Project: **1261**
Project title: **N4E - NFDI4Earth**
Principal investigator: **Ivonne Anders**
Report period: **2021-11-01 to 2022-10-31**

Introduction:

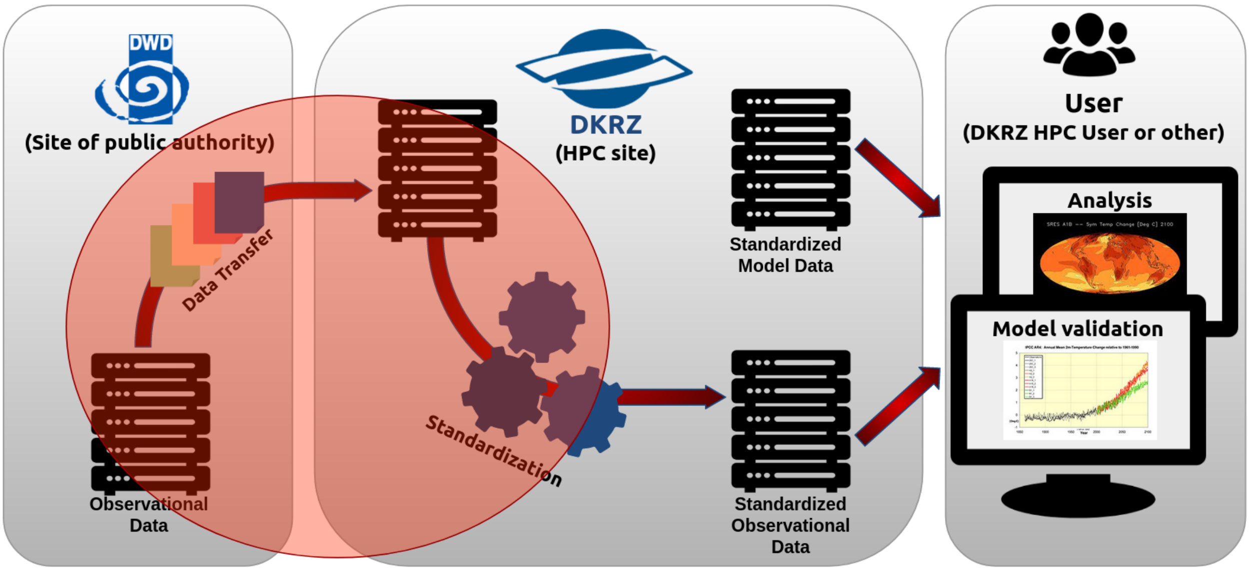
The National Research Data Infrastructure (NFDI) is intended to systematically develop, sustainably secure and make accessible the data holdings of science and research and to network them (inter)nationally. It will be established in a process driven by the scientific community as a networked structure 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 will subsequently start in September. 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, we don't have another pilot who has used the resources. This has to do with postponements of the runtime and that the pilots first do preliminary work on a topic-by-topic basis. We expect the resources to be used by the end of 2022 but mainly in 2023.

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 XX 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 in the example on the reanalysis data of the German Weather Service COSMO-REA6. 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: 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. Thereby it was recorded which temporal resolution is desirable in each case, with which priority. The reason for use was also queried. With regard to the 2nd version of the reanalysis, it was also asked which parameters are needed that are not included in version 1. 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. The survey was sent out to the stakeholders of the pilot (the ClimXtreme and RegiKlim projects) but also to different communities from Global and Regional Climate Modeling and Climate Services who are particularly using this data. We received about 20 responses from individuals but also on behalf of communities, allowing us to make a well-founded parameter selection that on the one hand meets the user demands while on the other hand limiting the storage resources required for this project.

Data Preparation: 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 and has been downloaded to the levante file system, the parameters in the tape archive will be retrieved in the coming months in cooperation with the DWD, and require an extensive aggregation. We plan to archive this data via the DKRZ HSM (double), since the time consuming retrieval from the ECMWF tape archive can only be a one time effort.

Process of CMORization: The project stakeholders selected the CORDEX data standard as the target of the CMORization process. The data processing scripts are under gitlab version control, are continuously extended as well as tested and already include the complete list of 2D parameters. Eventually, the data will be CMORized in two chunks - the open data and the ECMWF parameter set. For the CMORization process we make use of the *cdo cmor operator* (https://code.mpimet.mpg.de/projects/cdo/wiki/CDO_CMOR_Operator) and *standardization WebGUI* (https://c6dreq.dkrz.de), which have been developed within the BMBF funded project *CMIP6 DICAD*.