Proposal for an NFDI4Earth Pilot

Combined Analysis and Publication of Ice Sheet data CAPICE

Angelika Humbert[1], Jonas Eberle[2], Philipp S. Sommer[3]

[1] Alfred-Wegener-Institut Helmholtz Zentrum für Polar- und Meeresforschung (AWI)
[2] Deutsches Zentrum für Luft- und Raumfahrt e. V. (DLR)
[3] Helmholtz-Zentrum Hereon

Date of submission 30 March 2023

Abstract: This project is concerned with data driven ice sheet modelling that requires efficient ingestion of climate forcing data and satellite data products. The goal of this pilot is to establish an efficient workflow for simulations on distributed system. The expected result is a blueprint for workflows of similar kind. Users of this workflow are Earth system compartment modelers, with similar demands. The outcome of this pilot can enhance RDM workflows for satellite data driven modelling and the publication of respective simulation data.

I. Introduction

This NFDI4Earth Pilot addresses the goal "**Collaborative Analysis integrating different Sources".** It aims at implementing a technical solution to enable cross-centre interactive and semi-automatized work on complex data from Earth Observation (EO) and Earth System Modelling (EM), and prepare publication of resulting analysis data product in persistent and citable form. The needs for such a development have been identified by the NFDI4Earth interest group on FAIR HPC Data [1,2].

Domain models, such as ice sheet models, are requiring input data from climate models (global circulation models, GCMs), which are used as boundary conditions, which we call forcing. Forcing data is huge and normally the output of GCMs are not hosted on all HPC clusters due to their vast size. Next generation ice sheet models are using in addition to the climate forcing data also satellite observation data (or derived products) as input data. The satellite observation data is normally not stored on HPC systems and when AI methods are used, GPU based processing is preferable. As a consequence, the domain model is requiring data as runtime input from different platforms. In the ideal world, one computing centre may host the GCM data, another one the satellite data and the domain model would run on an HPC platform not directly attached to those. This brings of course several challenges with it, given that also performance should not break down during data ingestion into the simulation.

With this project we aim at implementing a distributed processing and modeling, so that GCM boundary data is extracted from a separate platform that also offers preprocessing capacity to reduce data volumes to a minimum before transfer. In this way the data management is very much automatized to allow for a modeling cycle without manual intervention on two sites. Furthermore, the same service operating at the climate data centre DKRZ in Hamburg will be used to transfer back the result data from the simulation computed in Munich at LRZ. In this case AWI uses the open source ice model ISSM [3]. This ice sheet code has been extensively applied for sea level projections and has contributed to the latest IPCC reports [4,5]. The code is widely applied in the community and is extended frequently with new features and capabilities. Figure 1 gives a schematic representation of the coupled infrastructures. On the left hand side the German Climate Computing Centers' CMIP6 data residing in Hamburg is depicted, on the right hand side the satellite imagery data depicted as stored at the DLR operated terrabyte infrastructure in Munich. The bottom displays the actual simulation that will be performed on LRZ systems in Munich as well.

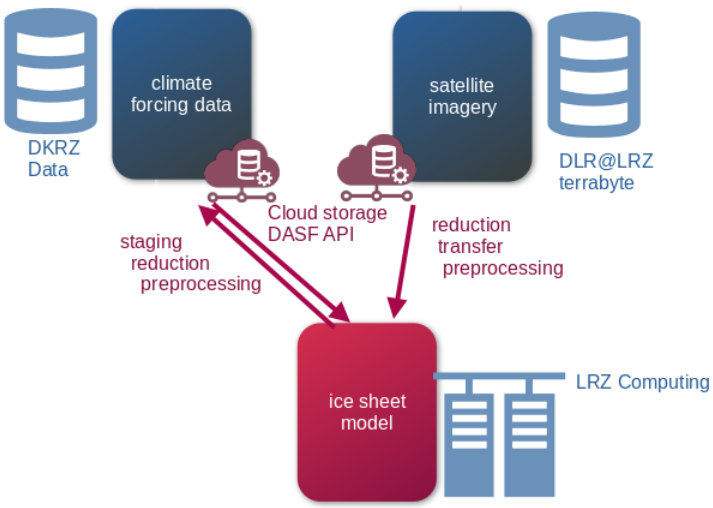

Figure 1: Coupling data and modeling infrastructures using the message broker within DASF

Key innovation is the use of a message broker within the Data Analyses Software Framework [6] (DASF), as well as cloud-based storage platform which is currently going to operation at DKRZ for collaboration projects and interoperability. The resulting services of the centres needed to allow for a distributed workflow will be kept alive for further use in NFDI, and the technical approach will be described in an NFDI4Earth article publication.

II. Pilot Description

The pilot develops a simulation and data workflow as depicted in Figure 2. The already at LRZ running Ice model runtime environment will be augmented by two communication paths to 1) initiate preprocessing and transfer of data from DKRZ, and 2) to store back simulation results for later publication and early access of the group of researchers to the cloud storage resources.

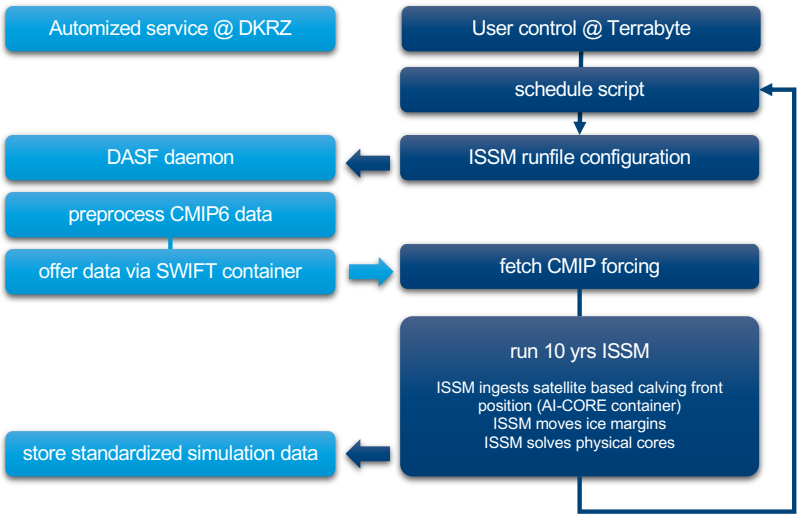

Figure 2: Workflow scheme of computational tasks distributed between LRZ/DLR Terrabyte and DKRZ.

Simulation data will be stored continuously at DKRZ, where a group of users may further analyse the results with restricted access in their collaboration and prepare the data product for publication within the ISMIP6 consortium.Opening access to data and processing capacity are crucial for distributed and collaborative scientific workflows. The workflow-centric approach of the pilot follows the principle of “Thinking in Workflows” that has been established in the Digital Earth community [7]. Generalisation on an abstraction layer are a key element to secure reusability of the developed entities. One such entity is a software container for computing iceberg calving positions by machine learning methods that has been developed during the HGF funded AI-CORE project and is available on Terrabyte.

The communication between the two computing environments DKRZ and Terrabyte is established through the Data Analytics Software Framework [6] (DASF), a messaging framework that leverages scientific python-code for a so-called remote procedure call (RPC). In our setup, the scientist writes a python API that runs at the DKRZ and extracts the CMIP6 data that is required to run the ice sheet model. This data is then uploaded to a swift container[1]. At LRZ, the user can use the exact same script, just that DASF leverages the API call to extract the CMIP6 data and processes it at the DKRZ. The python script at the LRZ will receive the link to download the data from the swift container. The connection within DASF between the two HPC resources is established through a so-called message broker that transmits the requests from one HPC resource to the other.

Data publication and sharing for ISMIP6
The Ice Sheet Model Intercomparison Project 6 (ISMIP6) has set up standards for data sharing [8] for downstream applications of simulation output, such as impat models or emulators, as well as for analysis of the benchmark. We will follow their implemented standards for data formats, grid, output variables. For the final data publication, it will first be investigated whether the available result data can be published within the framework of the Earth System Grid Federation (ESGF) [9] according to CMIP6 [10] data standards. Data published in the ESGF are prepared for final long-term archiving in the certified World Data Centre for Climate (WDCC) [11]. A DOI will be assigned to them there. Supplementary data can be included in the basic long-term archiving (DOKU) of the DKRZ. Before final publication the simulation data is already gathered at DKRZ during the run of the computational workflow. It is intended to be shared among the members of the PI's workgroup and also non-DKRZ users, demonstrating the storage capability for open collaboration.

Re-usage of standardised geospatial data-access/processing toolkits
Our workflows will re-use community-standard processing tools as far as possible. In particular, a standardized way of conducting federated processing of geospatial data is currently being discussed in the new Geo Data Cube API working group of the Open Geospatial Consortium (OGC). Also, in the recent years, a web service specification named OpenEO [12] has been developed – funded by the European Union's Horizon 2020 program as well as the European Space Agency, which allows for standardized processing federated and across data centers. We will follow such developments which are supposed to accelerate these years involving also large data centers and data providers, such as ESA, EUMETSAT, ECMWF. In particular, we plan to evaluate OpenEO towards the sustainability phase of our effort.

Collaboration between workflow development and scientific application
The work plan (Section V) has been carefully constructed to ensure close collaboration between the teams conducting workflow development and infrastructure provisioning, and the science application team. WPs 1 and 3 (requirements collection and actual runs) will be realised in close and comprehensive collaboration of application providers, computing centres, and providers of innovative system components such as DASF (see work plan). The latter two groups will collaborate on WP2 for a proper infrastructure set-up, while WP4 – worked on by AWI, DKRZ and LRZ - will care about sustainability including data publication. In all this effort, the scientific application team will provide a setup that is currently already running on Terrabyte and develop the time slice scheme for the production runs. Also the code for preprocessing of CMIP6 output at DKRZ will be provided by the science application team. This is the basis for the workflow development.

III. Relevance to NFDI4Earth
This pilot is a result of the previous work in the NFDI4Earth Interest Group "IG High-performance Computing in Earth System Sciences" and combines the domains of climate modelling with satellite observation. Compared to the traditional execution method, a data reduction takes place close to the climate data, which considerably reduces the data transfer to the ice sheet model. At the same time, the execution of the processes on the participating data centres is closely interlocked by a message broker in order to avoid unwelcome waiting states of the climate model on the mainframe and to minimise the storage requirements for intermediate results.

However, this pilot does not only address the immediate use case. By using (large-scale) computing capacities at different locations and a wide variety of storage solutions such as HPC and cloud storage,

[1]The swift object storage is operated by the DKRZ and is used as a platform to exchange the netCDF data. https://docs.dkrz.de/doc/datastorage/swift/index.html

the pilot also improves the approach of data-centric computing by executing algorithms close to the data.

DASF is already in use within the Helmholtz framework for selected usage scenarios. With the help of this pilot, it is now being transferred to another use case. It promotes (and requires) collaboration between HPC centres and serves as a blueprint for further use cases in NFDI4Earth that require distributed stored data sets. It also has the potential to serve as a stimulus and template in further NFDI consortia via the NFDI Common Infrastructures section. The intended integration of the existing algorithms in OpenEO (alongside the message broker DASF) enables the use of another standard Earth system science framework in NFDI4Earth. The scientific data generated in this pilot will be prepared for archiving in repositories represented in NFDI4Earth, such as the WDCC, to enable long-term reuse in NFDI4Earth and beyond. The embedding of this pilot in the IG HPC in Earth System Science, in which several HPC centres collaborating in NFDI4Earth are represented, provides optimum conditions for the workflow setup to be re-used and further built upon. Technically, this will be supported by packaging workflow components and deployment descriptions. A special work package is envisaged to support sustainability of the results within NFDI4Earth and beyond.

IV. Deliverables

D1.1 Short report on requirements analysis (M2)
D1.2 Documentation of data & computing systems made available (M4)
D2.1 Report on DASF setup (M8)
D3.1 Report on simulation runs (M11)
D4.1 Datasets published/archived (M12)
D4.2 Sustainability concept – short report (M12)

V. Workplan and requested funding

WP1 (DKRZ, AWI, DLR, LRZ) Requirements for distributed workflow and availability of storage and compute systems (M1-M4)
Considering our distributed workflow concepts (Sec. II), a requirements analysis is conducted regarding storage/compute systems (M1-M2). To ensure generalisability, systems with standard interfaces are allocated at each involved site, as well as IaaS infrastructure for DASF (M1-M4).

WP2 (HEREON, DKRZ, LRZ, DLR) Installation and adaptation of data and workflow interfaces (M4-M12)
The DASF stack for workflow automatisation is installed at the sites involved (M4-M7). Data systems access need to be made technically accessible (M4-M7), possibly via interfacing middleware such as MinIO. After the set-up of DASF, accommodating job submission peculiarities such as 2-factor authentication (M6-M8), system usage can begin. DASF is continuously maintained (M8-M12).

WP3 (AWI, LRZ, DLR, DKRZ, HEREON) Workflow Runs (M8-M11)
This work package runs the simulation workflow on the distributed systems after it has installed the necessary software packages. By M10, production of scientifically relevant output data can be expected, which are kept and enriched with basic (e.g. DataCite) metadata where relevant. As a more IT-centric WP component, the use of OpenEO in the workflow can be evaluated.

WP4 (LRZ, AWI, DKRZ) FAIR data publication, sustainability, generalisation plans (M10-M12)
In M10-M12, we will demonstrate best-practice data publication of an annotated data set from WP3, via repositories/mechanisms endorsed by NFDI4Earth. We will also develop a perspective for generalising our workflows and for sustainability (follow-up projects, packaging, deployment descriptions, publications – including code).

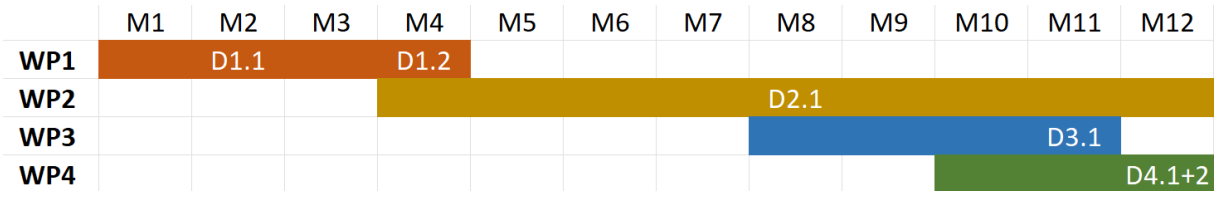

Fig. 3 Gantt chart of CAPICE

Requested Funding

Participating institutes contribute within their NFDI4Earth activities on their own. We request funding for 1 FTE at AWI Bremerhaven (PI A. Humbert) for the entire run-time of the project, for a person ideally familiar with our simulations but, above all, with skills in computational workflows to drive the project forward.

References

1. https://www.nfdi4earth.de/2participate/get-involved-by-interest-groups/ig-high-performance-computing-in-earth-system-sciences
2. Frickenhaus, S., Fritzsch, B., Hachinger, S., Humbert, A., Koldunov, N.V., Müller-Pfefferkorn, R., Munke, J., Trumpik, N., & Thiemann, H. (2022). FAIR@HPC – Improving HPC usage in ESS by FAIR data and compute services (1.0). Zenodo. doi: https://doi.org/10.5281/zenodo.6565405
3. https://issm.jpl.nasa.gov/, Larour, E., Seroussi, H., Morlighem, M., and Rignot, E. (2012), Continental scale, high order, high spatial resolution, ice sheet modeling using the Ice Sheet System Model (ISSM), *J. Geophys. Res.*, 117, F01022, doi:10.1029/2011JF002140
4. ISMIP Greenland: Goelzer, H., Nowicki, S., Payne, A., Larour, E., Seroussi, H., Lipscomb, W. H., Gregory, J., Abe-Ouchi, A., Shepherd, A., Simon, E., Agosta, C., Alexander, P., Aschwanden, A., Barthel, A., Calov, R., Chambers, C., Choi, Y., Cuzzone, J., Dumas, C., Edwards, T., Felikson, D., Fettweis, X., Golledge, N. R., Greve, R., Humbert, A., Huybrechts, P., Le clec'h, S., Lee, V., Leguy, G., Little, C., Lowry, D. P., Morlighem, M., Nias, I., Quiquet, A., Rückamp, M., Schlegel, N.-J., Slater, D. A., Smith, R. S., Straneo, F., Tarasov, L., van de Wal, R., and van den Broeke, M.: The future sea-level contribution of the Greenland ice sheet: a multi-model ensemble study of ISMIP6, The Cryosphere, 14, 3071–3096, https://doi.org/10.5194/tc-14-3071-2020, 2020
5. ISMIP Antarctica: Seroussi, H., Nowicki, S., Payne, A. J., Goelzer, H., Lipscomb, W. H., Abe-Ouchi, A., Agosta, C., Albrecht, T., Asay-Davis, X., Barthel, A., Calov, R., Cullather, R., Dumas, C., Galton-Fenzi, B. K., Gladstone, R., Golledge, N. R., Gregory, J. M., Greve, R., Hattermann, T., Hoffman, M. J., Humbert, A., Huybrechts, P., Jourdain, N. C., Kleiner, T., Larour, E., Leguy, G. R., Lowry, D. P., Little, C. M., Morlighem, M., Pattyn, F., Pelle, T., Price, S. F., Quiquet, A., Reese, R., Schlegel, N.-J., Shepherd, A., Simon, E., Smith, R. S., Straneo, F., Sun, S., Trusel, L. D., Van Breedam, J., van de Wal, R. S. W., Winkelmann, R., Zhao, C., Zhang, T., and Zwinger, T.: ISMIP6 Antarctica: a multi-model ensemble of the Antarctic ice sheet evolution over the 21st century, The Cryosphere, 14, 3033–3070, https://doi.org/10.5194/tc-14-3033-2020, 2020
6. Eggert, D. et al., (2022). DASF: A data analytics software framework for distributed environments. Journal of Open Source Software, 7(78), 4052, doi: https://doi.org/10.21105/joss.04052
7. Bouwer. L. et. al (2022): Integrating Data Science and Earth Science doi: https://doi.org/10.1007/978-3-030-99546-1
8. https://www.climate-cryosphere.org/wiki/index.php?title=ISMIP6-Projections2300-Antarctica and https://www.climate-cryosphere.org/wiki/index.php?title=ISMIP6-Projections-Greenland
9. https://esgf.llnl.gov/
10. https://pcmdi.llnl.gov/CMIP6/Guide/dataUsers.html
11. https://www.wdc-climate.de/
12. Schramm, M., Pebesma, E., Milenković, M., Foresta, L., Dries, J., Jacob, A., Wagner, W., Mohr, M., Neteler, M., Kadunc, M., Miksa, T., Kempeneers, P., Verbesselt, J., Gößwein, B., Navacchi, C., Lippens, S., & Reiche, J. (2021). The openEO API–Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities. *Remote Sensing*, *13*(6), [1125]. doi: https://doi.org/10.3390/rs13061125