A workflow for the Climate Digital Twin

Aina Gaya-Àvila[1] (aina.gayayavila@bsc.es) , Leo Arriola i Meikle[1], Francesc Roura Adserias[1], Bruno De Paula Kinoshita[1], Daniel Beltrán Mora[1], Rohan Ahmed[1], Miguel Andrés-Martínez[2] and Miguel Castrillo [1]

1 : Barcelona Supercomputing Center (BSC), Barcelona, Spain 2 : Alfred Wegener Institut (AWI), Bremerhaven, Germany

The **Climate Change Adaptation Digital Twin**[1] aims to deliver multi-decadal simulations at kilometer scale at high temporal resolution. For that purpose, it uses three coupled models (ICON, IFS-NEMO, and IFS-FESOM) that output in a standardized way (Generic State Vector, GSV), and runs on EuroHPC platforms, such as LUMI and MareNostrum5. The Digital Twin Engine includes essential components such as climate models, other software related to data streaming (GSV interface, one-pass algorithms[2], MARS) and model output management (MultIO[3], FDB[4]), and data consumers. This requires a workflow management solution to run portable and scalable climate simulation workflows across different HPC systems. We developed a **flexible**, **fully parametrized**, and **platform-agnostic** workflow, using **Autosubmit**[5]. It has served as the primary instrument for methodically configuring, managing, and conducting experiments. The developed workflow handles the deployment of the climate model, simulation triggering and monitoring, data processing, result generation by the use cases, and the archival of the data. Some of the components are run with containers to ensure consistent deployment across HPCs.

Workflow overview

The workflow includes three distinct modes: **end-to-end**, which generates model data and simultaneously streams it to data consumers; **model**, which executes the climate model components at high resolution; and **applications**, utilizing pre-existing data offline.
There are some common initial steps when running the workflow in any of these three modes, such as model and/or application deployment on HPC infrastructure, and rigorous checks that ensure subsequent steps' efficiency, crucial due to their resource-intensive nature and queuing requirements.

Data streaming

The workflow includes immediate and continuous data processing, promoting scalability in temporal and spatial resolution[6]. This approach ensures the efficient handling of intricate climate models, meeting the demands for **high-resolution temporal and spatial data**, while enhancing user accessibility and adaptability across different computational environments.
For the end-to-end mode, a **data listening mechanism** activates concurrently with model execution: upon commencement, a task awaits data arrival, triggering subsequent tasks responsible for data retrieval and computation of statistics requested by applications. Following data processing, applications execute and yield results.
The data consumers that run along with the model can **use the data as soon as it is produced**. To achieve this, two new features were added to Autosubmit:

- Start conditions in the dependencies: the data listening mechanism is triggered when the model starts running, so the data is consumed as soon as it is produced.
- Dependencies between splits (finer level of granularity).

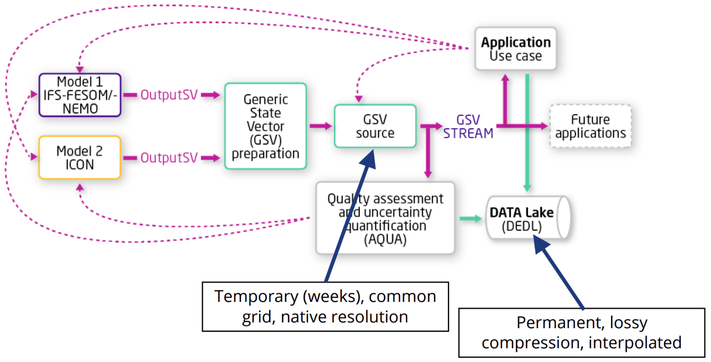

Figure 1: Workflow overview

Workflow configuration

We've optimized the configuration process by creating a user-friendly structure that **simplifies the interface for scientists**, hiding the model's inherent complexity in setting up simulations. The user can run a simulation by specifying just a small set of keys in a single YAML file, that act as switches for the background configuration. If more customization is desired, the user can overwrite any parameter of the configuration.

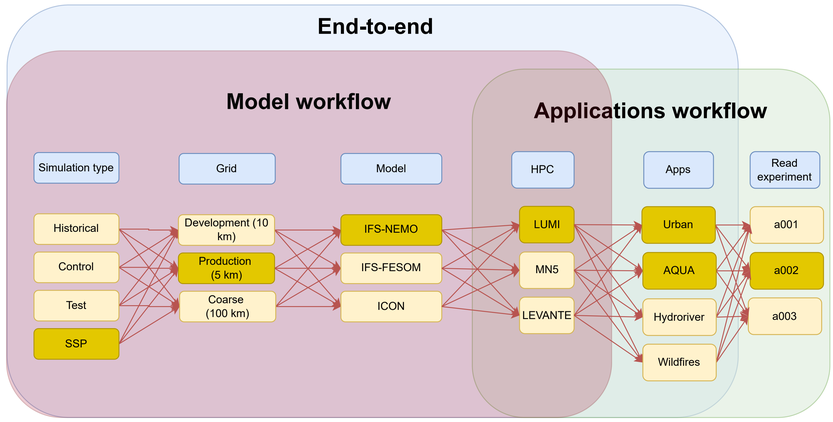

Figure 2: Configuration scheme. The user chooses between three different modes and then chooses the parameters for the experiment.

Traceability, robustness and efficiency

One experiment is conducted by executing a series of simulations. The experiment configuration is stored and preserved, as well as the logs of each command and the output logs of the tasks.

Benefits from Autosubmit

- Ensures the **tracking of processes** and the ability to **reproduce results** accurately.
- Automatic collection of **performance metrics**, which measure the efficiency of the simulation performed.
- If the HPC experiences instabilities or failures, Autosubmit is capable of **resubmitting the failed tasks**.
- **Job aggregation** capabilities (wrappers) allow to reduce the time-to-solution by reducing queuing times and optimizing resource usage.

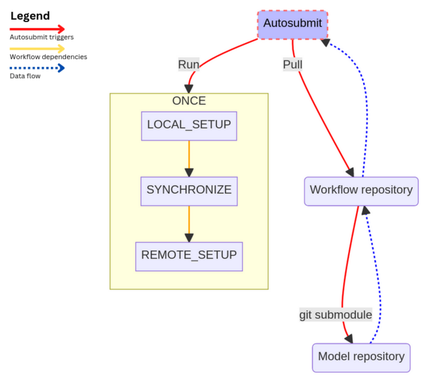

Figure 3: The workflow code and configuration are stored in a repository, and the model sources are included as git submodules. Git submodules have a fixed commit, so each release of the model has fixed model sources.

Conclusions

The described workflow has been able to successfully perform simulations for Phase 1 of the Climate Change Adaptation Digital Twin. It is a robust solution that ensures the traceability and reproducibility of the results obtained. It has been widely used by scientists and technicians to run simulations, reproducibility tests or performance analysis in a standardized way.

References

[1] J. Kontkanen et al. Climate digital twin to support climate change adaptation efforts. In *EGU General Assembly 2023*, Vienna, Austria, 2023. EGU23-13018.

[2] K. Grayson et al. One-pass algorithms for streamed climate data. In *EGU General Assembly 2024*, Vienna, Austria, 2024.

[3] D. Sármány et al. Multio: Message-driven data routing for distributed earth-system models. In *EGU General Assembly 2023*, Vienna, Austria, 2023. EGU23-13856.

[4] Simon D. Smart et al. A high-performance distributed object-store for exascale numerical weather prediction and climate. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '19, New York, NY, USA, 2019. Association for Computing Machinery.

[5] Domingo Manubens-Gil et al. Seamless management of ensemble climate prediction experiments on hpc platforms. In *2016 International Conference on High Performance Computing & Simulation (HPCS)*, pages 895–900, 2016.

[6] F. Roura-Adserias et al. The data streaming in the climate adaptation digital twin: a fundamental piece to transform climate data into climate information. In *EGU General Assembly 2024*, Vienna, Austria, 2024.

Acknowledgements

Destination Earth is a European Union funded initiative and is implemented by ECMWF, ESA, and EUMETSAT.

Funded by the European Union Destination Earth implemented by ECMWF esa EUMETSAT