Project: **1017**

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

Principal investigator: **Stephan Kindermann**

Report period: **2020-11-01 to 2021-08-31**

The storage resources allocated to the CMIP-DKRZ-Datenpool (5 PByte) are nearly fully booked with the national CMIP6 contributions and replicated core data collections from the overall global CMIP6 data holding (currently ~12 PByte). The following figures try to illustrate the current status (a dataset corresponds to a complete time series of a specific variable, thus datasets correspond to a set of netcdf files):

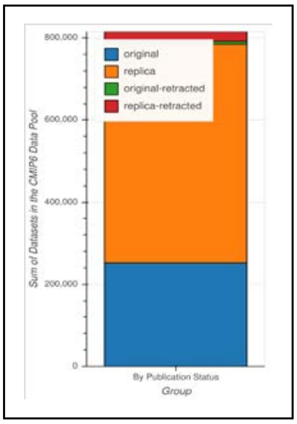


Fig 1: An internationally agreed and communicated subset of CMIP6 is re-published into ESGF as “replica”. Modeling groups continue to update and correct data, outdated data is “retracted” and after some time removed from the pool. Currently the datasets sum up to more then 4 PBytes.

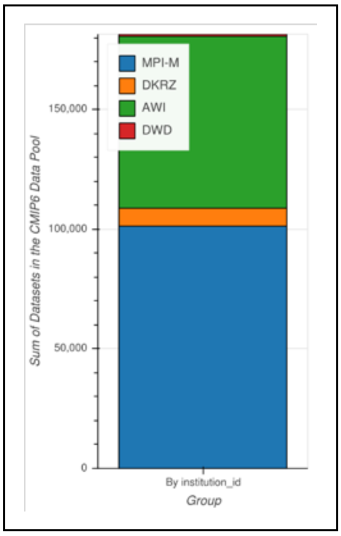
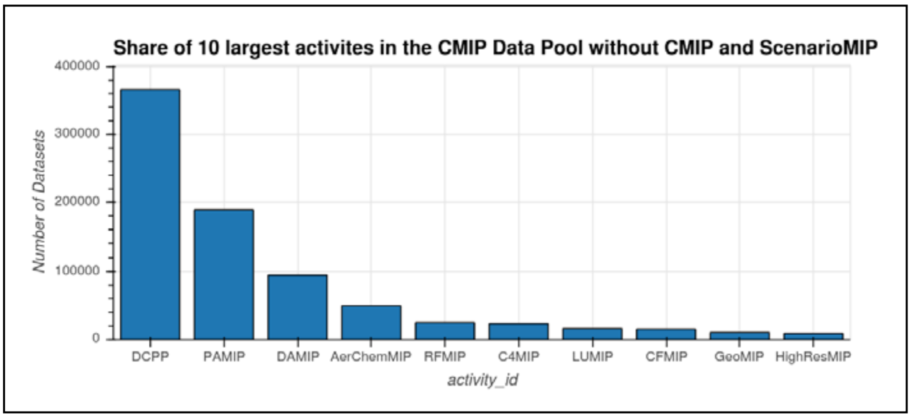


Fig 2: The CMIP6 pool contains all German CMIP6 contributions. This figure illustrates the number of datasets made available by the individual German CMIP6 modeling groups. The 180.000 datasets correspond to around 1.5 PByte.



The data pool content on one hand is 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. 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 was strongly improved during the reporting period by:

- Automatic generation and provisioning of intake data catalogs, directly usable in jupyter notebooks (e.g. by using the DKRZ jupyter hub)
- Provisioning of detailed documentation which is continuously updated, major updates of the pool are communicated via the DKRZ Blog [1] and documentation is published in the DKRZ documentation site [2].
- Providing a Training Event for the European Research Community (jointly organized with the European IS-ENES3 and EOSC-hub projectes) [3] and providing a dedicated DKRZ tech talk [4].

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.

As the DKRZ CMIP data pool content is directly accessible to all DKRZ users, no detailed overview can be given on the different analysis activities which were supported during the reporting period. 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:

- IPCC working group members, preparing figures for the IPCC report
- 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.

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

[2] DKRZ CDP documentation: https://doc.dkrz.de/doc/cmip-data-pool/index.html

[3] CDP data analysis training event: https://www.dkrz.de/de/kommunikation/aktuelles/training-datenanalytik

[4] DKRZ data pool tech talk: https://www.dkrz.de/up/news-and-events/tech-talks/dkrz-tech-talk-the-cmip-data-pool