Project: **1261**
Project title: **N4E - NFDI4Earth**
Principal investigator: **Ivonne Anders**
Report period: **2022-11-01 to 2023-10-31**

Introduction:

The National Research Data Infrastructure (NFDI) is intended to systematically develop, sustainably secure and make accessible the data holdings of science and research and to network them (inter)nationally. It will be established in a process driven by the scientific community as a networked structure of consortia acting on their own initiative. DKRZ is co-applicant of NFDI4Earth (https://www.nfdi4earth.de/), whose funding began in October 2021 and will initially run for 5 years. Uni-HH, Hereon, AWI, MPI BGC-Jena and other DKRZ user institutes are further partners of this consortium. NFDI4Earth currently supports 14 small flexible pilot projects (1-year projects) to integrate a broad community and innovative developments into the NFDI4Earth. The pilots cover a broad range of topics (technical implementations, new methods, new standards and FAIRness, reproducibility, interoperability, and others). Current pilots began in April 2022, some delayed due to staffing and contractual issues. Therefore, the mentioned pilots will end by summer 2023. A new round of pilots started in September 2023 and supports 7 diverse pilots within NFDI4Earth. Their duration is also limited to one year.
This report is about the progress in the Pilot „OcMOD – Observations closer to model data“ which is initiated by DKRZ and DWD. So far, other pilots did not make use of the offered resources.

Report on the Pilot OcMod:

Working with climate model output nearly always includes a validation of this data compared to reference data from observations, to ensure that the model data chosen is suitable for the individual research question or application. Model data and commonly used observational data have to be obtained from different sources: model data from e.g. the DKRZ and observational data from e.g. servers from public authorities. All data have to be prepared to be in the same formats and standards before it can be used for further analysis. A broad range of users (e.g. climate modelers, impact modelers, climate scientists, providers for climate services and education) are dealing with the same effort and troubles with the same sets of data. The aim of the pilot is to bring observational data close to the model output, to easily access data from public authorities and increase the number of users of various disciplines, and to provide this data in standardized formats for easy usage. Figure 1 illustrates this: the current workflow for the user in the background and, marked by a red circle, those parts, which OcMOD will focus on.

As described in the application, we focus on the example of the reanalysis data COSMO-REA6 of the German Weather Service . As a quasi-observation (model simulation with assimilation of observation data), this is an important data set for evaluating models in comparison with past simulations for Europe. The reanalysis provides numerous meteorological fields of the past, but not all are needed in equal measure by all users.

*Survey (https://c6dreq.dkrz.de/files/ocmod_dreq.php)**:*** Since the amount of data of the reanalysis is very large, a survey was first created in close cooperation with the meteorological service, which parameters of the data set are needed by the users. In this way we were able to make a selection of parameters, which data are most important for the users and can be processed with the available resources.

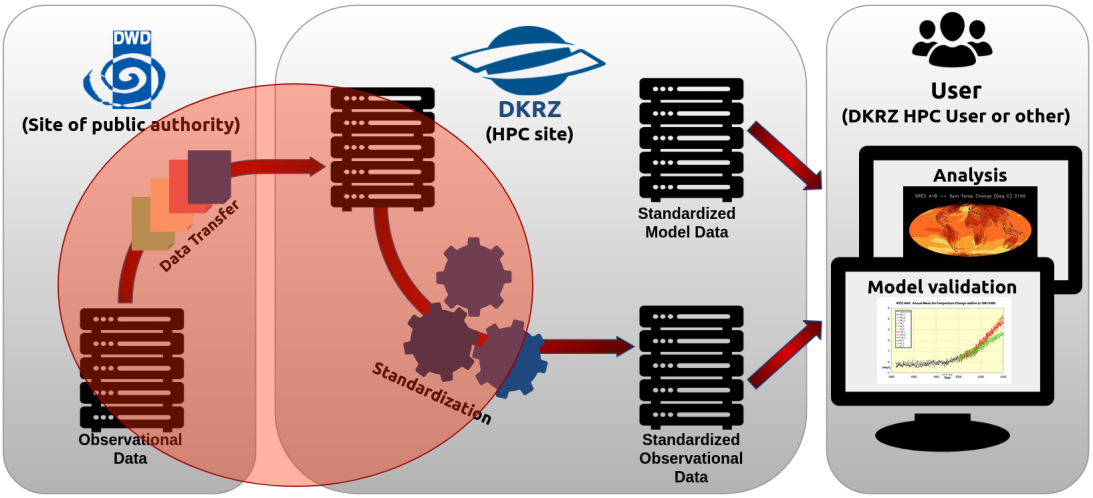

Figure 1: Typical workflow of a DKRZ HPC user that requires external datasets of public authorities such as the DWD.

Data Preparation (https://gitlab.dkrz.de/dicad-pp/ocmod): The selected parameters are partly available via the open data service of the DWD, and partly stored in the ECMWF tape archive. While the data from the DWD server is already in an aggregated form, the parameters in the ECMWF tape archive required an extensive aggregation. The processing as well as the eventual quality control revealed several problems of the data provided via the DWD open data service, which have been corrected by the DWD. Thus this work contributed to significant improvements in terms of data quality for the COSMO-REA6 datasets. Since the CDOs were not capable of dealing with the vertical axis of the COSMO model (hybrid height axis), to process the 3D data, the FIELDEXTRA tool of the COSMO community had to be licensed for our use case, then installed on levante and applied. We plan to archive the resulting data via the DKRZ HSM and DOKU, since the time, disk space and nodehour consuming retrieval and FIELDEXTRA processing of the 3D data can only be a one time effort.

CMORisation: The project stakeholders selected the CORDEX data standard as the target of the CMORization process. The data processing scripts and CMOR tables are under gitlab version control, publicly available, and are momentarily tested and adjusted at DWD for the standardisation of the 2nd generation of COSMO-REA6, "R6G2", that is based on ERA5 rather than ERA-Interim. For the CMORisation process we make use of the *cdo cmor operator* (https://code.mpimet.mpg.de/projects/cdo/wiki/CDO_CMOR_Operator) and *standardisation WebGUI* (https://c6dreq.dkrz.de), which have been developed within the BMBF funded project *CMIP6 DICAD*.

Outlook: Due to the long processing time of the 3D variables, the publication process is estimated to extend until winter 2023/2024. While writing this report, the 2D variables and a small set of 3D variables had already been ESGF published, with their WDCC publication being in progress. Furthermore, the monthly mean datasets have been transferred to the DKRZ cloud, to serve as data source for a Hackathon "DataXplorers" in Jena organised by NFDI4Biodiversity, NFDI4Earth, and NFDI4Microbiota (https://www.nfdi.de/dataxplorers-hackathon/). The publication of the other important result of the OcMOD pilot, a general blue print text and schema for the standardisation and utilisation of datasets of public authorities, is expected by November 2023.