

EGU24-17150, updated on 18 Mar 2025  
<https://doi.org/10.5194/egusphere-egu24-17150>  
EGU General Assembly 2024  
© Author(s) 2025. This work is distributed under  
the Creative Commons Attribution 4.0 License.



## Data access patterns of km-scale resolution models

Janos Zimmermann<sup>1</sup>, Florian Ziemer<sup>1</sup>, and Tobias Kölling<sup>2</sup>

<sup>1</sup>Deutsches Klimarechenzentrum (DKRZ) GmbH, Hamburg, Germany

<sup>2</sup>Max-Planck-Institut für Meteorologie, Hamburg, Germany

Climate models produce vast amounts of output data. In the nextGEMS project, we have run the ICON model at 5 km resolution for 5 years, producing about 750 TB of output data from one simulation. To ease analysis, the data is stored at multiple temporal and spatial resolutions. The dataset is now analyzed by more than a hundred scientists on the DKRZ levante system. As disk space is limited, it is crucial to obtain information, which parts of this dataset are accessed frequently and need to be kept on disk, and which parts can be moved to the tape archive and only be fetched on request.

By storing the output as zarr files with many small files for the individual data chunks, and logging file access times, we obtained a detailed view of more than half a year of access to the nextGEMS dataset, even going to regional level for a given variable and time step. The evaluation of those access patterns offers the possibility to optimize various aspects such as caching, chunking, and archiving. Furthermore, it provides valuable information for designing future output configurations.

In this poster, we present the observed access patterns and discuss their implications for our chunking and archiving strategy. Leveraging an interactive visualization tool, we explore and compare access patterns, distinguishing frequently accessed subsets, sparsely accessed variables, and preferred resolutions. We furthermore provide information on how we analyzed the data access to enable other users to follow our approach.