phase, DKRZ's new tape archival system installed during 2022 was finally operationally ready in late January 2023 to carry out operational jobs of magnitude 10TB/day per job. As one of the pioneer flagship activities the back-up of the remaining large volume of 2.5 PB of replicated CMIP6 data could be initiated. As an interim result, additional 200TB were successfully achieved as early as March 2023 thanks to this continuous data transfer.

Upon completion, the entire archived CMIP6 data set will be supported with an intake catalog to simplify data access. Additional measures like tests of checksums will be taken to ensure complete and correct data archival.

4. Integration into WDCC

Integration of the CMIP6 data collections archived on tape into WDCC is an ongoing task. Due to delays in the provision of the CMIP6 input dataset list by the WGI TSU and its poor quality, the DDC AR6 Reference Data archival took longer than expected. Nevertheless, the workflows for metadata ingest from ESGF and ingest of citation information from CMIP6 citation service into WDCC are well established, as well as the archiving procedures. Thus the integration of the remaining tape archived data collections into the WDCC is ensured and will be completed mostly in 2023.

Conclusion and Future work

An evolving collection of CMIP6 data was archived during 2022 thus ensuring the accessibility of important and often used CMIP6 datasets beyond the lifetime of the original datasets published on ESGF. The archival process and exposure of associated metadata as part of the WDCC is nearly completed for the IPCC DDC AR6 Reference Data Archive. The archival process now continues with the CMIP6 subset in the DKRZ CMIP data pool (nearly 5 PBytes of data) and which was exploited as part of the IS-ENES3 data and compute service activities.