Earth Sciences Department

Generic State Vector: Streaming and accessing high resolution climate data from models to end users

Iker Gonzazlez-Yeregi[1] (iker.gonzalez@bsc.es) , Pierre-Antoine Bretonnière[1], Aina Gaya-Àvila[1], Francesc Roura-Adserias [1]

1 : Barcelona Supercomputing Center (BSC), Barcelona, Spain

The Climate Adaptation Digital Twin (ClimateDT) is a contract under the Destination Earth initiative (DestinE) that aims to develop a **digital twin to account for climate change adaptation**[1]. This is achieved by running **high-resolution simulations** with different climate models by making use of the **different EuroHPC platforms** (Marenostrum5, Lumi, ...). In addition to the climate models, **applications that consume data** from models are also developed under the contract. A **common workflow** is used to execute the whole pipeline **from the model launching to the data consumption by the applications** in a user-friendly and automated way[2]. One of the challenges of this complex workflow is to handle the **different outputs** that each of the climate models initially offered. Each model works with its own **grid, vertical levels, and variable set**. These differences in format make it very complicated for applications to consume and compare data coming from different models in an automated and timely manner. This issue is resolved by introducing the concept of **Generic State Vector** (GSV), which defines a **common output portfolio for all models** to ensure a homogeneous output between models. The conversion from the model's native output to the GSV happens before the data is written in the HPC and it is automated in the workflow allowing **transparent access to the data** changing only the name of the model in the call.

Generic State Vector Concept

The Generic State Vector (GSV) defines the common format in which all the models have to provide the data. A **simulated day** of the GSV for a single model around **occupies around 110GB**.

The main characteristics of the GSV format are the following:

- Global data at **5km scale** (sfc and 3D)
- **Hourly** frequency for the **atmosphere**.
- **Daily** frequency for the **ocean**.
- Data stored in **GRIB2 format**.
- Same **set of variables** for all models, with the **same names and GRIB paramIds**. 66 variables in total.
- Data interpolated to **HEALPix grid**

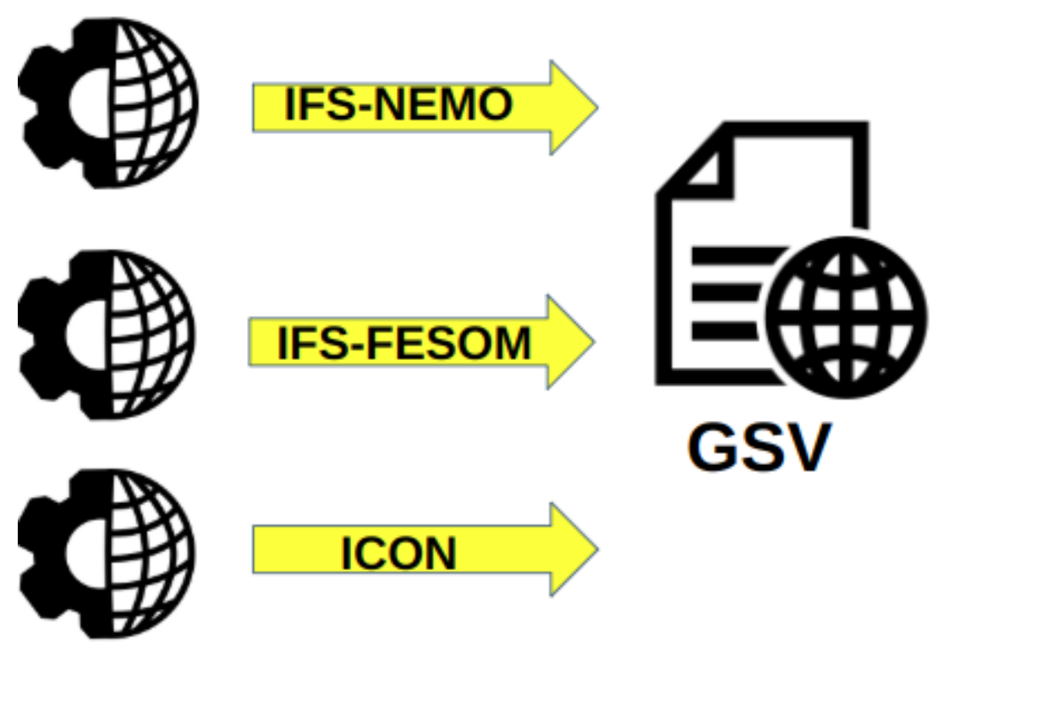


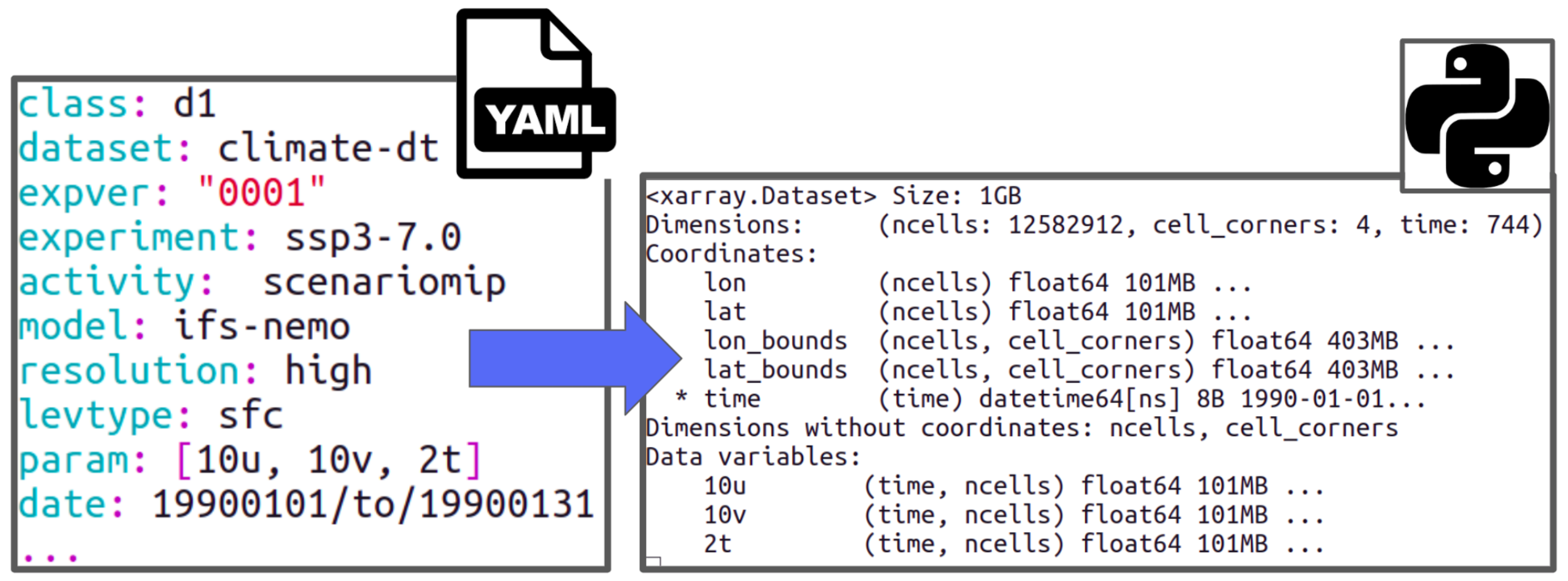
Figure 1: Each of the models native output is converted to a unified GSV format output.

GSV Interface

The GSV Interface is a dedicated Python tool developed to read the data generated by the ClimateDT models[3].

It provides an **interface** to the user to access ClimateDT data using a **MARS-like syntax**. Data is **converted from GRIB2 to xarray** on the fly. It also provides some optional utilities such as **interpolation to regular grids**, or **selection of specific areas**.

The tool is released as **Open Source software** with an Apache license, and **can be installed via pip**.



```yaml
class: d1
dataset: climate-dt
expver: "0001"
experiment: ssp3-7.0
activity:  scenariomip
model: ifs-nemo
resolution: high
levtype: sfc
param: [10u, 10v, 2t]
date: 19900101/to/19900131
...
```

```
<xarray.Dataset> Size: 1GB
Dimensions:     (ncells: 12582912, cell_corners: 4, time: 744)
Coordinates:
    lon         (ncells) float64 101MB ...
    lat         (ncells) float64 101MB ...
    lon_bounds  (ncells, cell_corners) float64 403MB ...
    lat_bounds  (ncells, cell_corners) float64 403MB ...
  * time        (time) datetime64[ns] 8B 1990-01-01...
Dimensions without coordinates: ncells, cell_corners
Data variables:
    10u        (time, ncells) float64 101MB ...
    10v        (time, ncells) float64 101MB ...
    2t         (time, ncells) float64 101MB ...
```

Figure 2: Example of a GSV request.

Data Flow Overview

Data from the ClimateDT models is stored in two different databases: the **HPC FDB** (short-term storage) and the **Databridge FDB** (long-term storage).

- The **HPC FDB** is the first location in which the models write the output data. It contains the full data with no reduction. Data in the HPC FDB can be accessed by applications running inside the ClimateDT contract in a streaming mode, where data is consumed at the same time is being generated by the models[4]. Data in the HPC FDB is only available for a **short time window**.
- From time to time, data is transferred from the HPC FDB to the **Databridge FDB** and **wiped from the HPC FDB**. The transfer happens regularly, based on disk space occupation. The Databridge FDB acts as a **long-term storage**. Some data reduction may be applied to account for storage constraints. Users **internal or external to ClimateDT** can access the data from the databridge via the **Destination Earth Service Platform (DESP)**[5].

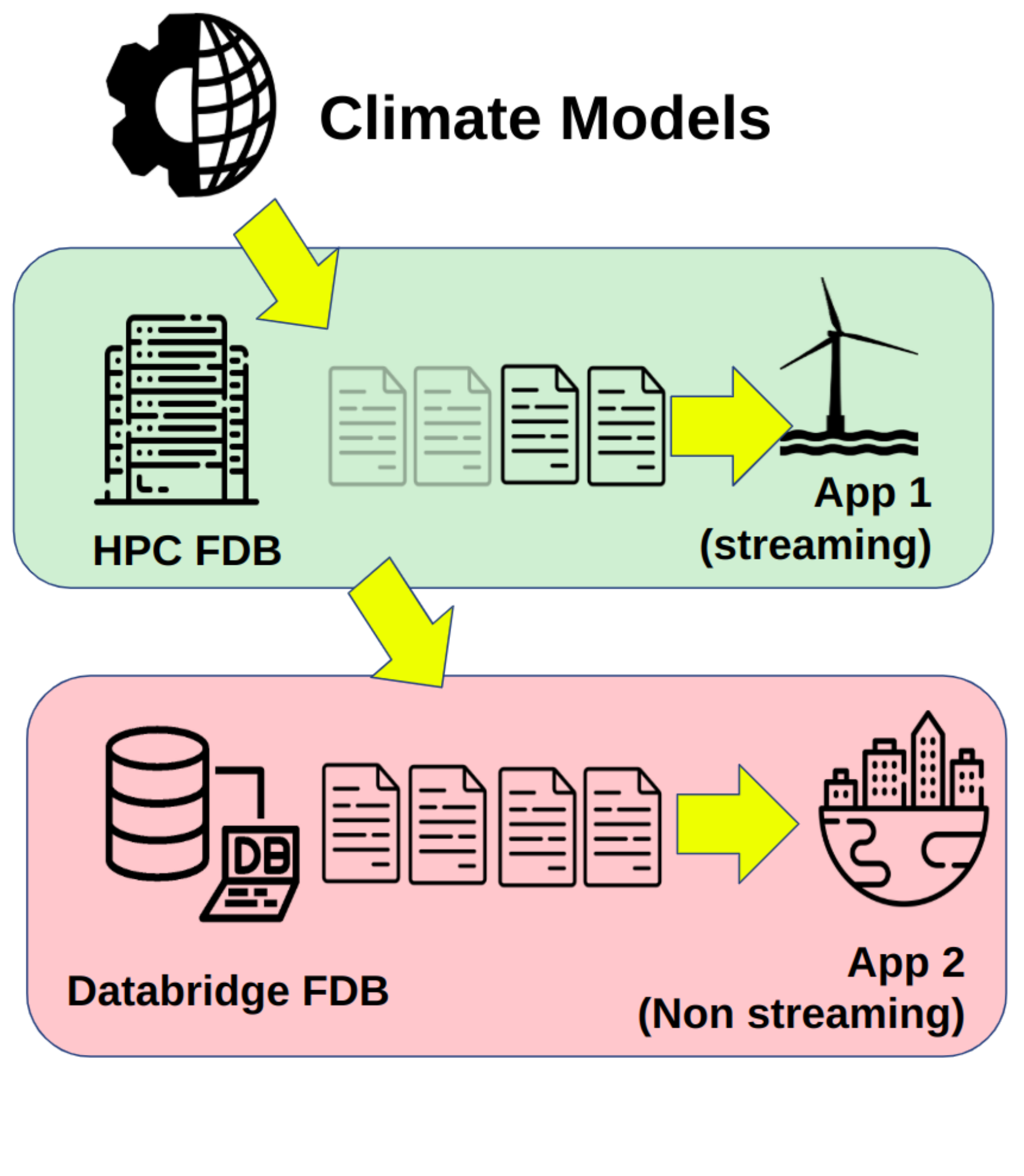


Figure 3: Overview of the data flow in the ClimateDT.

Accessing Data with GSV Interface

The GSV Interface can be used to retrieve data from the models in a user-friendly xarray format. Data is identified using its MARS keys.

Depending on the type of user, two ways of retrieving data are available:

- Users **internal** to ClimateDT can retrieve data **directly from the HPC FDB**. This allows internal app developers to run in streaming using the ClimateDT workflow.
- Users **external** to ClimateDT can retrieve data **from the databridge through DESP**.

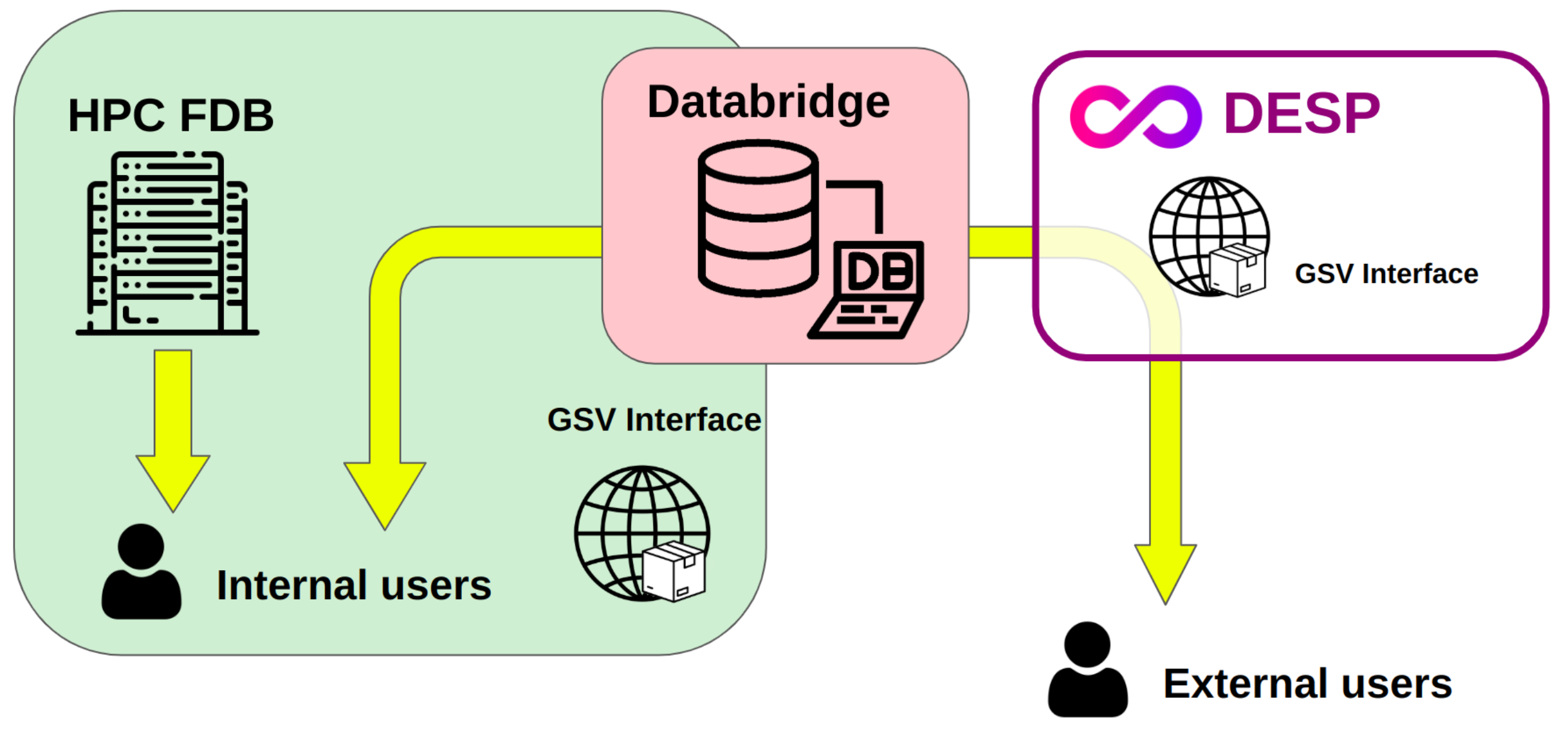


Figure 4: The GSV interface can be used to retrieve data directly from the HPC or from the databridge through DESP.

Additionally, it can also be used to read from GRIB files directly.

Conclusions

The concept of the Generic State Vector has played a key role in ClimateDT by allowing applications to transparently read from any model with a common workflow.

It allows applications to read data from different models transparently, with a uniform interface and in an automated way.

The GSV Interface is currently integrated in all the workflow components that require reading the model output (such as applications or data quality checker).

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] A. Gaya-Àvila et al. A workflow for the climate digital twin. In *EGU General Assembly 2024*, Vienna, Austria, 2024. EGU24-2533.

[3] Gsv interface python package. https://pypi.org/project/gsv-interface/. Accessed: 2025-04-24.

[4] 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. EGU24-2164.

[5] Destination earth service platform (desp). https://platform.destine.eu/. Accessed: 2025-04-24.

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