Facilitating global access to a high-volume flagship climate model dataset: the MPI-M Grand Ensemble experience

Karsten Peters\textsuperscript{1,3}, Michael Botzet\textsuperscript{2}, Veronika Gayler\textsuperscript{2}, Estefania Montoya Duque\textsuperscript{2}, Nicola Maher\textsuperscript{2}, Sebastian Milinski\textsuperscript{2}, Katharina Berger\textsuperscript{1}, Fabian Wachsmann\textsuperscript{1}, Laura Suarez-Gutierrez\textsuperscript{2}, Dirk Olonscheck\textsuperscript{3}, and Hannes Thiemann\textsuperscript{1}

\textsuperscript{1}Deutsches Klimarechenzentrum GmbH (DKRZ), Datamanagement, Hamburg, Germany
\textsuperscript{2}Max-Planck-Institut für Meteorologie, Hamburg, Germany
\textsuperscript{3}peters@dkrz.de

In a collaborative effort, data management specialists at the German Climate Computing Centre (Deutsches Klimarechenzentrum, DKRZ) and researchers at the Max Planck Institute for Meteorology (MPI-M) are joining forces to achieve long-term and effective global availability of a high-volume flagship climate model dataset: the MPI-M Grand Ensemble (MPI-GE, Maher et al. 2019\textsuperscript{1}), which is the largest ensemble of a single state-of-the-art comprehensive climate model (MPI-ESM1.1-LR) currently available. The MPI-GE has formed the basis for a number of scientific publications over the past 4 years\textsuperscript{2}. However, the wealth of data available from the MPI-GE simulations was essentially invisible to potential data users outside of DKRZ and MPI-M.

In this contribution, we showcase the adopted strategy, experiences made and the current status of FAIR long-term preservation of the MPI-GE dataset in the World Data Center for Climate (WDCC), hosted at DKRZ. The importance of synergistic cooperation between domain-expert data providers and knowledgeable repository staff will be highlighted.

Recognising the demand for MPI-GE data access outside of its native environment, the development of a strategy to make MPI-GE data globally available began in mid 2018. A two-stage dissemination/preservation process was decided upon.

In a first step, MPI-GE data would be published and made globally available via the Earth System Grid Federation (ESGF) infrastructure. Second, the ESGF-published data would be transferred to DKRZ's long-term and FAIR archiving service WDCC. Datasets preserved in the WDCC can be made accessible via ESGF - global access via the established system would thus still be ensured.

To date, the first stage of the above process is completed and data are available via the ESGF\textsuperscript{3}. Data published in the ESGF has to comply with strict data standards in order to ensure efficient data retrieval and interoperability of the dataset. Standardization of the MPI-GE data required selection of an applicable data standard (CMIP5 in this case) and an appropriate variable subset, adaptation and application of fit-for-purpose DKRZ-supplied post-processing software and of course the post-processing of the data itself. All steps required dedicated communication and
collaboration between DKRZ and MPI-M staff and required significant time resources. Currently, some 87 TB, comprised of more than 55,000 records, of standardized MPI-GE data are available for search and download from the ESGF. About three to four thousand records with an accumulated volume of several hundred GB are downloaded by ESGF users each month.

The long-term archival of the standardized MPI-GE data using DKRZ’s WDCC-service is planned to begin within the first half of 2020. All preparatory work done so far, especially the data standardization, significantly reduces the effort and resources required for achieving FAIR MPI-GE data preservation in the WDCC.


2http://www.mpimet.mpg.de/en/grand-ensemble/publications/

3https://esgf-data.dkrz.de/projects/mpi-ge/