Project: **1017**

Project title: **Konsortialdatenprojekt CMIP-DKRZ-Datenpool**

Principal investigator: **Stephan Kindermann**

Report period: **2023-11-01 to 2024-10-31**

The storage resources allocated to the CMIP-DKRZ-Datenpool (5 PByte) are fully booked with the national CMIP6 contributions and replicated core data collections from the overall global CMIP6 data holding.

The data pool content on one hand builds up the DKRZ storage space accessible via the global ESGF data federation based on ESGF data nodes and the ESGF portal deployed at DKRZ (https://esgf-data.dkrz.de) and provides the largest CMIP data space integrated into the European ENES data infrastructure. Fig 1a and Fig 1b show the number of downloads of CMIP6 data from the ESGF.

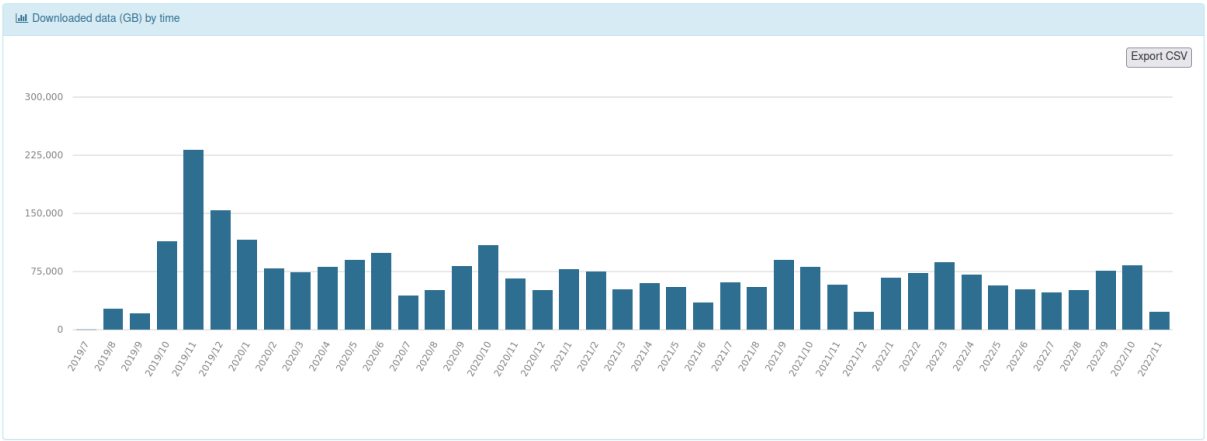

Fig 1a: Downloaded CMIP6 data hosted at DKRZ by size in GB over time.

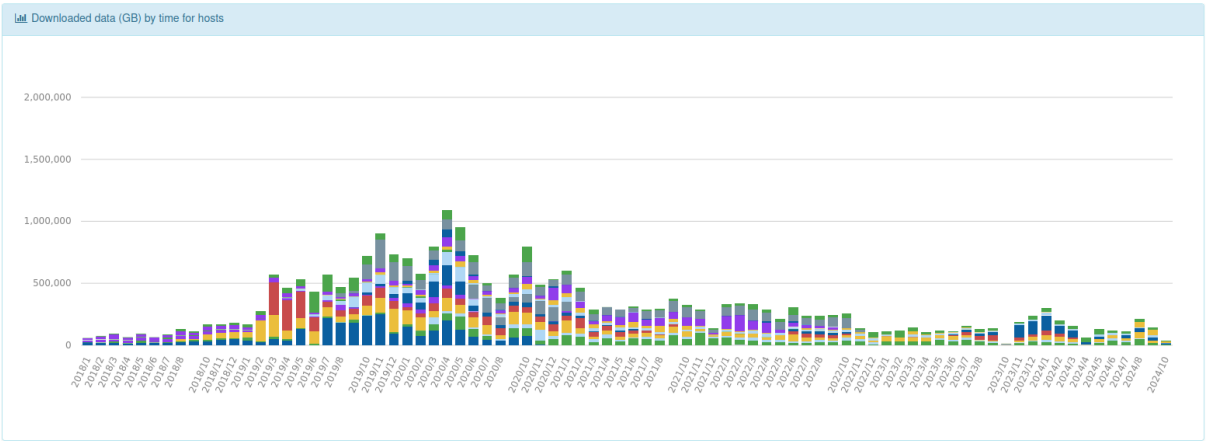

Fig 1b: Downloaded CMIP6 data by size in GB over time for all ESGF nodes.

On the other hand, the pool is accessible to all DKRZ users (via /pool/data). The direct exploitation of this data pool based on the DKRZ HPC resources is supported by:

- Generation and provisioning of intake data catalogs, directly usable in jupyter notebooks (e.g. by using the DKRZ jupyter hub)

- Provisioning of detailed documentation with updates communicated via DKRZ documentation[1] and the DKRZ Blog[2]

During the reporting period the management of the data pool content (replication, update of data collections, removal of data collections) was driven by different requirements:

- Satisfying the requests of researchers to include additional data replicas into the data pool to support their analysis activity.
- Updating the data pool by removing retracted data and updating data with new versions of CMIP6 data.

For non DKRZ users and to support coordinated exploitation of the data pool at the European level (supported by the IS-ENES3 project) the DKRZ project 1088 was heavily used and showed interest from various user groups:

- Climate model evaluation activities (e.g. using the ESMValTool)
- Climate impact researchers interested generating derived data products
- Climate service center members and commercial climate service providers exploring the possibilities of having a direct access to CMIP6 via jupyterhub.
- Interdisciplinary user groups in the context of projects related to the European Open Science cloud.
- Other DKRZ projects making use of the CDP include ClimExtreme, DAKI RegIKlim, CLINT and CLICCs etc.

Kerchunk

First efforts have been made to generate a data cube representation of CDP highlights. This means, single files are virtually aggregated to single large datasets that can be accessed with the zarr library. The resulting virtual representations are called ‘kerchunks’, are stored in json or parquet files, simplify the data access and reduce preparation tasks on the user side. Variables which share dimensions and are sampled on the same frequency, are merged and concatted on their time dimension. For an ensemble, a corresponding ‘ensemble’ dimension is newly inserted which collects the single realizations of an experiment. The resulting dataset allows users to access 10k underlying files and terabytes of data through one virtual representation.

The kerchunks are openly accessible under /work/ik1017/CMIP6/meta/kerchunk . In the future, a kerchunk architecture that integrates well with stac catalogs is planned for ESGF and thus the CDP.

[1] DKRZ CDP documentation: https://cmip-data-pool.dkrz.de

[2] DKRZ CDP blog: z.B. https://blog.dkrz.de/dkrz-cdp-updates-july-21.html