

EGU21-2442, updated on 06 Dec 2021

<https://doi.org/10.5194/egusphere-egu21-2442>

EGU General Assembly 2021

© Author(s) 2021. This work is distributed under the Creative Commons Attribution 4.0 License.



## Transfer Data from NetCDF on Hierarchical Storage to Zarr on Object Storage: CMIP6 Climate Data Use Case

**Marco Kulüke**<sup>1</sup>, Fabian Wachsmann<sup>1</sup>, Georg Leander Siemund<sup>2</sup>, Hannes Thiemann<sup>1</sup>, and Stephan Kindermann<sup>1</sup>

<sup>1</sup>German Climate Computing Center, Data Management, Hamburg, Germany

<sup>2</sup>University of Hamburg, Hamburg, Germany

This study provides a guidance to data providers on how to transfer existing NetCDF data from a hierarchical storage system into Zarr to an object storage system.

In recent years, object storage systems became an alternative to traditional hierarchical file systems, because they are easily scalable and offer faster data retrieval, as compared to hierarchical storage systems.

Earth system sciences, and climate science in particular, handle large amounts of data. These data usually are represented as multi-dimensional arrays and traditionally stored in netCDF format on hierarchical file systems. However, the current netCDF-4 format is not yet optimized for object storage systems. NetCDF data transfers from an object storage can only be conducted on file level which results in heavy download volumes. An improvement to mitigate this problem can be the Zarr format, which reduces data transfers, due to the direct chunk and meta data access and hence increases the input/output operation speed in parallel computing environments.

As one of the largest climate data providers worldwide, the German Climate Computing Center (DKRZ) continuously works towards efficient ways to make data accessible for the user. This use case shows the conversion and the transfer of a subset of the Coupled Model Intercomparison Project Phase 6 (CMIP6) climate data archive from netCDF on the hierarchical file system into Zarr to the OpenStack object store, known as Swift, by using the Zarr Python package. Conclusively, this study will evaluate to what extent Zarr formatted climate data on an object storage system is a meaningful addition to the existing high performance computing environment of the DKRZ.