Data spaces for climate data analysis

Ezequiel Cimadevilla Álvarez[1], Antonio S. Cofiño[1]

[1] Meteorology Group, Instituto de Física de Cantabria (IFCA, CSIC-UC), Santander, Spain

This work has been partially supported by:

- IS-ENES3 – InfraStructure for the European Network for the Earth System Modelling - IS-ENES3 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084.
- Grant PID2020-116595RB-I00 funded by MCIN/AEI/10.13039/501100011033.
- Grant PRE2021-097646 funded by MCIN/AEI/10.13039/501100011033.

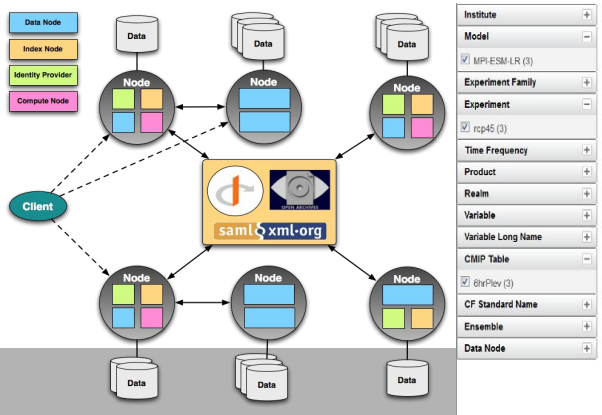
Project CORDyS (PID2020-116595RB-I00) funded by:

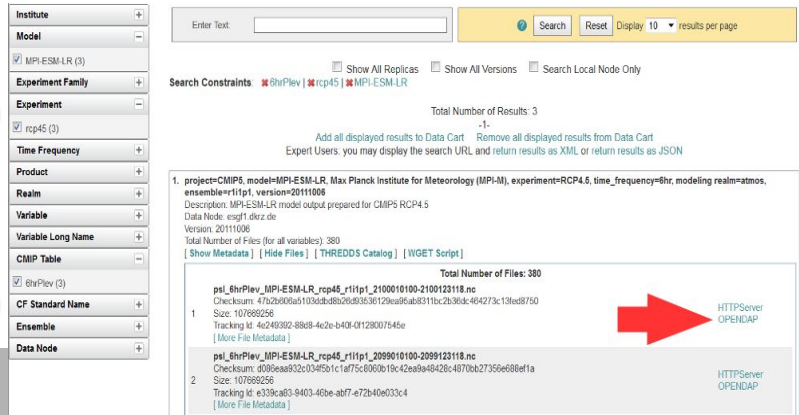
Outline

- Introduction
- HPC
- Cloud
- Data spaces
- Multidimensional data
- SantanderMetGroup Climate Data Service (SCDS)
- Conclusions

Introduction

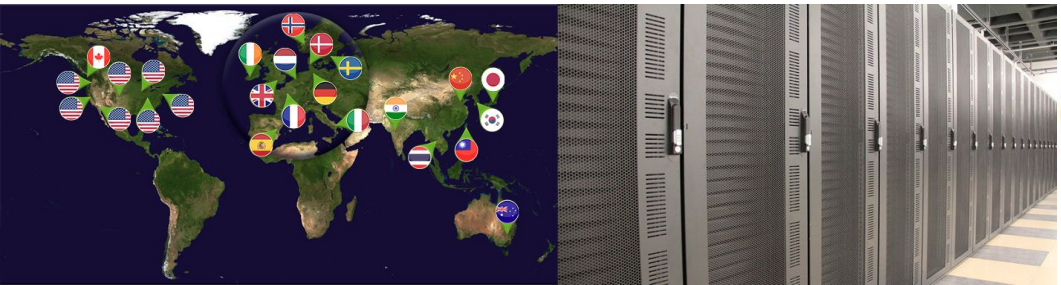
- The sheer volume of scientific data will continue to increase over the next 20 years (2012).
- Increasing demand for researchers to use each other's data, seek reproducibility in research results.
- Data stored historically either on storage media in labs or in department, university, or organizational data centers. "Download and analyze" model.
- Move computation to the data and provide remote data services.
- See Hey et al. (2012).

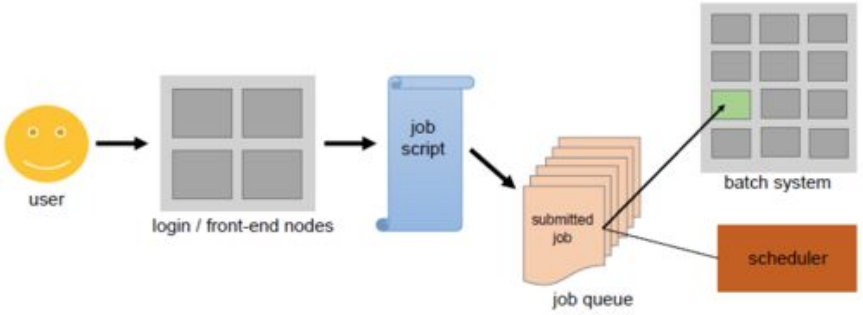




HPC

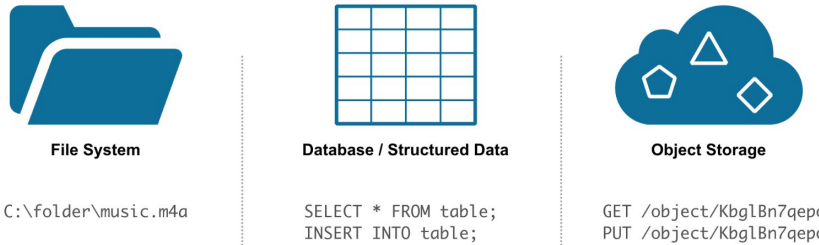
- Regarding climate science, the Coupled Model Intercomparison Project (**CMIP**) coordinates the design and distribution of global climate model simulations (20 PB of CMIP6 data).
- Sixth Assessment Report (**AR6**) of the Intergovernmental Panel on Climate Change (**IPCC**).
- Contributions to the **increase in data volume** include the systematic increase in model resolution and complexity of the experimental protocol and data request.
- The Earth System Grid Federation (**ESGF**) is a global infrastructure and network of internationally distributed sites that together work as a federated data archive, supporting the distribution of global climate model simulations.
- Data is downloaded to **local HPC infrastructures** and analyzed using job queues and ad hoc tools for data processing.





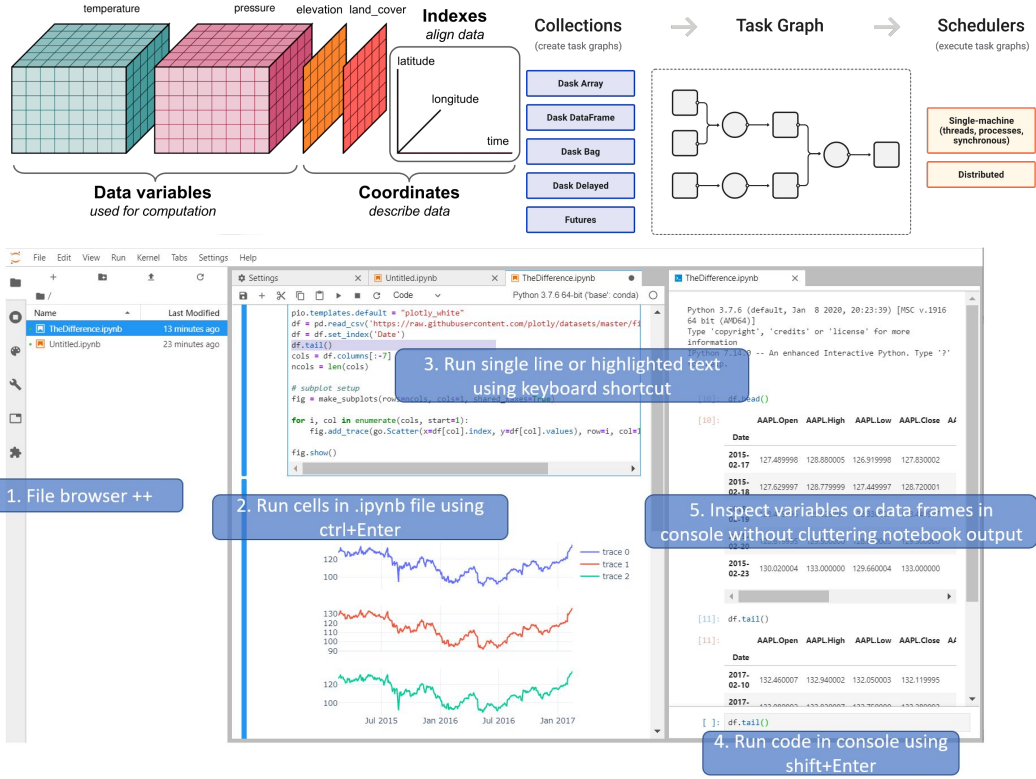
Cloud

- Cloud computing has a big potential - Created to deal with the Internet's big data challenges.
- The path to the cloud computing is unclear for majority of the science and application community.
  - New technological stack, libraries and data formats.
- Pangeo - A community that cultivates an ecosystem in which the next generation of open-source analysis tools for ocean, atmosphere and climate science can be developed, distributed, and sustained.
- Nearly a petabyte of NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC) data products have been moved to NASA's Earthdata Cloud—hosted in the Amazon Web Services (AWS) cloud.
- Google Earth Engine - Combines a multi-petabyte catalog of satellite imagery and geospatial datasets with planetary-scale analysis capabilities.



Data spaces

- Simplify access to **data** and **computation**. Move from the "Download and analyze" model.
  - Provide remote access to the data using remote data access services.
  - Provide computation services based on web interfaces (**Jupyter/Python**).
- A data space can be deployed both in HPC and cloud infrastructures.
- Improve **sophistication** and **speed** of information systems.

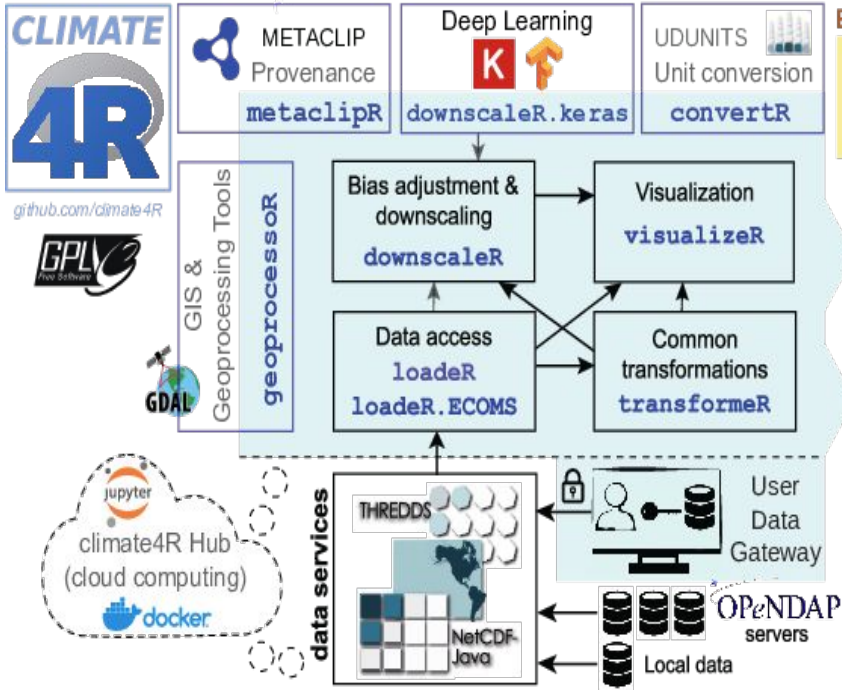


Multidimensional data

- Climate datasets are usually provided in separate files that facilitate dataset management in climate data distribution systems.
  - In ESGF a time series of a variable is split into smaller pieces of data in order to reduce file size.
- This enhances usability for data management in the ESGF distribution system (i.e. file publication and download).
- For data analysis is convenient to rearrange multiple files as a single data source, in order to obtain a “data analysis ready” **- <u>Scientific ETL process.</u>**
- **Virtual Datasets** allow to produce data analysis ready datasets with no storage penalty.

SantanderMetGroup Climate Data Service (SCDS)

- Home page: http://www.meteo.unican.es/
- Remote data access based on THREDDS data server.
  - Publicly accessible from the User Data Gateway (UDG).
- JupyterHub deployment on internal cloud.
  - Currently restricted for internal purposes.
- climate4R package for climate data analysis.
  - Integrated with the UDG.
  - Training resources available here. Launch the examples in binder.
- Use your own data analysis tools via DAP protocol (e.g. xarray).



Conclusions

- Data spaces are important tools for acceleration of climate scientific research.
  - Need to move from the “Download and analyze” model.
- Data spaces can be deployed either on HPC or cloud infrastructures.
  - Differences in ecosystems between infrastructures need to be taken into account.
- Data science and data management are fundamental disciplines for effective data spaces.
  - Scientific ETL (Extract/Transform/Load) processes.
- Data spaces are mainly implemented under the Python ecosystem.
  - Data formats - HDF5, netCDF, Zarr
  - Data analysis - numpy, xarray, pandas
  - Computing - Dask
  - Packaging - Conda
- Cloud providers data spaces - Google Colab, Kaggle, Amazon SageMaker and more.

Data spaces for climate data analysis

Ezequiel Cimadevilla Álvarez[1], Antonio S. Cofiño[1]

[1] Meteorology Group, Instituto de Física de Cantabria (IFCA, CSIC-UC), Santander, Spain

This work has been partially supported by:

- IS-ENES3 – InfraStructure for the European Network for the Earth System Modelling - IS-ENES3 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084.
- Grant PID2020-116595RB-I00 funded by MCIN/AEI/10.13039/501100011033.
- Grant PRE2021-097646 funded by MCIN/AEI/10.13039/501100011033.

Project CORDyS (PID2020-116595RB-I00) funded by: