DESTINATION EARTH

Climate Digital Twin (DE340)

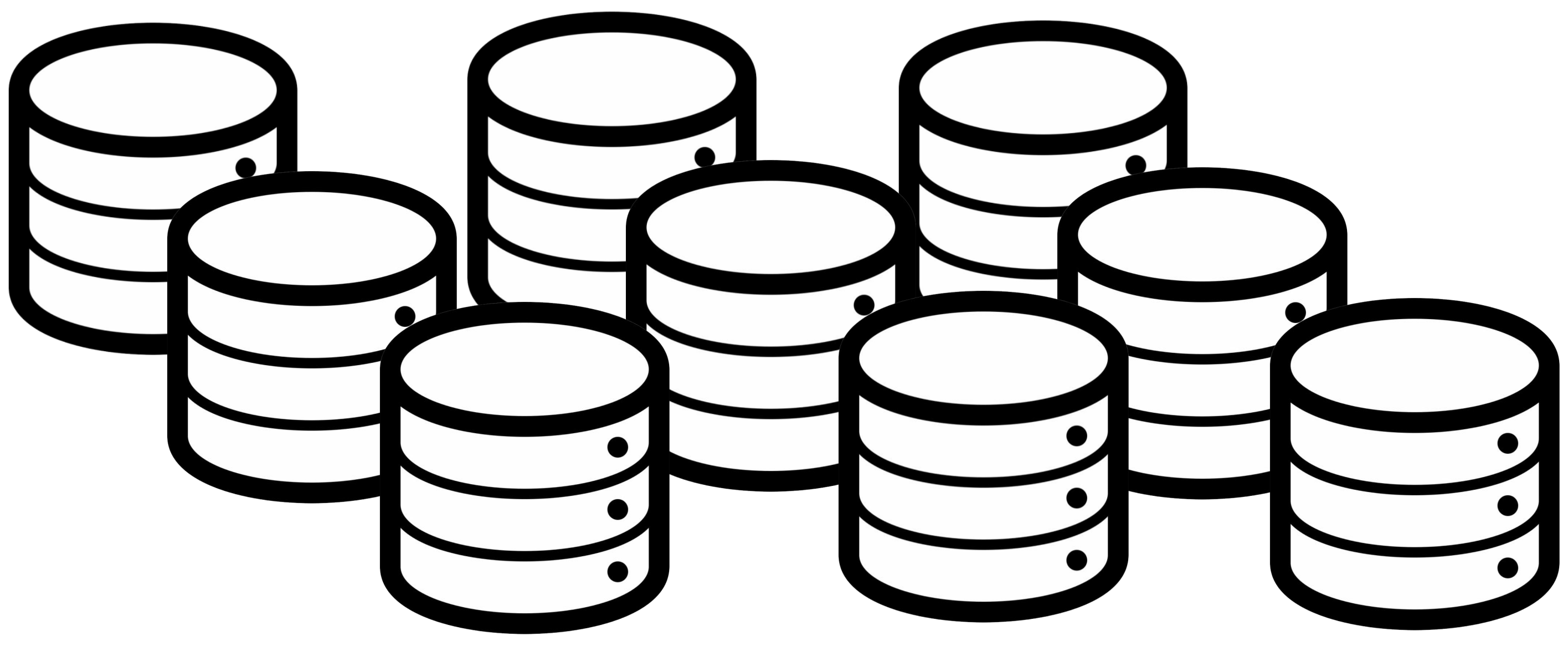
Earth Sciences Department

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

The Generic State Vector definition and the streaming concept

Roura-Adserias, F., Gonzalez, I., Grayson, K. and Lacima, A.

Why do we need streaming?

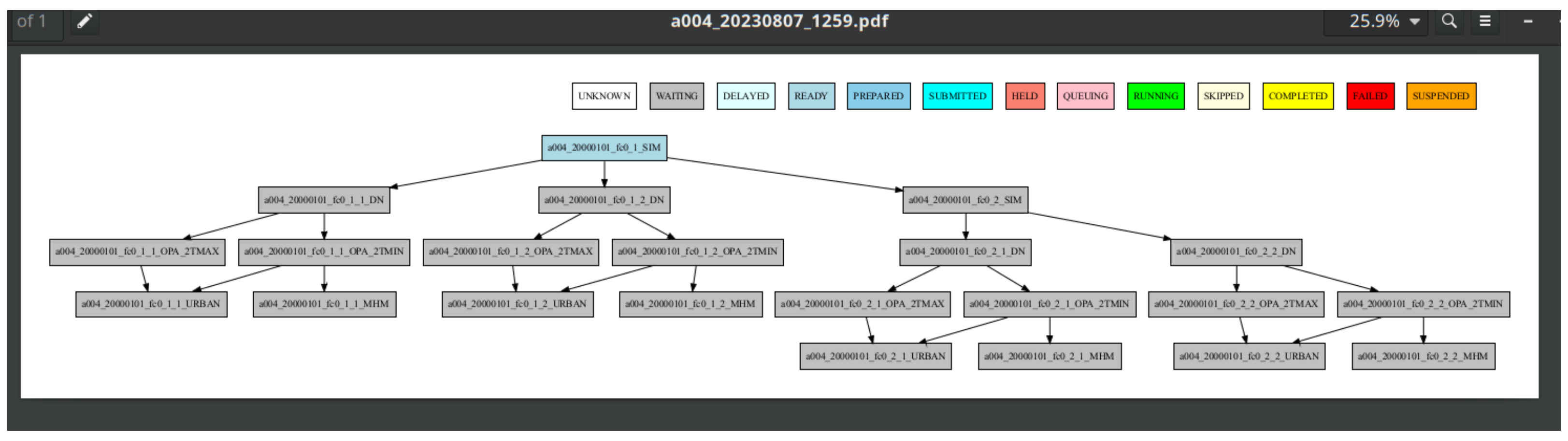


The climate crisis poses extreme threat to our society and we need tools to **take informed decisions**. In order to take these decisions, the improvement of global climate models is key. However, **high resolution models** that can be run in a more **agile and interactive way** produce an amount of data that can not be permanently stored (~1 PiB/day).

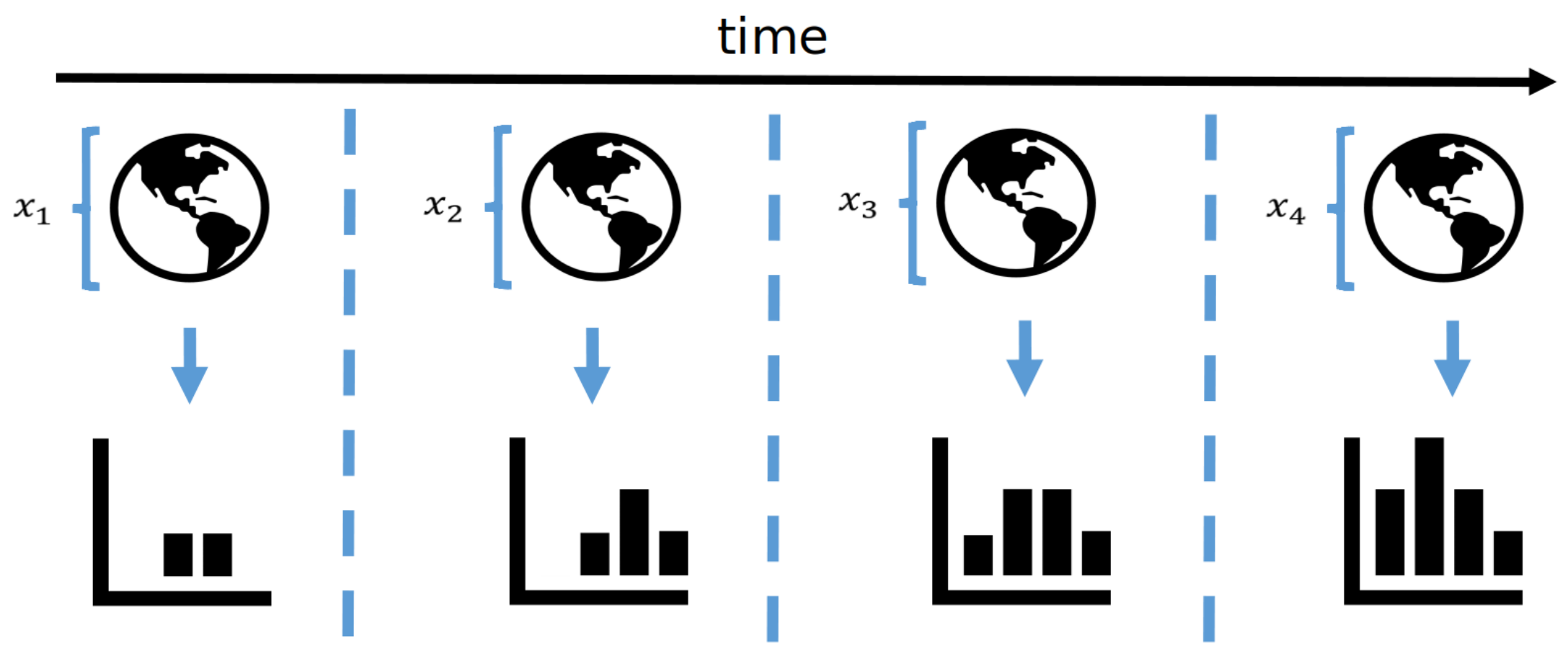
To mitigate this long-term data storage issue we employ **data streaming**; providing the data directly from the models to the users, without needing to store vast amounts of data to disk.

How do we achieve streaming?

- **Automated workflow manager**
  - Handles all the tasks in the workflow, generating model data and transferring it to the impact models (use cases)
- **Data listening mechanism**
  - Software that automatically notifies the downstream workflow that data is avalaible
- **One – pass algorithms**
  - Mathematical algorithms that compute statistics required by the user on the streamed data (storage saving).


of 1
a004_20230807_1259.pdf
25.9%
UNKNOWN
WAITING
DELAYED
READY
PREPARED
SUBMITTED
HELD
QUEUING
RUNNING
SKIPPED
COMPLETED
FAILED
SUSPENDED
a004_20000101_fc0_1_SIM
a004_20000101_fc0_1_1_DN
a004_20000101_fc0_1_2_DN
a004_20000101_fc0_2_SIM
a004_20000101_fc0_1_1_OPA_2TMAX
a004_20000101_fc0_1_1_OPA_2TMIN
a004_20000101_fc0_1_2_OPA_2TMAX
a004_20000101_fc0_1_2_OPA_2TMIN
a004_20000101_fc0_2_1_DN
a004_20000101_fc0_2_2_DN
a004_20000101_fc0_1_1_URBAN
a004_20000101_fc0_1_1_MHM
a004_20000101_fc0_1_2_URBAN
a004_20000101_fc0_1_2_MHM
a004_20000101_fc0_2_1_OPA_2TMAX
a004_20000101_fc0_2_1_OPA_2TMIN
a004_20000101_fc0_2_2_OPA_2TMAX
a004_20000101_fc0_2_2_OPA_2TMIN
a004_20000101_fc0_2_1_URBAN
a004_20000101_fc0_2_1_MHM
a004_20000101_fc0_2_2_URBAN
a004_20000101_fc0_2_2_MHM

Example of automated workflow tasks.



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

What is GSV and why is it needed?

Different **global climate models** (GCMs) write in different grids and non-uniform metadata. Using a **Generic State Vector** (GSV), we convert all the different GCMs output to a common format with a common grid, **making all the data accessible to downstream users.** Data from different models will be distinguishable only through its metadata.

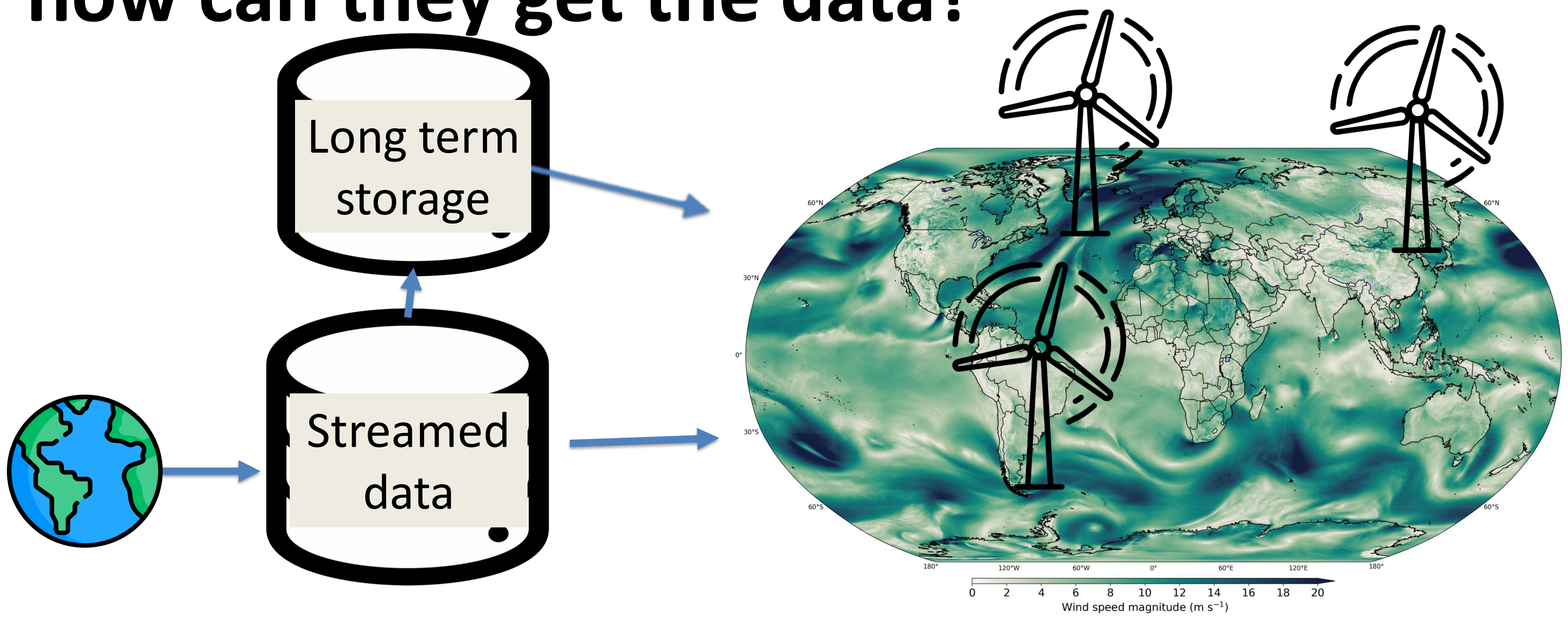
Notorious features of the streaming

- **Computer agnostic:** ClimateDT will be able to run in 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 translated into ecFlow and it follows the FAIR principles.
- **The workflow adapts to the data request.**

What can GSV interface currently do?

The GSV interface can **read experiment data** from FDB and convert it to xarray format, **from several model grids**, in HEALPix format of any resolution. Additionally, it can also read data in regular latlon, and some model native grids (tco79, tco399, tco1279, tco2559, eORCA1). It can also **interpolate** from the original grid **to a latlon grid** of choice.

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



We can go **from raw data to climate information** in a more efficient way.
We will give the users the possiblity to get data from the latest versions of the models, as well as to ask for variables that are not usually available, in an interactive way. We will provide **local information at global scale**.

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

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

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