DESTINATION EARTH

Climate Adaptation Digital Twin (DE340)

The data streaming in the Climate Adaptation Digital Twin: a fundamental piece to transform climate data into climate information

Francesc Roura-Adserias[1] (francesc.roura@bsc.es), Aina Gaya i Avila[1], Leo Arriola i Meikle[1], Miguel Andrés-Martínez[2], Dani Beltran Mora[1], Iker Gonzalez Yeregui[1], Katherine Grayson[1], Bruno De Paula Kinoshita[1], Rohan Ahmed[1], Aleksander Lacima-Nadolnik[1], and Miguel Castrillo[1]

[1] Barcelona Supercomputing Center (BSC), Earth Sciences, Barcelona, Catalonia, Spain, [2] Alfred Wegener Institute for Polar and Marine Research, Computing and Data Centre, Bremerhaven, Germany

Earth Sciences Department

Barcelona Supercomputing Center, Centro Nacional de Supercomputación

Why do we need data streaming?

The climate crisis poses extreme threat to our societies, and we need tools **to take informed decisions** (IPCC, 2022). To take these decisions, the improvement of global climate models (GCMs) is key. However, **high resolution models** that can be run in a more **agile and interactive way** produce an amount of data that can not be permanently stored (~80 Tb per real day).

To mitigate this long-term data storage issue, we employ data streaming; providing the data directly from the models to the users, without needing to store vast amounts of data to disk.

How do we achieve streaming?

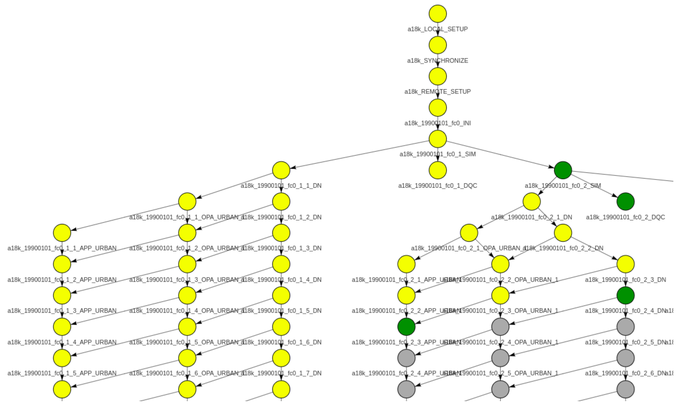

Figure 1: Example of how the automated workflow tasks are shown in the Autosubmit GUI.

- **Automated workflow manager:** Autosubmit (Manubens-Gil et al., 2016), orchestrates the workflow, transferring model data to the impact models (use cases) (Fig 1).
- **Data listening mechanism:** Workflow components that automatically notify the downstream workflow that data is available.
- **One – pass algorithms** (Grayson at al., in prep.): They allow large spatial domains from high-resolution GCMs to be incrementally processed, as each time step is combined with the rolling statistic summary, allowing for vast reductions in memory requirements in the temporal dimension (Fig 2).

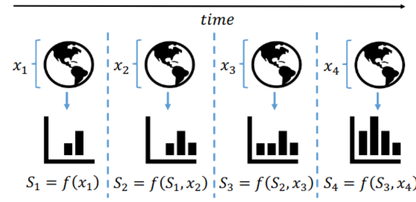

Figure 2: Example of how the one-pass algorithms compute statistics on streamed data, without having the whole data-set in view.

Why is this useful for the user and how can they get the data?

We can move **from global raw data to local climate information** in a more efficient and interactive way.
We will allow users to obtain data from the latest versions of the models and to interactively request variables that are not normally available. We provide **local information on a global scale**.

Users will retrieve the data directly **through online streaming** or **from the Data Lake** (long term storage), which will contain part of the streamed data (Fig 3).

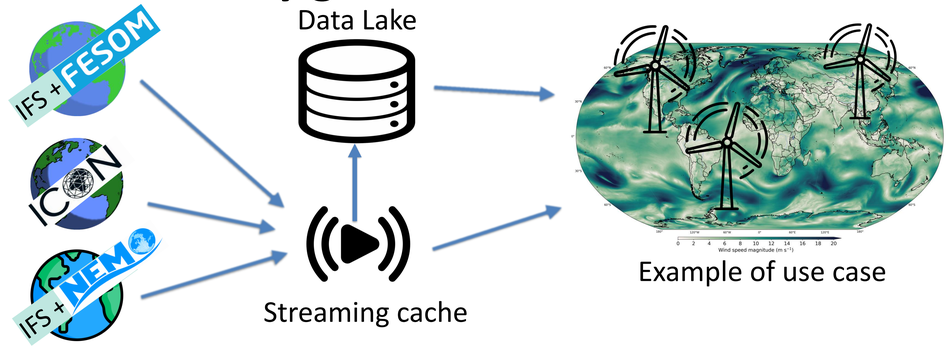

Figure 3: Diagram that shows the data flow from the to the data consumers, in this case an application related to wind energy.

Icons from https://www.flaticon.com/

Notorious features of the streaming

- **Platform agnostic:** ClimateDT is able to run on any platform, thanks to the use of containers and its internal logic.
- **Independent of the workflow engine.** The main workflow engine is Autosubmit but it can be ported to ecFlow and it follows the FAIR principles.
- **User interactivity:** as the workflow can satisfy user data requests, we can provide climate information in very high resolution interactively.

Managing streaming

The data listening mechanism together with Autosubmit allows **a timely execution** of the One - pass algorithms as well as the applications that depend on them. The workflow allows execution of applications at different frequencies (Fig 4).

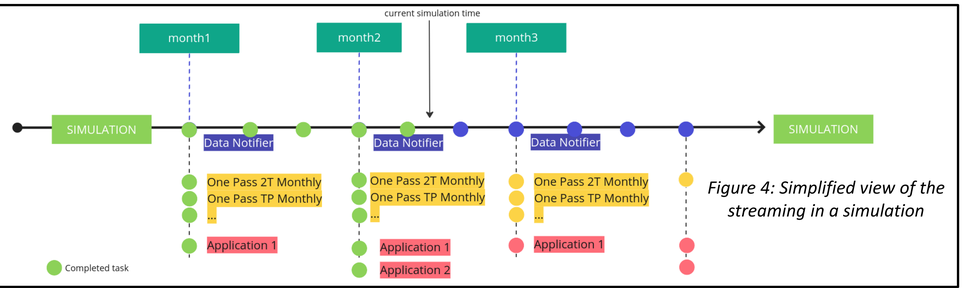

Figure 4: Simplified view of the streaming in a simulation

SCAN ME

Take home messages

- Streaming presents an **efficient solution** for addressing data storage concerns.
- **User interactivity** will be reached thanks to the digital twin.
- A robust and flexible workflow infrastructure is **essential to** configure, orchestrate, track and **interact with the digital twin execution**.

References

D. Manubens-Gil, J. Vegas-Regidor, C. Prodhomme, O. Mula-Valls, and F. J. Doblas-Reyes. Seamless management of ensemble climate prediction experiments on hpc platforms. *In 2016 International Conference on High Performance Computing Simulation (HPCS)*, pages 895–900, 2016. doi: 10.1109/HPCSim.2016.7568429.

Grayson, K., Thober, S., Sharifi, E., Lacima-nadolnik, A., Lledó, L., & Doblas-reyes, F. Statistical summaries for streamed data from climate simulations. 1–26. Under review in Geoscientific Model Development, 2024

IPCC, 2022: Summary for Policymakers [H.-O. Pörtner, D.C. Roberts, E.S. Poloczanska, K. Mintenbeck, M. Tignor, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem (eds.)]. In: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 3-33, doi:10.1017/9781009325844.001.

Acknowledgements

Destination Earth is a European Union funded initiative launched in 2022, with the aim to build a digital replica of the Earth system by 2030. The initiative will be jointly implemented by three entrusted entities: the European Centre for Medium-Range Weather Forecasts (ECMWF) responsible for the creation of the first two 'digital twins' and the 'Digital Twin Engine', the European Space Agency (ESA) responsible for building the 'Core Service Platform', and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), responsible for the creation of the 'Data Lake'.

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