DKRZ CMIP Data Pool

Motivation

The DKRZ CMIP data pool (DKRZ CDP) provides often needed collections of climate model data in the context of climate model intercomparison and evaluation projects. This pool is hosted as part of the German Climate Computing Center (DKRZ) data infrastructure and wants to support user groups in high volume climate data collection, access and processing. The hosted data concentrates on model data generated as part of larger climate model intercomparison projects e.g. CMIP and CORDEX.
This document summarizes the initial service offerings around this data pool (especially with respect to data ingest and data processing) and will be updated regularly to reflect the evolution of these services.

Registration

- To access the pool data, the user needs to be register at DKRZ, see "Register a new website account."
  - Registration can be done at https://luv.dkrz.de.
  - This registration assigns a DKRZ web account.
- After registration users need to join specific groups to be assigned to personal storage and compute resources. Please, select the specific group

DKRZ_MIP_POOL_Analysis (number 1088) by clicking on “join existing project” here https://luv.dkrz.de/projects/ or contacting data-pool@dkrz.de.

Data pool content

The overall data pool provides space for around 5 PBytes of data stored as part of the DKRZ HPC Lustre file system. A dedicated board decides on the prioritization of data storage based on end user requirements as well as international agreements (e.g. with respect to data replication). Based on current estimations around 2 PByte are reserved for providing replicated CMIP data from other ESGF data nodes around the world. Around 100 TByte are currently reserved for storing derived data products.

In case you miss some ESGF CMIP(6) data available at other ESGF nodes you can request a data replication by contacting esgf-replication@dkrz.de. You will be provided with a password to access an interactive data request form. The specification of the specific data collection you want to be replicated and made locally accessible is based on synda selection files (http://prodiguer.github.io/synda/).

In case you have requirements with respect to storage of derived data products or the inclusion of non-ESGF accessible data sets please contact data-pool@dkrz.de.

Data access from pool associated resources

The data pool is efficiently accessible from HPC resources as well as virtual machines:

- HPC resources:

  See https://www.dkrz.de/up/my-dkrz/getting-started/getting-started-at-dkrz for an introduction

  - Users can directly login into front end nodes of the DKRZ HPC system (“login nodes”) and run interactive processing scripts. The data pool is directly accessible under local directory paths (directly mounted):
    - **/work/kd0956** holds around 1.2 PBytes of CMIP5 data and CORDEX data
    - **/work/ik1017** will provide access to around 3 PBytes of CMIP6 data
  - compute intensive parallel data analysis is supported by the submission of batch compute jobs to the DKRZ HPC computer. The pool data is accessible from all compute nodes. The need for such compute intensive compute loads needs to be requested in the application as well as the amount of cpu time.
- Virtual machines: Users can request virtual machines with direct access to the data pool. Requests need to be sent to data-pool@dkrz.de. These requests are evaluated and the

resources are granted based on the result of this evaluation. Users can then directly log into these virtual machines and install and run their analysis codes.

An overview of the pre-installed libraries and tools please refer to

- How to compile and run your programs.
- Pre-installed software list

Data access via Jupyter hub:

DKRZ currently provides a pre-production jupyter hub service for interested users : https://jupyterhub.dkrz.de
Users need a DKRZ account and associated project assignment (see above) to launch jupyter notebooks with different dedicated amounts of CPU resources. The notebooks are running in the DKRZ HPC environment - thus the CMIP data pool is directly accessible as described above.
Documentation is available at https://www.dkrz.de/up/de-systems/de-jupyterhub-dkrz.de-1 and there is also a demo notebook.

Data access from remote resources

Data from the data pool can be replicated to remote e.g. institutional resources using different methods:

- Use the synda replication tool (recommended as this method supports consistent version updates etc. to keep the copy in sync with the pool content)
- Use the ESGF interfaces directly (wget - http, or gridftp)
- Direct copy via rsync, gridftp etc.

Data search

There are different possibilities to search for data items in the data pool:

- Use the replica search interface hosted at https://cmip-esmvaltool.dkrz.de/solr/data-browser/ (use the guest login tab)
- Use the search index on the login nodes
    - Log into the interactive nodes at DKRZ
    - Module load cmip6-dicad/1.0
    - freva --databrowser --help
- Use the DKRZ ESGF portal http://esgf-data.dkrz.de and click on the CMIP6 Data Search link in the “Search all projects” section. Once you are there, select “show all

replicas” as well as “esgf3.dkrz.de” from the “data node” tab. This will show all CMIP6 data replicated to DKRZ which is already re-published to ESGF (thus a subset of the data found in the first option above (using the data-browser), which also included not yet republished replicas available at DKRZ.

Technical information with respect to the DKRZ CMIP Data Pool

Data store:

- A ~5 PByte disk pool, hosted at DKRZ as part of the DKRZ HPC Lustre storage system

Compute resources:

- Dedicated compute servers
- Virtual machines

User support:

- data-pool@dkrz.de

Status

- The data pool is currently containing
    - a near complete copy of CMIP5 data (~1.2 Petabyte) as well as a collection of CORDEX data, besides some CMIP related projects (Obs4Mips,..)
    - ~ 600 TByte CMIP6 replicas (beginning of June 2019), growing ~ 50 TByte per week
- For dkrz users the data is accessible in the file system at **/work/kd0956** (CMIP5) and **/work/ik1017** (CMIP6)