EGU24-19677, updated on 20 Mar 2025
https://doi.org/10.5194/egusphere-egu24-19677
EGU General Assembly 2024

A cascaded framework for unified access to and analysis of kilometer scale global simulations across a federation of data centers

Kameswarrao Modali[1], Karsten Peters-von Gehlen[1], Florian Ziemen[1], Rajveer Saini[2], Simon Grasse[2], and Martin Schultz[2]

[1]Deutsches Klima Rechenzentrum,Hamburg, Germany (modali@dkrz.de)

[2]Jülich Supercomputing Center(JSC), Jülich, Germany(m.schultz@fz-juelich.de)

As the High Performance Computing (HPC) marches into the exascale era, earth system models have transformed into a numerical regime wherein simulations with a 1 km spatial resolution on a global scale are a reality and are currently being performed at various HPC centers across the globe. In this contribution, we provide an overview of the strategy and plans to adapt the data handling services and workflows available at the German Climate Computing Center (DKRZ) and the Jülich Supercomputing Center (JSC) to enable efficient data access, processing and sharing of output from such simulations using current and next generation Earth System Models. These activities are carried out in the framework of projects funded on an EU as well as national level, such as NextGEMS, WarmWorld and EERIE.

With the increase in spatial resolution comes the inevitable jump in the volume of the output data. In particular, the throughput due to the enhanced computing power always surpasses the capacity of single-tier storage systems made up of homogeneous hardware and necessitates multi-tier storage systems consisting of heterogeneous hardware . As a consequence, new issues arise for an efficient, user-friendly data management within each site. Sharing of model outputs that may be produced at different data centers and stored across different multi-tier storage systems poses additional challenges, both in terms of technical aspects (efficient data handling, data formats, reduction of unnecessary transfers) and semantic aspects (data discovery and selection across sites). Furthermore, there is an increasing need for scientifically operational solutions, which requires the development of long-term strategies that can be sustained within the different data centers. To achieve all of this, existing workflows need to be analyzed and largely rewritten. On the upside, this will allow the introduction of new concepts and technologies, for example using the recent zarr file format instead of the more traditional netCDF format.

More specifically, in WarmWorld, the strategy is to create an overarching user interface, to enable the discovery of the federated data, and implement the backend infrastructure for handling the movement of the data, across the storage tiers (SSD, HDD, tape, cloud), within as well as across the HPC centers, as necessitated by the analytical tasks. This approach will also leverage the benefits of community efforts in redesigning the way km-scale models provide their output, i.e. on

hierarchical grids and in relatively small chunks.

We present specific ongoing work to implement this data handling strategy across HPC centers and outline the vision for the handling of high-volume climate model simulation output in the exascale era to enable the efficient analysis of the information content from these simulations.