Project: **1261**
Project title: **N4E - NFDI4Earth**
Principal investigator: **Ivonne Anders**
Report period: **2023-11-01 to 2024-10-31**

Introduction

The National Research Data Infrastructure (NFDI) aims to systematically develop, securely maintain, and provide access to data resources in science and research, facilitating both national and international collaboration. Driven by the scientific community, the NFDI will be implemented as an interconnected structure of independently operating consortia. The German Climate Computing Center (DKRZ) is a co-applicant for the NFDI4Earth initiative (https://www.nfdi4earth.de/), which received funding in October 2021 for an initial five-year period. Additional partners in this consortium include Uni-HH, Hereon, AWI, MPI BGC-Jena, and other institutions that use DKRZ resources.
NFDI4Earth currently supports the third round of 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 start in will start between October 2024 and January 2025. The previous round of pilots started in September/October 2023 and supported 7 diverse pilots within NFDI4Earth. DKRZ supported one specific pilot in the framework of NFDI4Earths User Support Network activities, to learn about user needs.
This report is mainly about the progress in the Pilot „ OcMOD – Observations closer to model data “, which is initiated by DKRZ and DWD.

Report on the Pilot OcMod

When working with climate model output, it is almost always necessary to validate this data against observational reference data to ensure its suitability for the specific research question or application. Model data and commonly used observational data are sourced from different providers—for example, model data from the DKRZ and observational data from public authority servers. All data must be converted to consistent formats and standards before further analysis. A diverse group of users, including climate modelers, impact modelers, climate scientists, climate service providers, and educators, face similar challenges with these data sets. This pilot project aims to streamline access to observational data alongside model output, making data from public authorities readily accessible, increasing usage across various disciplines, and ensuring data is available in standardized formats for ease of use. 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.

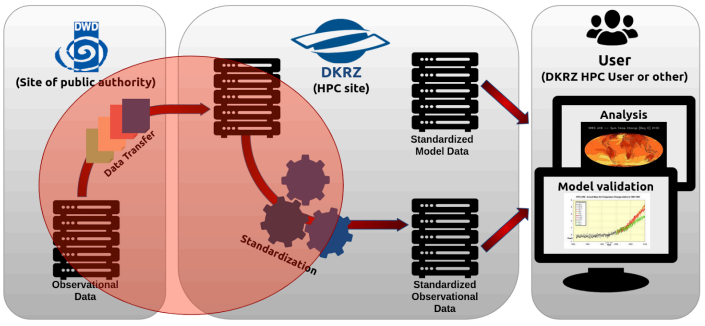

Figure 1: Typical workflow of a DKRZ HPC user that requires external datasets of public authorities such as the DWD.

After key prerequisites were created in 2023, such as data transfer, data checking, correction and adaptation of standards and other important aspects, the data could be standardized and published this year. This includes the following subtasks:

Compression and archiving of the raw data from COSMO-REA6 The COSMO-REA6 raw data was compressed for space-saving and efficient storage and then packed in tar-format. This data was stored in `/arch`. The data is now centrally archived and can be retrieved and processed more easily if required.

Republication of data in ESGF under new license The DWD is bound by the GeoNutzV (\'Verordnung zur Festlegung der Nutzungsbestimmungen für die Bereitstellung von Geodaten des Bundes\'). In an internal DWD coordination process, it was decided for this specific case study that publication under CC BY 4.0 is possible. This is an important step, as it enables more flexible use and distribution of the data and ensures that all legal requirements for broad and transparent use of the data are met. Selected and aggregated data was republished in ESGF under the new CC BY 4.0 license as part of an adaptation to current licensing requirements.

Errata for the data in ESGF Errata were identified and documented during the publication process. These can be viewed under the following link: Errata overview (see tab: “errata”). This documentation serves as a central point of contact to provide users with important information about any data errors and their correction processes.

Publication and archiving of the data in the WDCC with DOI In addition to the ESGF, the reanalysis data was also published and archived in the World Data Center for Climate (WDCC). Each dataset was assigned a DOI to enable clear referencing and citation. The publication is available at the following link: COSMO-REA6 DOI. Archiving and assigning DOIs significantly improves the traceability and citation of scientific work. The German Meteorological Service can thus also trace when the data is used for scientific purposes.

Data moved to a new pool project The ESGF data previously stored in the /work directory has been moved to a new /pool/data project to improve data structuring and availability. Further information and the project directory are documented here: Pool data projects

Preparation and cmorization of the 3D variables for DOKU publication/archiving For the variables `ua,va,ta,pfull` and `hus` (now on 41 vertical levels instead of 6) a preparation and cmorization for DOKU publication and archiving was performed. After a subsequent review and testing process (RT), the data will be released for publication by the end of this year.

Creation of a blueprint The project's insights and lessons learned were incorporated into a blueprint, providing guidance on making data from other authorities accessible and usable for both research and the public. Overall, the entire process can be divided into 5 sub-steps: (1) **determination and classification of the need**, (2) **survey of the feasibility**, (3) **implementation**, (4) **feedback and follow-up**, (5) **dissemination**. This blueprint outlines generalizable steps and aspects applicable across domains and collaborators, offering a framework for optimizing the use of governmental data in diverse fields. The detailed blueprint is available via this link: Blueprint.

Support for the DWD in the follow-up project R6G2 The German Weather Service (DWD) is being supported in the follow-up project R6G2 in the process of cmorizing the data and preparing it for publication on the DWD's ESGF node. This support will ensure the use and availability of the project on ESGF and thus make it accessible to other users.

Report on the Pilot “Propagating complex uncertainties in data cubes”

The pilot “Propagating complex uncertainties in data cubes” coordinated at TU Tübingen dealt with how uncertainties in proxy data in the field of paleoclimate research can be merged with model data and integrated accordingly in a datacube. The DKRZ project provided support here with advice on computing resources at the DKRZ. These were ultimately not utilized. Further information and the pojects proposal can be found here.