Project: **1017**

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

Principal investigator: **Stephan Kindermann**

Report period: **2024-11-01 to 2025-10-31**

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.

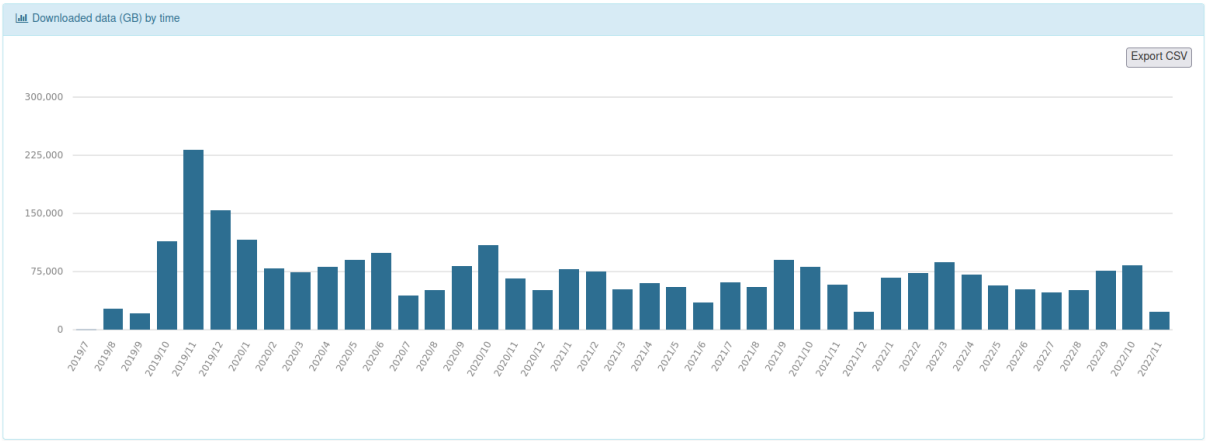

Fig 1a: Downloaded CMIP6 data hosted at DKRZ by size in GB over time.

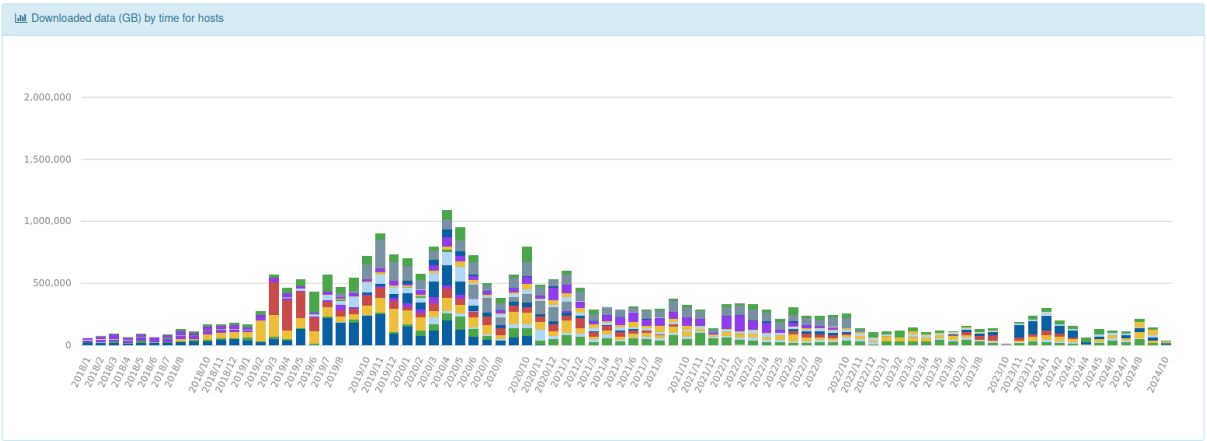

Fig 1b: Downloaded CMIP6 data by size in GB over time for all ESGF nodes.

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.

Archival, CMIP7 preparation and US-Backup

More than 90% of all CMIP6 Pool datasets have been successfully archived in WDCC, enabling a significant release of disk space.To simplify user access, a Python-based client to the WDCC API, detailed notebook guidance for its use, and a browser-based application for dataset exploration and retrieval were developed and provided. Together, these tools simplify workflows and enhance discoverability of archived climate data not limited to CMIP.

In preparation for CMIP7, targeted cleanup and optimization activities are underway. Rarely accessed CMIP6 content not required by the community is being removed to free storage for incoming CMIP7 datasets. In parallel, support for *intake-esgf* - a lightweight replication tool - has been established, with practical notebook guidance provided and integration into the replication service completed. Automatic weekly replication of INPUT4MIPs and OBS4MIPs data was set-up and users find it on Levante under /work/kd0956/INPUT4MIPS/data/input4MIPs/ and in ESGF.

An ongoing collaborative effort with European ESGF node partners focuses on replicating up to 1 PB of CMIP6 data currently hosted only on US nodes. This initiative strengthens data availability and fail-safety within the distributed ESGF federation, ensuring robust long-term access to critical climate model outputs.

---

[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