

# Data spaces for climate data analysis



**Ezequiel Cimadevilla Álvarez<sup>1</sup>, Antonio S. Cofiño<sup>1</sup>**

<sup>1</sup> 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:

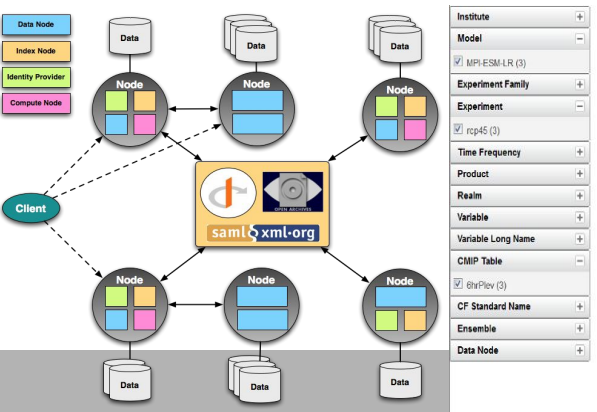


## Outline

- 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. \(2012\)](#).



Display 10 results per page

Search Constraints ■ 9hrPlev | ■ rcp45 | ■ MPI-ESM-LR

Total Number of Results: 3

-1-

Add all displayed results to Data Cart    Remove all displayed results from Data Cart

Expert Users, you may display the search URL and return results as XML or return results as JSON

---

1. project=CMIP5, model=MPI-ESM-LR, Max Planck Institute for Meteorology (MPI-M), experiment=RCP4.5, time\_frequency=6hr, modeling\_realm=atmos, ensemble=rf1r1, version=2011006

Description: MPI-ESM-LR model output prepared for CMIP5 RCP4.5

Data Node: esgf1.dkrz.de

Version: 20111006

Total Number of Files (for all variables): 380

[ Show Metadata ] [ Hide Files ] [ THREDDS Catalog ] [ WGET Script ]

---

Total Number of Files: 380

ps1\_6hrPlev\_MPI-ESM-LR\_rcp45\_r11r1\_2100010100-2100123118.nc

Checksum: a058ea932c33495b1c1a7f5c606b619c42ea9a49426c4970bb27350e6080efa

Size: 107666256

Tracking Id: 4e249992-8808-4e2e-b404-01128007544e

[ More File Metadata ] ➔ HTTPServer OPENDAP

---

ps1\_6hrPlev\_MPI-ESM-LR\_rcp45\_r11r1\_2099010100-20999123118.nc

Checksum: a058ea932c33495b1c1a7f5c606b619c42ea9a49426c4970bb27350e6080efa

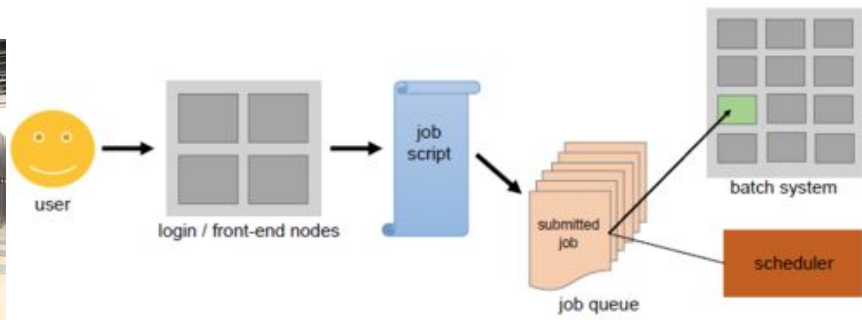
Size: 107666256

Tracking Id: e339ca83-9403-46be-ab7f-e72b40e033c4

[ More File Metadata ] ➔ HTTPServer OPENDAP

## 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.



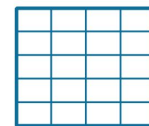
## Cloud

- 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.



File System

C:\folder\music.m4a



Database / Structured Data

```
SELECT * FROM table;  
INSERT INTO table;
```

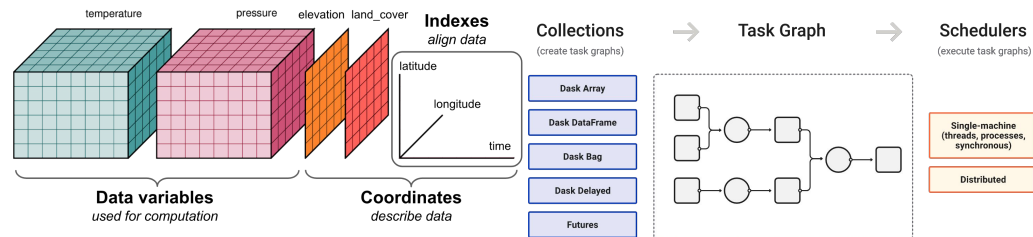


Object Storage

```
GET /object/Kbg1Bn7qepo  
PUT /object/Kbg1Bn7qepo
```

## 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.



The screenshot shows a Jupyter notebook interface with the following workflow steps highlighted:

1. File browser ++
2. Run cells in .ipynb file using `ctrl+Enter`
3. Run single line or highlighted text using keyboard shortcut
4. Run code in console using `shift+Enter`
5. Inspect variables or data frames in console without cluttering notebook output

The notebook code includes:

```

pio.templates.default = "plotly_white"
df = pd.read_csv("https://raw.githubusercontent.com/plotly/datasets/master/tesla.csv")
df.set_index("Date", inplace=True)
df.tail()
cols = df.columns[7:]
nrows = 100
ncols = len(cols)

# subplot setup
fig = make_subplots(rows=nrows, cols=ncols)

for i, col in enumerate(cols, start=1):
    fig.add_trace(go.Scatter(x=df.index, y=df[col].values), row=i, col=i)
fig.show()
    
```

The console output shows data for AAPL stock prices across different dates, such as 2015-02-17 and 2017-02-10.

## 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” - **Scientific ETL process.**
- Virtual Datasets** allow to produce data analysis ready datasets with no storage penalty.

2. CMIP6.CMIPBCC.BCC-CSM2-MR.historical.r1i1p1f1.3hr.tas.gn  
 Data Node: cmip.bcc.cma.cn  
 Version: 20181127  
 Total Number of Files (for all variables): 22  
 Full Dataset Services: [ Show Metadata ] [ Hide Files ] [ WGET Script ] [ LAS ] [ Show Citation ] [ PID ] [ Globus Download ]

Total Number of Files: 22

1 tas\_3hr\_BCC-CSM2-MR\_historical\_r1i1p1f1\_gn\_195001010000-195212312100.nc  
 checksum: b5f270ed53e3ae7cbaa362b8cc1e3961e25750fecbb07ec074bd401ec9d02748  
 size: 1794284104  
 tracking\_id: hdl:21.14100/b880830c-7104-44d5-a02f-59bd431e816d  
 [ More File Metadata ]

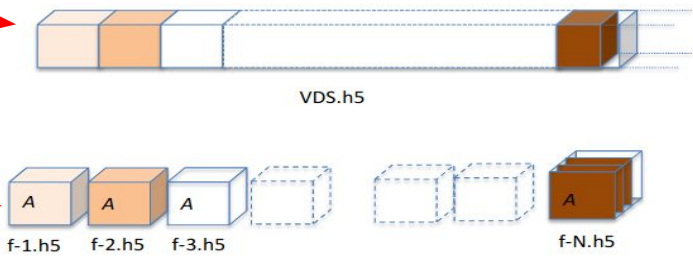
2 tas\_3hr\_BCC-CSM2-MR\_historical\_r1i1p1f1\_gn\_195301010000-195512312100.nc  
 checksum: 44d89d66384dd5be2d42fa83abeb358a25e79017f39625f3a472b8d7c5d592f  
 size: 1794284104  
 tracking\_id: hdl:21.14100/02bcfe96-5445-4454-99f3-448f161f7887  
 [ More File Metadata ]

Dataset

Files in Dataset

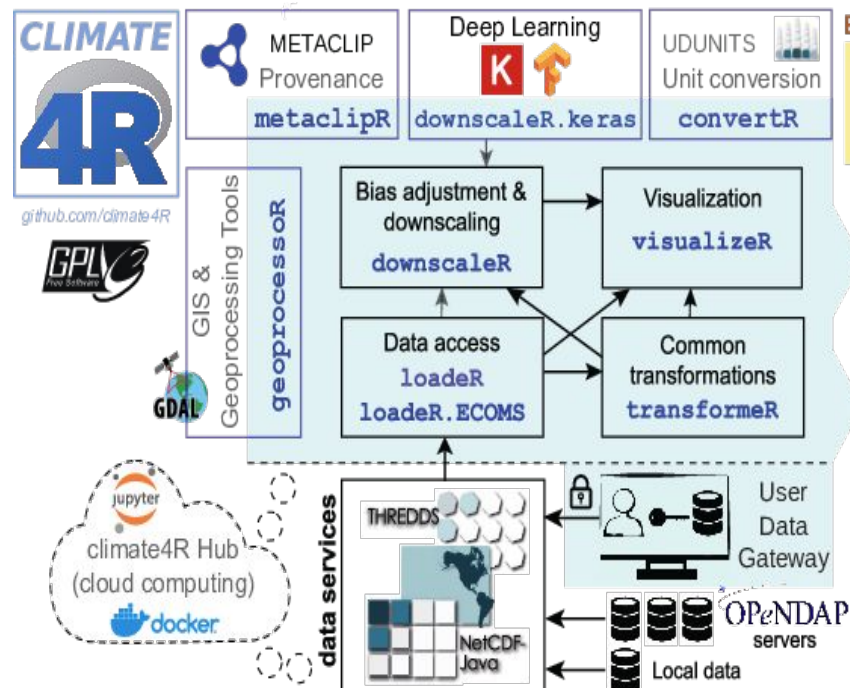
Virtual Dataset VDS

VDS.h5



## 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<sup>1</sup>, Antonio S. Cofiño<sup>1</sup>**

<sup>1</sup> 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:

