



MINISTERIO
DE CIENCIA
E INNOVACIÓN



AGENCIA
ESTATAL DE
INVESTIGACIÓN

Management of new high resolution data sets

Jesus Fernandez, Antonio S. Cofiño



CSIC

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS



UNIVERSIDAD
DE CANTABRIA



IFCA

Instituto de Física de Cantabria

Outline

- The problem
 - Typical data set size
 - Access patterns
- Data processing
 - Traditional vs newer trends
- Some practical low level details
- FAIRness and reproducibility of results

Some CP simulation data sets

- CORDEX FPS-CONV
 - Alpine domain @ 3km (10 yr evaluation + 20 yr scenario+hist, ~8 models)
 - 500 MB to 5 GB per hourly 2D variable and year
 - Forschungszentrum Jülich server
- CORDEX FPS-SESA
 - Central South America @ 4km (3 yr evaluation, 2 models)
 - 500 MB to 10 GB per hourly 2D variable and year
 - Santander MetGroup server (data.meteo.unican.es)
- WRF SAAG (South America affinity group)
 - South American continent @ 4km (20 yr evaluation, 1 model)
 - 100 GB per hourly 2D variable and year
 - NCAR globus server
- EUCP CP simulations
 - Multiple European domains @ 3km. Currently on private DMI server

FPS-CONV

<https://doi.org/10.1007/s00382-021-05708-w>

“Due to the large amount of data produced by these kilometer-scale simulations, the analysis and the calculation of the indices is performed by each group individually using scripts provided by the corresponding author. Only the final results have been shared.”

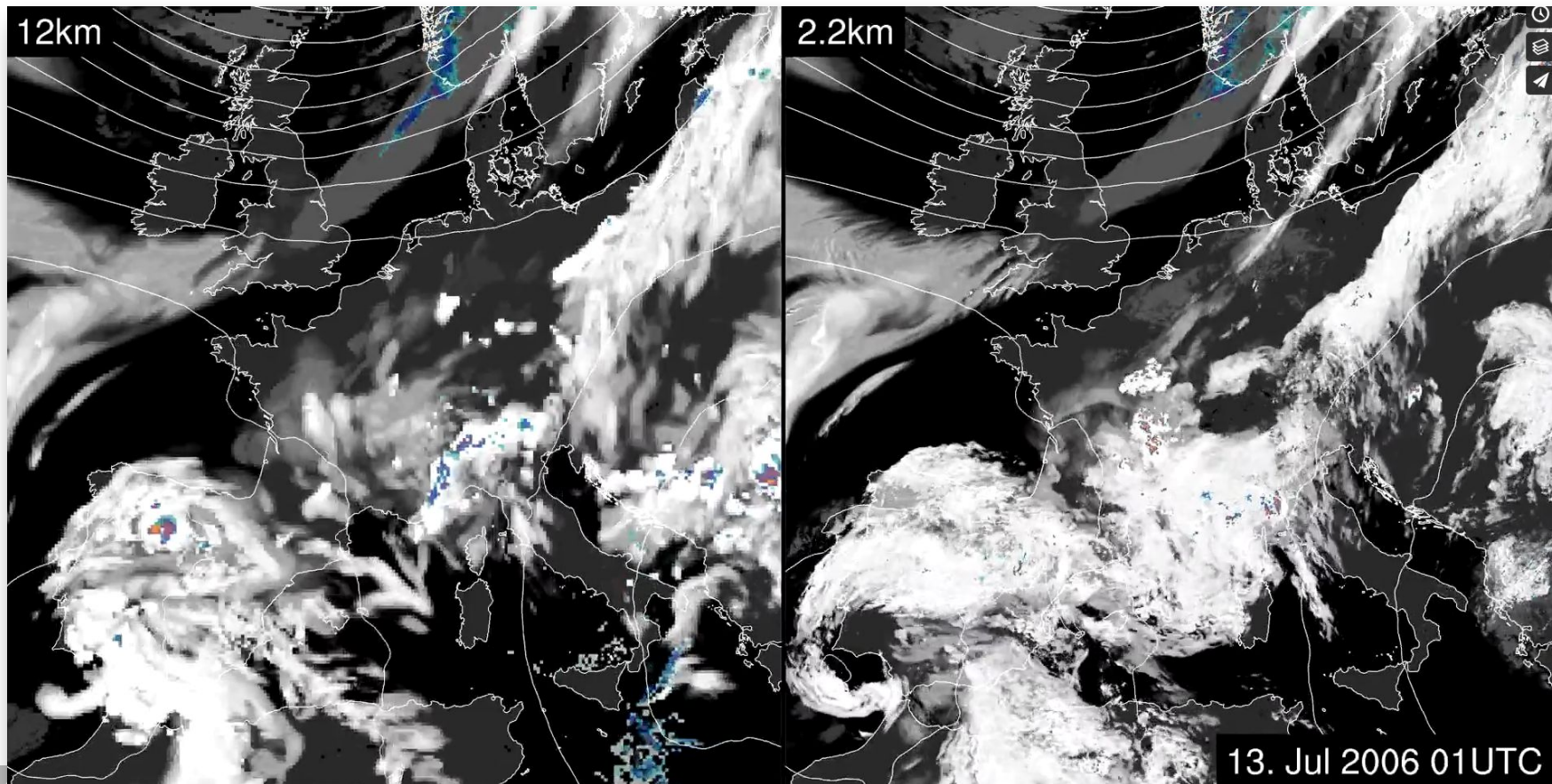
Climate Dynamics (2021) 57:275–302
<https://doi.org/10.1007/s00382-021-05708-w>

The first multi-model ensemble of regional climate simulations at kilometer-scale resolution, part I: evaluation of precipitation

Nikolina Ban¹ · Cécile Caillaud² · Erika Coppola³ · Emanuela Pichelli³ · Stefan Sobolowski⁴ · Marianna Adinolfi⁵ · Bodo Ahrens⁶ · Antoinette Alias² · Ivonne Anders⁷ · Sophie Bastin⁸ · Danijel Belušić⁹ · Ségolène Berthou¹⁰ · Erwan Brisson² · Rita M. Cardoso¹¹ · Steven C. Chan¹² · Ole Bøssing Christensen¹³ · Jesús Fernández¹⁴ · Lluís Fita¹⁵ · Thomas Frisius¹⁶ · Goran Gašparac¹⁷ · Filippo Giorgi³ · Klaus Goergen^{18,19} · Jan Erik Hauqen²⁰ ·

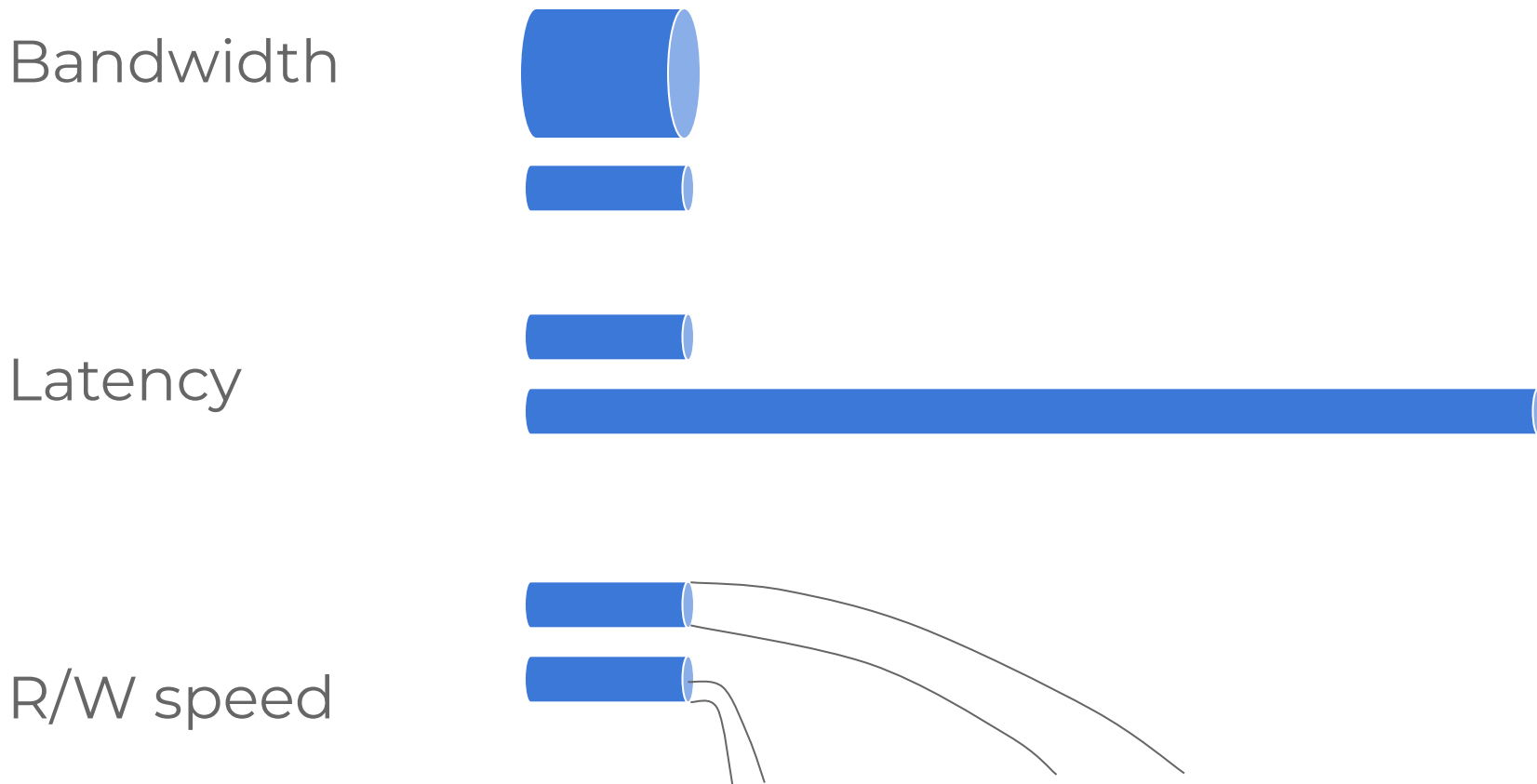
Detail beyond eyesight

Source: [ETHZ crCLIM gallery](#)



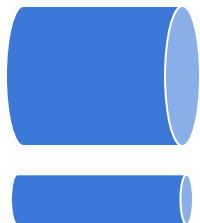
13. Jul 2006 01UTC

Data access: transfer speed limiting factors



Data access: transfer speed limiting factors

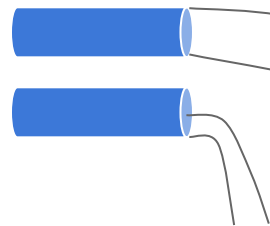
Bandwidth

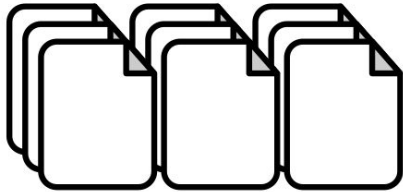
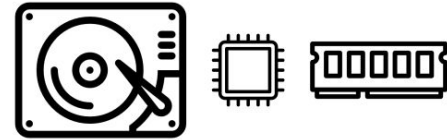


Latency

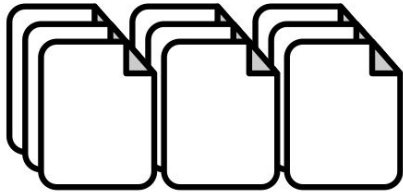
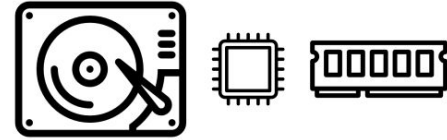


R/W speed

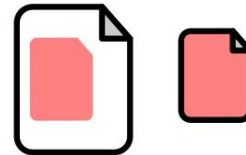
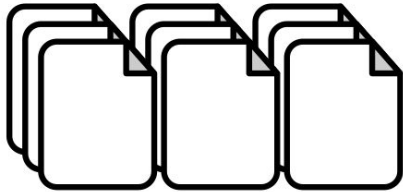
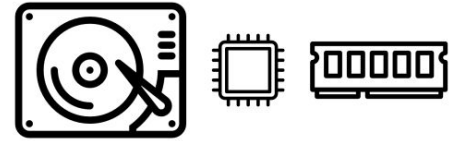




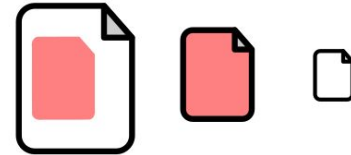
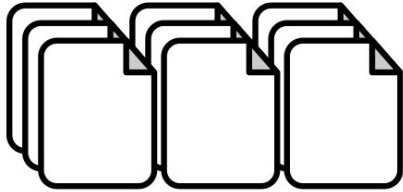
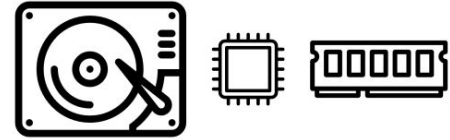
This image has been designed using resources from Flaticon.com



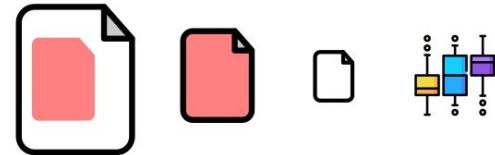
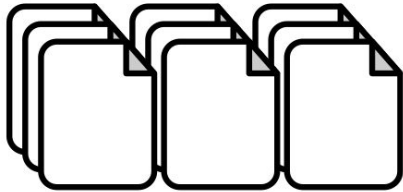
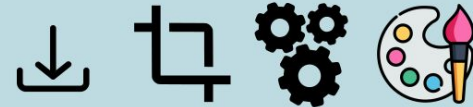
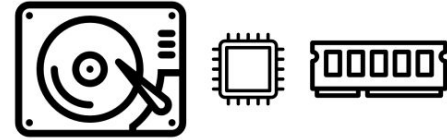
This image has been designed using resources from Flaticon.com



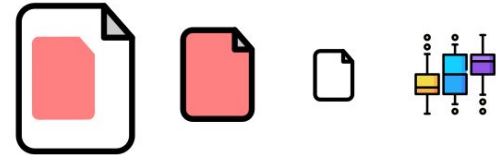
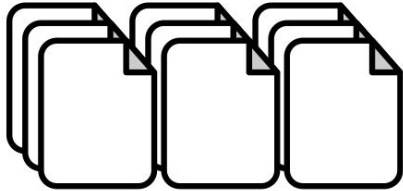
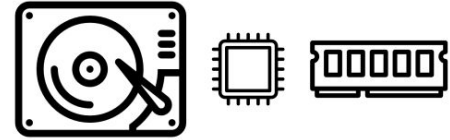
This image has been designed using resources from Flaticon.com



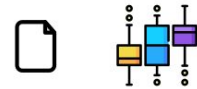
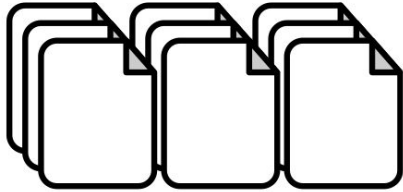
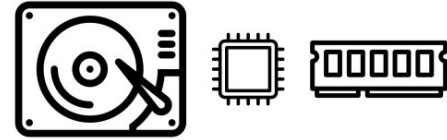
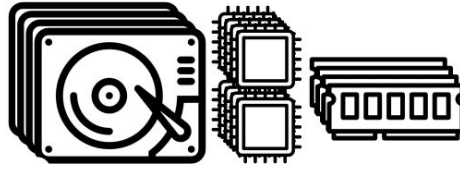
This image has been designed using resources from Flaticon.com



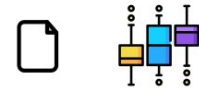
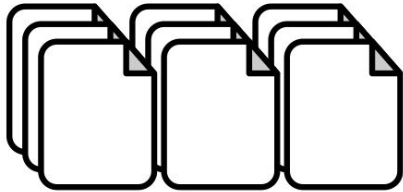
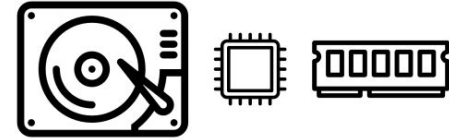
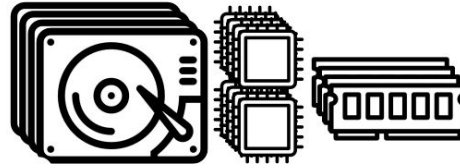
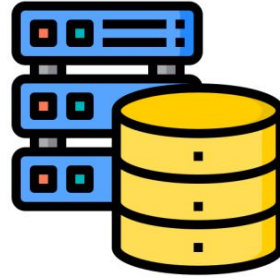
This image has been designed using resources from Flaticon.com



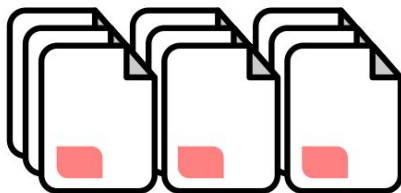
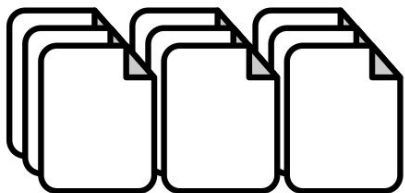
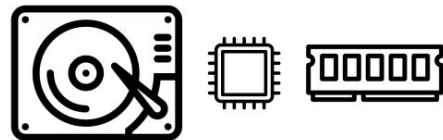
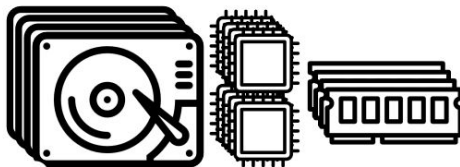
This image has been designed using resources from Flaticon.com



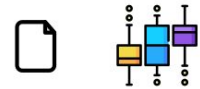
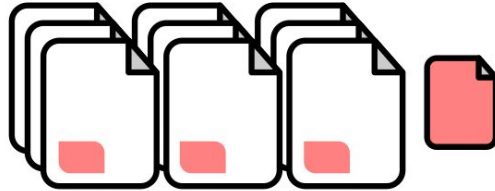
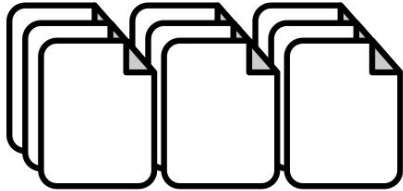
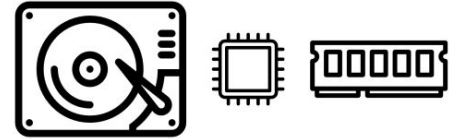
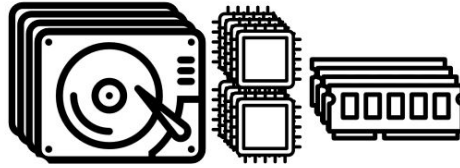
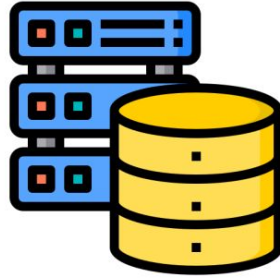
This image has been designed using resources from Flaticon.com



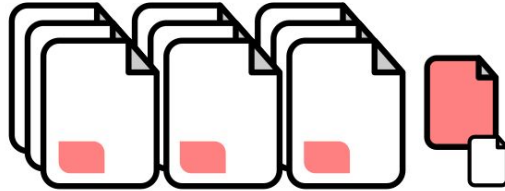
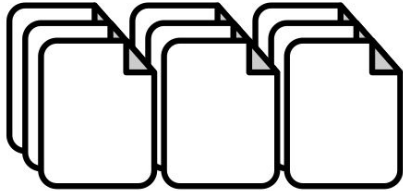
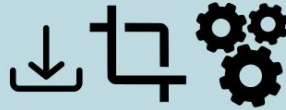
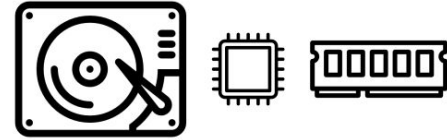
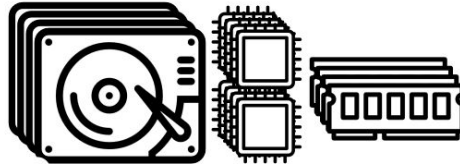
This image has been designed using resources from Flaticon.com



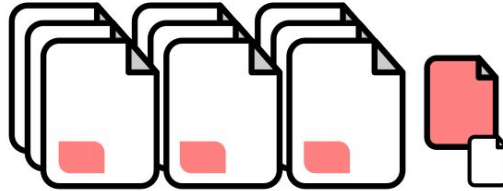
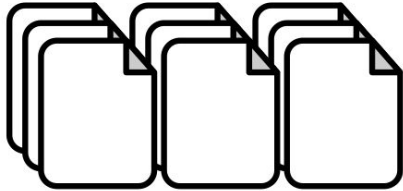
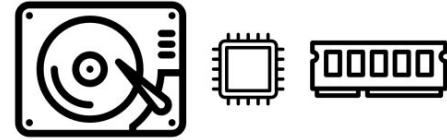
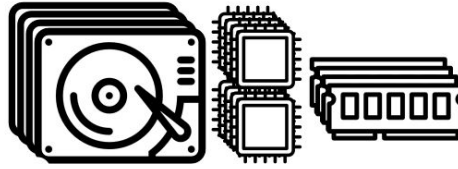
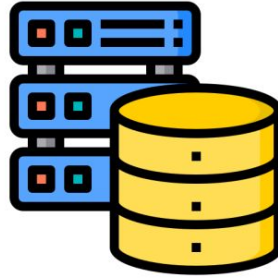
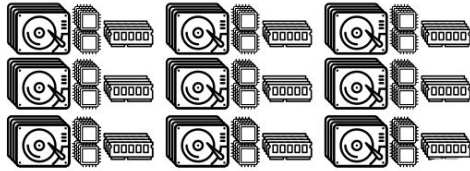
This image has been designed using resources from Flaticon.com



This image has been designed using resources from Flaticon.com

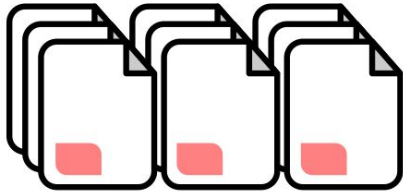
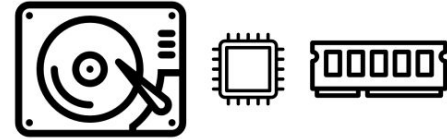
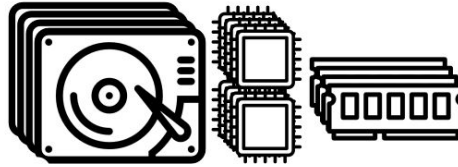
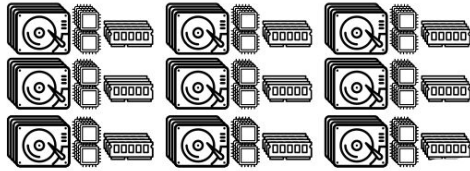


This image has been designed using resources from Flaticon.com



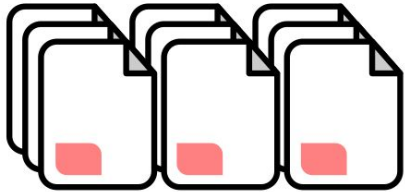
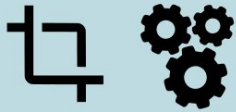
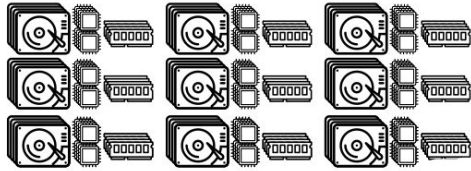
This image has been designed using resources from Flaticon.com

NCAR Research Data Archive (RDA) Copernicus CDS

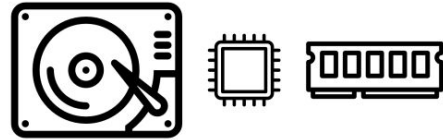
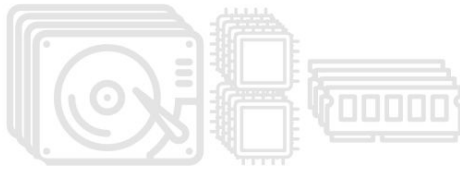


This image has been designed using resources from Flaticon.com

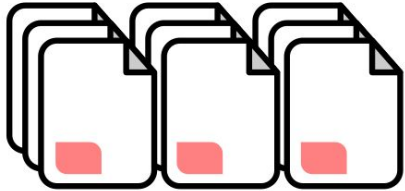
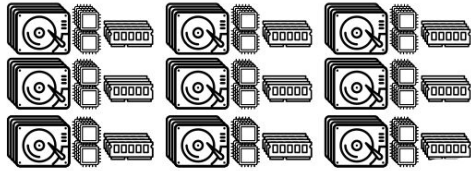
E.g. FPS-CONV FZJülich server



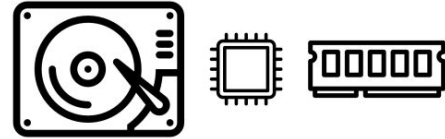
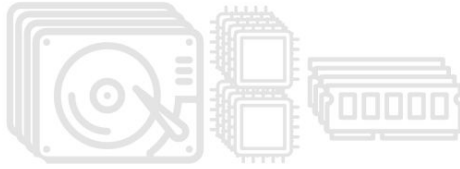
This image has been designed using resources from Flaticon.com



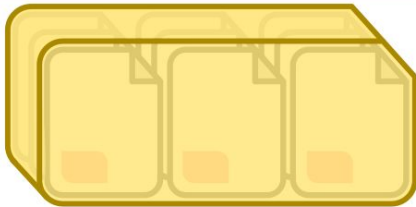
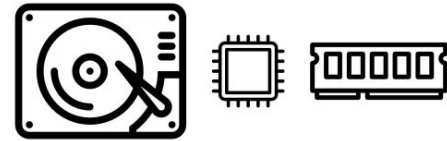
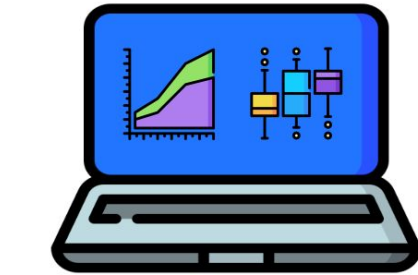
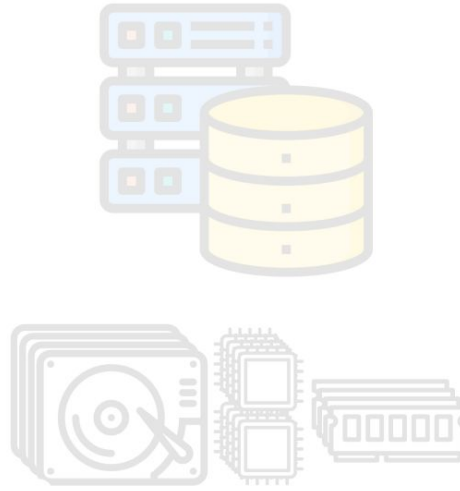
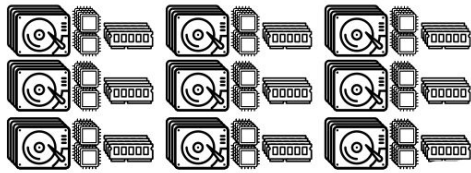
E.g. FPS-CONV FZJülich server



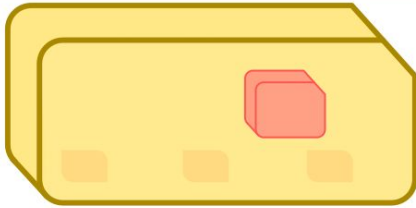
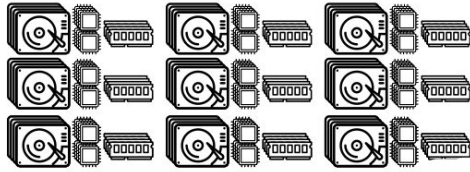
This image has been designed using resources from Flaticon.com



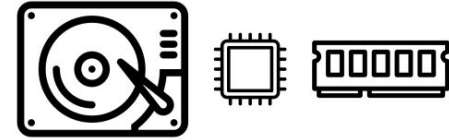
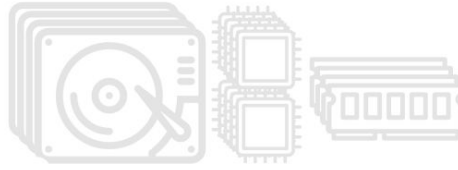
E.g. JASMIN (CEDA, UK)



This image has been designed using resources from Flaticon.com



This image has been designed using resources from Flaticon.com



Online Evaluation Dashboard

South America Affinity Group (SAAG) 4-km Test Simulations WRF Simulation Precipitation Evaluation

year

correction (1) data

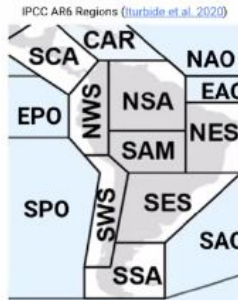
Type to search

undercatch corrected 3.9

original 3.6

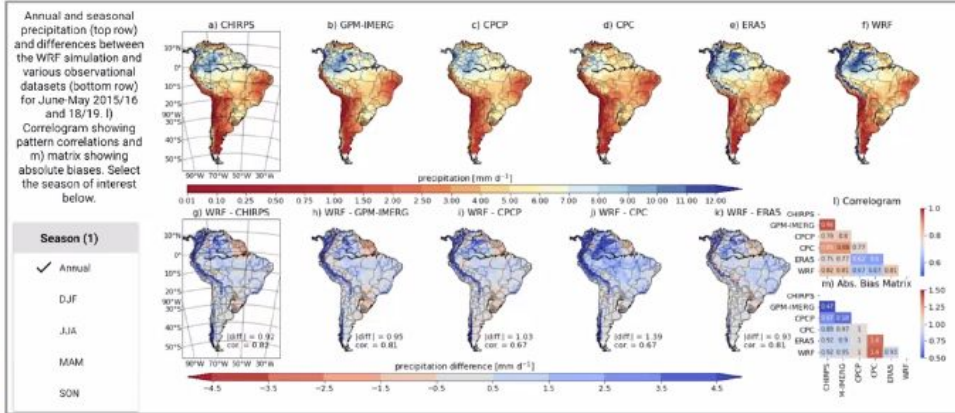
basin name

- NSA
- EAO
- SCA
- NWS
- SAM
- CAR
- NES
- SES
- NAO
- SSA
- SWS
- SPO
- EPO

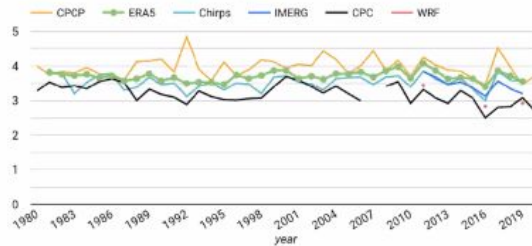


CAR - Caribbean
 SCA - S. Central-America
 NWS - N.W. South-America
 NSA - N. South-America
 SAM - South-America-Monsoon
 NES - N.E. South-America
 SES - S.E. South-America
 SWS - S.W. South-America
 SPO - S. South America

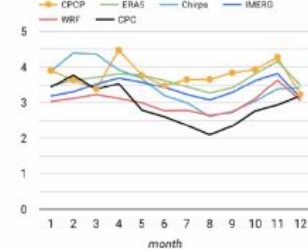
EPO - Equatorial Pacific Ocean
 SPO - S. Pacific-Ocean
 NAO - N. Atlantic-Ocean
 EAO - Equatorial Atlantic-Ocean
 SAO - S. Atlantic-Ocean



Annual average precipitation [mm/d] over Level 1 Ecoregions in South America (select region on the right). Averages run from June to May of the consecutive year. Undercatch corrected data can be shown by checking the box on the right.



Monthly average precipitation [mm/d] in Level 1 Ecoregions. Undercatch corrected data can be shown by checking the box on the right.



Contact: Andreas F. Prein (prein@ucar.edu)
 © University Corporation for Atmospheric Research (UCAR)

NCAR
 UCAR

<https://datastudio.google.com/reporting/33013d29-b61e-49d4-85f3-51efd96b7739>

Some low level details...

to work efficiently with climate data

NetCDF



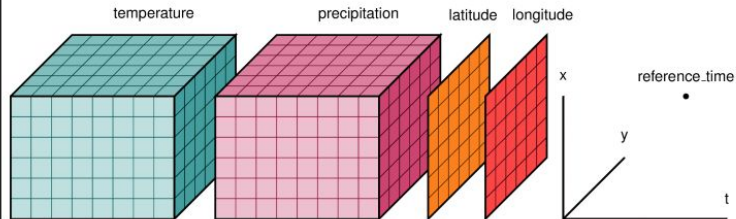
Software libraries and machine-independent data ~~format~~ **model** for array-oriented scientific data.

- Interfaces in many programming languages
- Self-describing via metadata
- Header and data
- Dimensions, coordinates, variables and attributes
- Lazy loading
- Native compression (lossless and lossy)
- Backward compatibility

NetCDF



```
netcdf pr_CSAM-4i_evaluation_UCAN-WRF433_1hr {
  dimensions:
    time = 3624 ;
    lon = 676 ;
    lat = 451 ;
  variables:
    double time(time) ;
      time:standard_name = "time" ;
      time:long_name = "Time" ;
      time:units = "days since 1949-12-01T00:00:00Z" ;
      time:calendar = "standard" ;
      time:axis = "T" ;
    double lon(lon) ;
      lon:standard_name = "longitude" ;
      lon:axis = "X" ;
      lon:long_name = "Longitude" ;
      lon:units = "degrees_east" ;
    double lat(lat) ;
      lat:standard_name = "latitude" ;
      lat:axis = "Y" ;
      lat:long_name = "Latitude" ;
      lat:units = "degrees_north" ;
    float pr(time, lat, lon) ;
      pr:standard_name = "precipitation_flux" ;
      pr:long_name = "Precipitation" ;
      pr:units = "kg m-2 s-1" ;
      pr:missing_value = 1.e+20f ;
}
```





NetCDF: storage formats

Software libraries and machine-independent data ~~format~~ **model** for array-oriented scientific data.

```
ncdump -k file.nc
```

[CDF-1] classic

[CDF-2] 64-bit offset (version ≥ 3.6)

[HDF5] netCDF-4 (version ≥ 4.0) and netCDF-4 classic model

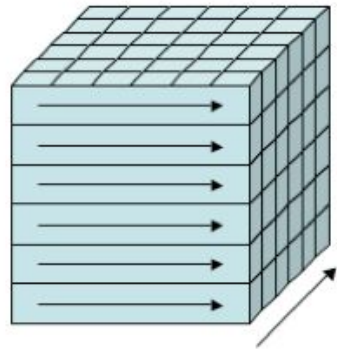
[CDF-5] 64-bit data (version ≥ 4.4) parallel

[Zarr] NCZarr (version ≥ 4.8) on S3 cloud storage

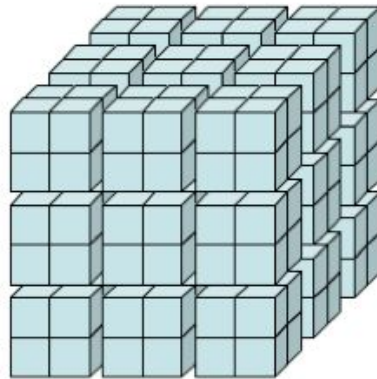
[https://docs.unidata.ucar.edu/ ... netcdf_format](https://docs.unidata.ucar.edu/...netcdf_format)

NetCDF chunking

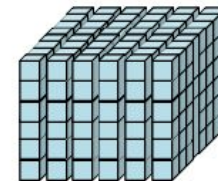
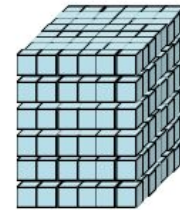
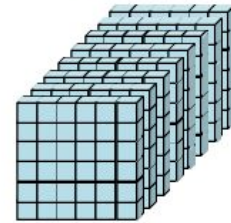
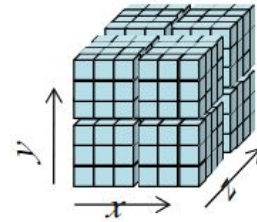
NetCDF-4 (classic or extended data model) allows for HDF5 chunks and compression filters of the data



index order



chunked



Source: Russ Rew (Uni

[https://www.unidata.ucar.edu/ ... chunking_data_why_it_matters](https://www.unidata.ucar.edu/... chunking_data_why_it_matters)



NetCDF chunking

```
$ ncdump -h -s file.nc
netcdf pr_CSAM-4i_evaluation_UCAN-WRF433_1hr {
dimensions:
    time = 3624 ;
    lon = 676 ;
    lat = 451 ;
Variables:

[...]

float pr(time, lat, lon) ;
    pr:standard_name = "precipitation_flux" ;
    pr:long_name = "Precipitation" ;
    pr:units = "kg m-2 s-1" ;
    pr:_FillValue = 1.e+20f ;
    pr:missing_value = 1.e+20f ;
    pr:cell_methods = "time: mean" ;
    pr:_Storage = "chunked" ;
    pr:_ChunkSizes = 168, 50, 50 ;
    pr:_DeflateLevel = 4 ;
    pr:_Shuffle = "true" ;
    pr:_Endianness = "little" ;
    pr:_NoFill = "true" ;
```

xarray



xarray borrows the NetCDF data model to annotate raw (NumPy) multidimensional arrays in the form of dimensions, coordinates and attributes.

```
xarray.Dataset
```

► Dimensions: (lat: 192, lon: 288, nbnd: 2, time: 600)

▼ Coordinates:

lat	(lat)	float64	-90.0 -89.06 -88.12 ... 89.06 90.0	📄	🗄️
lon	(lon)	float64	0.0 1.25 2.5 ... 356.2 357.5 358.8	📄	🗄️
time	(time)	object	1850-01-15 12:00:00 ... 1899-12-15 12:00:00	📄	🗄️

▼ Data variables:

time_bnds	(time, nbnd)	object	...	📄	🗄️
lat_bnds	(lat, nbnd)	float64	...	📄	🗄️
lon_bnds	(lon, nbnd)	float64	...	📄	🗄️
tas	(time, lat, lon)	float32	243.24796 243.24796 ... 247.15646	📄	🗄️

▼ Attributes:

Conventions :	CF-1.7 CMIP-6.2
---------------	-----------------

xarray



xarray borrows the NetCDF data model to annotate raw (NumPy) multidimensional arrays in the form of dimensions, coordinates and attributes.

It facilitates concise and error-free programming

```
ds.tas.mean('time')
ds.tas.groupby('time.season').mean('time')
ds.tas.sel(time = '2022-09-10')
ds.tas.sel(time = slice('2021-01-01', '2021-12-31'))
```

Dask



Dask is a python library implementing data collections such as parallel arrays, dataframes, and lists that extend common interfaces like NumPy, Pandas, or Python iterators to **larger-than-memory** or **distributed** environments. These parallel collections run on top of dynamic task schedulers.

Dask



Computations are lazy, just converted to task graphs

Collections

(create task graphs)



Task Graph



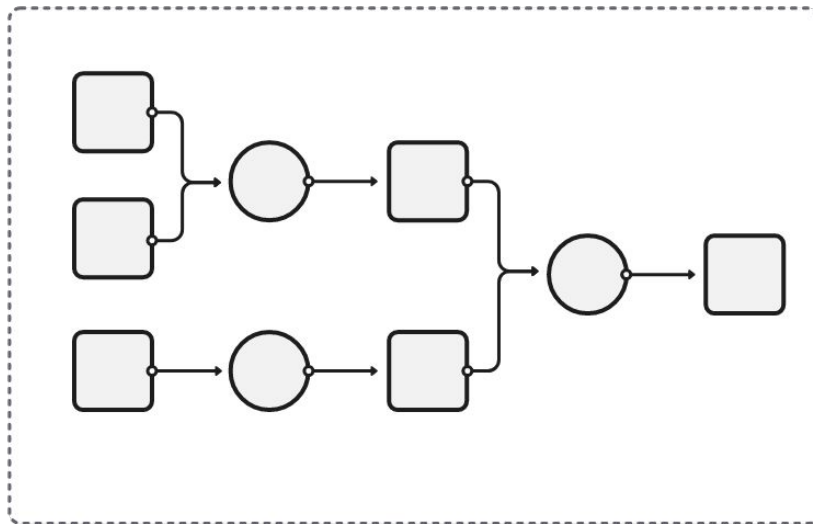
Schedulers

(execute task graphs)

`m = pr.mean('time')`

`m.compute()`

- Dask Array
- Dask DataFrame
- Dask Bag
- Dask Delayed
- Futures



Single-machine
(threads, processes,
synchronous)

Distributed

Source: <https://docs.dask.org>

Dask

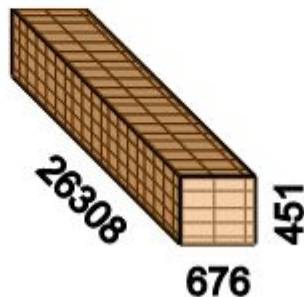


```
ds = xr.open_dataset(urls['UCAN-WRF433'], chunks = dict(time = 200, lon = 300, lat = 100))  
ds.pr
```

xarray.DataArray 'pr' (time: 26308, lat: 451, lon: 676)



	Array	Chunk
Bytes	29.88 GiB	22.89 MiB
Shape	(26308, 451, 676)	(200, 100, 300)
Count	1981 Tasks	1980 Chunks
Type	float32	numpy.ndarray



Dask

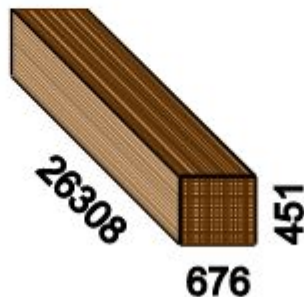


```
ds = xr.open_dataset(urls['UCAN-WRF433'], chunks = dict(lon = 30, lat = 30))
ds.pr
```

xarray.DataArray 'pr' (time: 26308, lat: 451, lon: 676)



	Array	Chunk
Bytes	29.88 GiB	90.32 MiB
Shape	(26308, 451, 676)	(26308, 30, 30)
Count	369 Tasks	368 Chunks
Type	float32	numpy.ndarray



[https://blog.dask.org/ ... choosing-dask-chunk-sizes](https://blog.dask.org/...choosing-dask-chunk-sizes)

Remote access

OPeNDAP (opendap.org)

Open-source Project for a Network Data Access Protocol (DAP)

DAP2 is a discipline-neutral means of requesting and providing data across the World Wide Web (HTTP).

The NetCDF-C library has a built-in DAP2 client

Drawbacks:

- Slow for large requests (it is not magic, it's remote)
- Potential unavailability (as any remote resource)

Cache your requests: Explore, request once and analyze many

Remote access

OPeNDAP (opendap.org)

Open-source Project for a Network Data Access Protocol (DAP)

DAP2 is a discipline-neutral means of requesting and providing data across the World Wide Web (HTTP).

The NetCDF-C library has a built-in DAP2 client

Drawbacks:

- Slow for large requests (it is not magic, it's remote)

```
$ ncdump -h http://dap-server.org/dataset.nc?time[10:1:20],lat[100:1:200],lon[100:1:300],  
var[10:1:20][100:1:200][100:1:300]
```

Remote access

OPeNDAP (opendap.org)

Open-source Project for a Network Data Access Protocol (DAP)

DAP2 is a discipline-neutral means of requesting and providing data across the World Wide Web (HTTP).

The NetCDF-C library has a built-in DAP2 client

Drawbacks:

- Slow for large requests (it is not magic, it's remote)

```
$ ncview http://dap-server.org/dataset.nc?time[10:1:20],lat[100:1:200],lon[100:1:300],  
var[10:1:20][100:1:200][100:1:300]
```


Remote access

OPeNDAP (opendap.org)

Open-source Project for a Network Data Access Protocol (DAP)

DAP2 is a discipline-neutral means of requesting and providing data across the World Wide Web (HTTP).

The NetCDF-C library has a built-in DAP2 client

Drawbacks:

- Slow for large requests (it is not magic, it's remote)

```
$ ncks ... http://dap-server.org/dataset.nc?time[10:1:20],lat[100:1:200],lon[100:1:300],  
var[10:1:20][100:1:200][100:1:300]
```

Remote access

OPeNDAP (opendap.org)

Open-source Project for a Network Data Access Protocol (DAP)

DAP2 is a discipline-neutral means of requesting and providing data across the World Wide Web (HTTP).

The NetCDF-C library has a built-in DAP2 client

Drawbacks:

- Slow for large requests (it is not magic, it's remote)

```
$ cdo info http://dap-server.org/dataset.nc?time[10:1:20],lat[100:1:200],lon[100:1:300],  
var[10:1:20][100:1:200][100:1:300]
```

Remote access



is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR CLIMATE DATA MODELING



Welcome, Guest. | Login | Create Account



You are at the **ESGF-DATA.DKRZ.DE** node

Home

Technical Support

- Project +
- Product +
- Domain -
- SAM-44 (36)
- Institute -
- SMHI (36)
- Driving Model +
- Experiment +
- Experiment Family +
- Ensemble +
- RCM Model +
- Downscaling Realisation +

Enter Text:

? Display results per page [\[More Search Options \]](#)

Show All Replicas Show All Versions Search Local Node Only (Including All Replicas)

Search Constraints: ~~×~~ SAM-44 | ~~×~~ pr | ~~×~~ 3hr | ~~×~~ SMHI

Total Number of Results: 36

-1- 2 3 4 Next >>

Please login to add search results to your Data Cart

Expert Users: you may display the search URL and [return results as XML](#) or [return results as JSON](#)

1. **cordex.output.SAM-44.SMHI.CSIRO-QCCCE-CSIRO-Mk3-6-0.historical.r1i1p1.RCA4.v3.3hr.pr**
 Data Node: esg-dn1.nsc.liu.se
 Version: 20180227
 Total Number of Files (for all variables): 55
 Full Dataset Services: [\[Show Metadata \]](#) [\[List Files \]](#) [\[THREDDS Catalog \]](#) [\[WGET Script \]](#)
2. **cordex.output.SAM-44.SMHI.IPSL-IPSL-CM5A-MR.rcp45.r1i1p1.RCA4.v3.3hr.pr**
 Data Node: esg-dn1.nsc.liu.se
 Version: 20180227

Go FAIR!

F
A
I
R



<https://www.go-fair.org/fair-principles>

Image: Australian National Data Service (ANDS)

Reproducibility

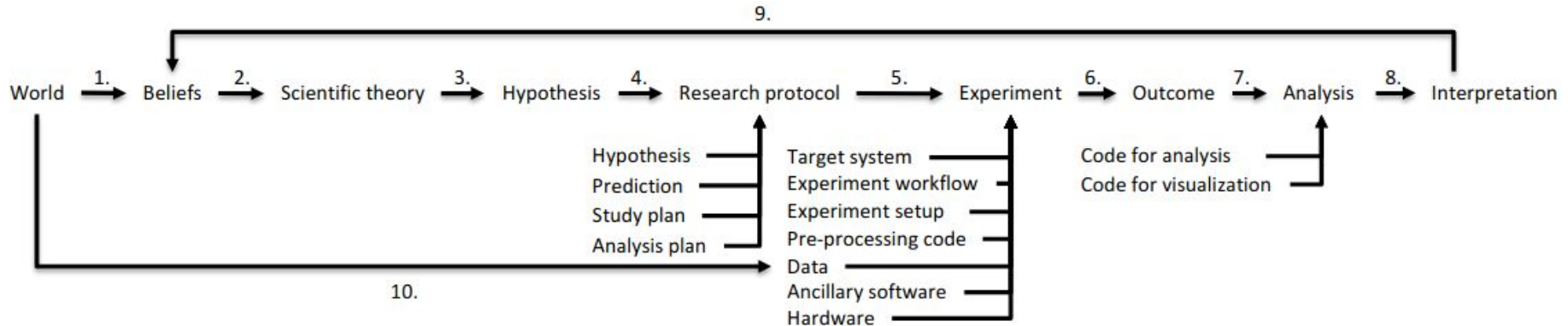


Figure 1. The scientific method as a ten step process: 1) *observe the world to form beliefs about it*; 2) *explain causes and effects by forming a scientific theory*; 3) *formulate a genuine test of the theory*; 4) *design an experiment to test the theory*; 5) *implement the experiment*; 6) *conduct the experiment*; 7) *analyse the outcome*; 8) *interpret the analysis*; 9) *update beliefs according to the result*; and 10) *observe the world systematically*.

Source: O. E. Gundersen (2020; <https://doi.org/10.1098/rsta.2020.0210>)

Reproducibility

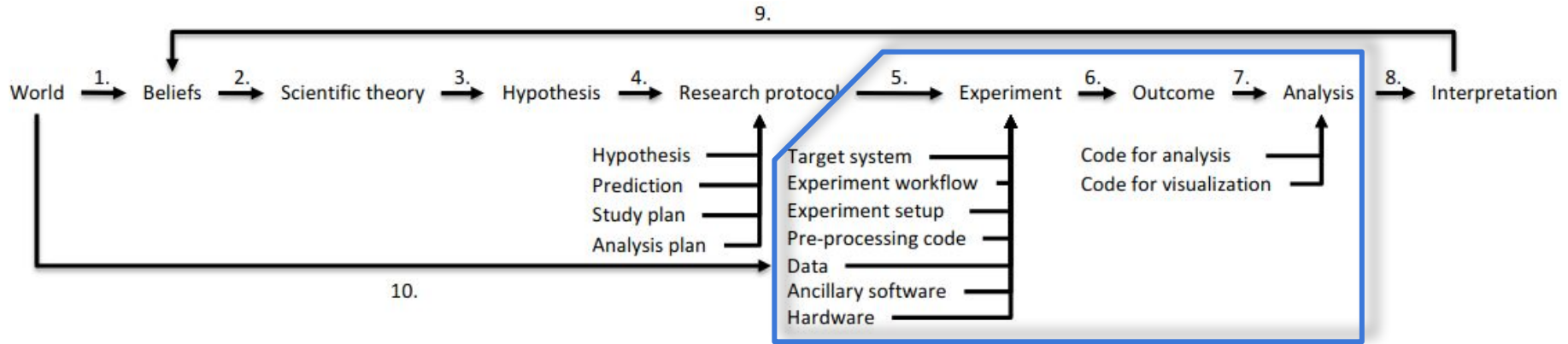


Figure 1. The scientific method as a ten step process: 1) *observe the world to form beliefs about it*; 2) *explain causes and effects by forming a scientific theory*; 3) *formulate a genuine test of the theory*; 4) *design an experiment to test the theory*; 5) *implement the experiment*; 6) *conduct the experiment*; 7) *analyse the outcome*; 8) *interpret the analysis*; 9) *update beliefs according to the result*; and 10) *observe the world systematically*.

Source: O. E. Gundersen (2020; <https://doi.org/10.1098/rsta.2020.0210>)

Reproducible environment

- Software (and libraries!) version

#	Name	Version		Version
	bash	5.1.16	libgfortran5	11.2.0
	bash_kernel	0.7.2	liblapack	3.9.0
	bzip2	1.0.8	libnetcdf	4.8.1
	cartopy	0.20.2	libpng	1.6.37
	cdo	1.9.10	libtiff	4.3.0
	cdsapi	0.5.1	libzlib	1.2.11
	cftime	1.6.0	matplotlib-base	3.5.2
	curl	7.83.0	mpich	4.0.2
	dask	2022.5.2	myproxyclient	2.1.0
	eccodes	2.25.0	mysql-libs	8.0.29
	esgf-pyclient	0.3.1	nco	5.0.6
	esmf	8.2.0	netcdf-fortran	4.5.4
	esmpy	8.2.0	netcdf4	1.5.8
	gsl	2.7	notebook	6.4.2
	hdf5	1.12.1	numpy	1.22.3
	ipython	8.3.0	oauthlib	3.2.0
	jasper	2.0.33	openssl	1.1.10
	jpeg	9e	pandas	1.2.4
	json5	0.9.5	proj	8.2.0

Reproducible environment

- Software (and libraries!) version
- Language-specific tools (pip, CRAN, ...)
- Multi-language environment management (conda)
- Full virtualization including OS (VMware, VirtualBox)
- OS-level virtualization (docker containers)
- Container orchestration and scaling (kubernetes)
- the Cloud ...

Conda (<https://conda.io>)



- Cross-platform package and environment manager
- Manages package versions and dependencies
- Isolates execution environments with different versions
- Available in different flavours
 - Anaconda, miniconda, mamba
- Manages Python, R, ... and many well known tools:
 - cdo, nco, ncview, ...
- No administrator rights required

Conda (<https://conda.io>)



- Cross-platform package and environment manager
- Manages package versions and dependencies
- Isolates execution environments with different versions
- Available in different flavours

Anaconda, miniconda, mamba

```
$ conda create -n myenv
$ conda activate myenv
$ conda install -c conda-forge cdo=1.9.8 nco ncview
$ conda install -c conda-forge esgf-pyclient
```

Jupyter notebook and lab (jupyter.org)

- Web-based interactive development environment for code notebooks and more.
- Notebooks integrate formatted text, formulas, code and code output, including plots.
- Support for over 40 programming languages



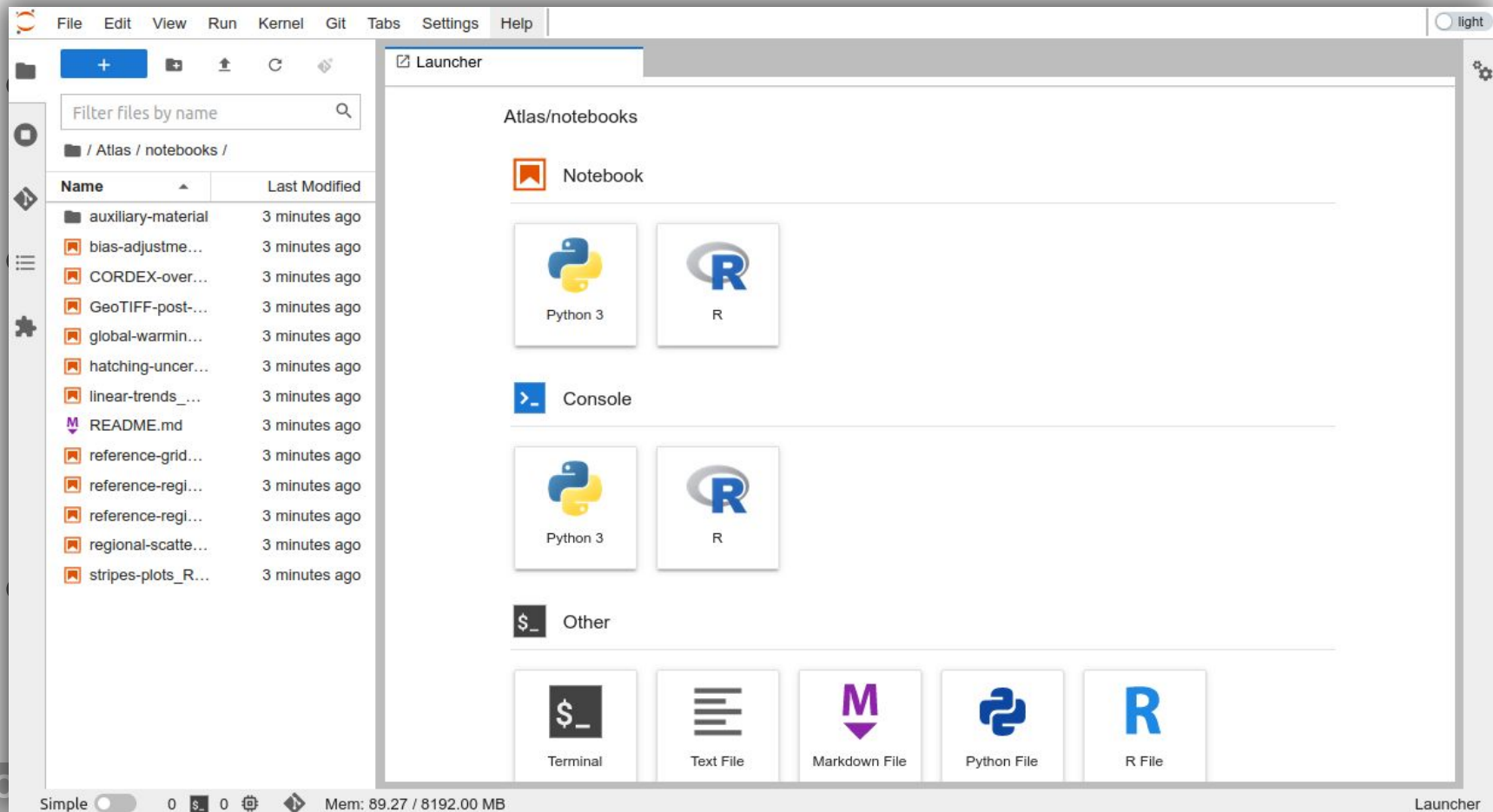
Jupyter notebook and lab (jupyter.org)

- Web-based interactive development environment for code notebooks and more.
- Notebooks integrate formatted text, formulas, code and code output, including plots.



```
$ conda install -c conda-forge jupyter jupyterlab  
$ jupyter lab
```

Jupyter notebook and lab (jupyter.org)



Jupyter notebook and lab (jupyter.org)

The screenshot displays the Jupyter Notebook interface. On the left, a file browser shows a directory structure under "/ Atlas / notebooks /". The file "stripes-plots_R..." is selected. The main area shows a code cell with the following text:

Stripes of global mean values for CMIP6

The next call of `computeStripes` produces a stripes figure for CMIP6 (`project = "CMIP6"`) annual (`season = 1:12`) mean precipitation (`var = "pr"`). The historical and ssp585 scenarios are considered (`experiment = "ssp585"`) for the global land surface (`region = "world"` and `area = "land"`).

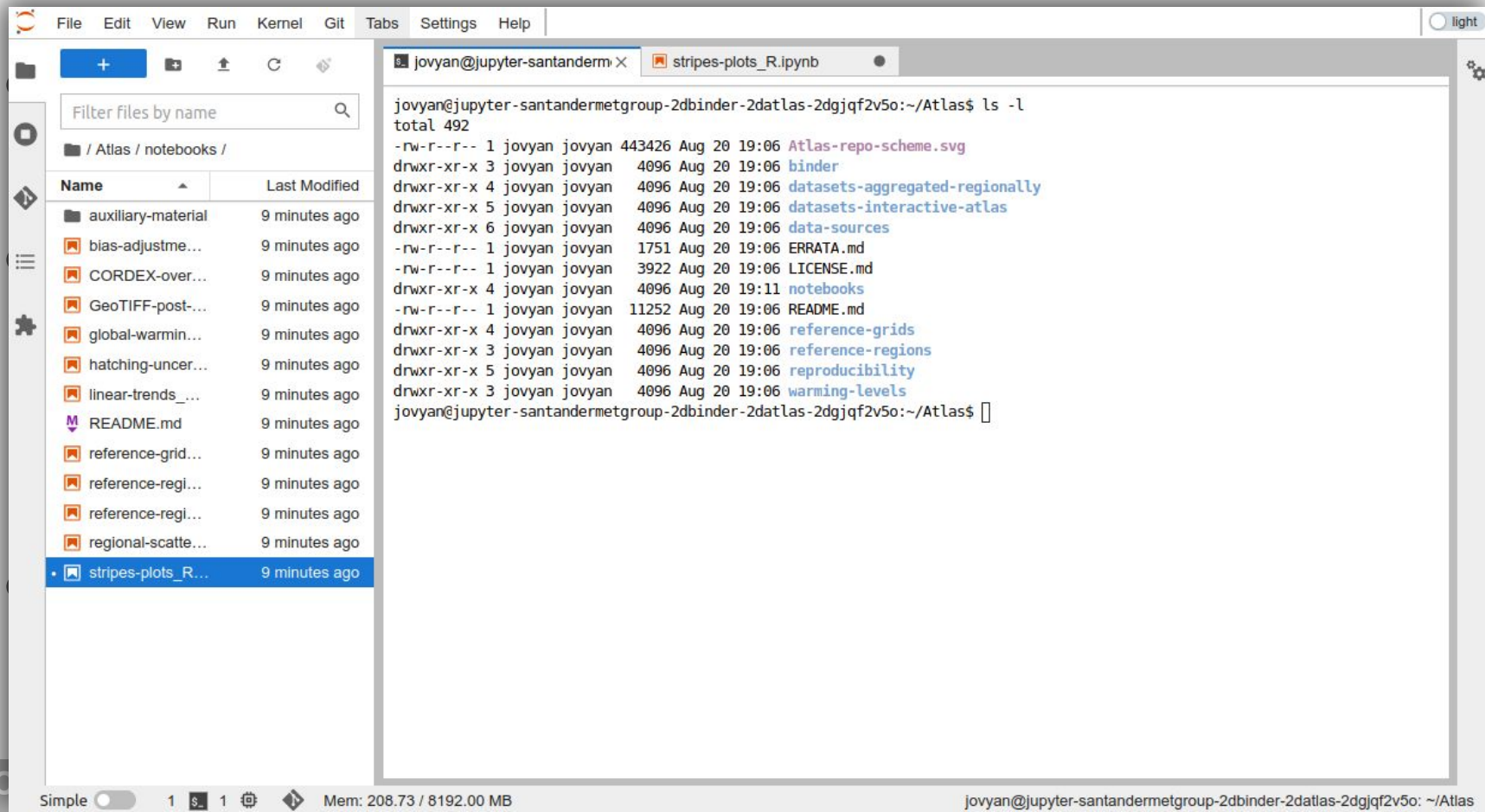
```
[3]: computeStripes(project = "CMIP6",
                    var = "pr",
                    experiment = "ssp585",
                    season = 1:12,
                    area = "land",
                    region = "world",
                    cordex.domain = NULL,
                    brewer.pal.name = "Blues",
                    rev.colors = TRUE)
```

proj:CMIP6 var:pr exp:ssp585 season:Annual area:land region:world

1st C

Simple 0 s 1 R | Idle Mem: 169.73 / 8192.00 MB Mode: Command Ln 1, Col 1 stripes-plots_R.ipynb (tina)

Jupyter notebook and lab (jupyter.org)



The screenshot shows the JupyterLab interface. On the left is a file browser for the directory `/ Atlas / notebooks /`. It lists various files and folders, including `auxiliary-material`, `bias-adjustme...`, `CORDEX-over...`, `GeoTIFF-post...`, `global-warmin...`, `hatching-uncer...`, `linear-trends_...`, `README.md`, `reference-grid...`, `reference-regl...`, `reference-regl...`, `regional-scatte...`, and `stripes-plots_R...`. The `stripes-plots_R...` file is selected.

The main area shows a terminal window with the following output:

```
jovyan@jupyter-santandermetgroup-2dbinder-2datlas-2dgjqf2v5o:~/Atlas$ ls -l
total 492
-rw-r--r-- 1 jovyan jovyan 443426 Aug 20 19:06 Atlas-repo-scheme.svg
drwxr-xr-x 3 jovyan jovyan 4096 Aug 20 19:06 binder
drwxr-xr-x 4 jovyan jovyan 4096 Aug 20 19:06 datasets-aggregated-regionally
drwxr-xr-x 5 jovyan jovyan 4096 Aug 20 19:06 datasets-interactive-atlas
drwxr-xr-x 6 jovyan jovyan 4096 Aug 20 19:06 data-sources
-rw-r--r-- 1 jovyan jovyan 1751 Aug 20 19:06 ERRATA.md
-rw-r--r-- 1 jovyan jovyan 3922 Aug 20 19:06 LICENSE.md
drwxr-xr-x 4 jovyan jovyan 4096 Aug 20 19:11 notebooks
-rw-r--r-- 1 jovyan jovyan 11252 Aug 20 19:06 README.md
drwxr-xr-x 4 jovyan jovyan 4096 Aug 20 19:06 reference-grids
drwxr-xr-x 3 jovyan jovyan 4096 Aug 20 19:06 reference-regions
drwxr-xr-x 5 jovyan jovyan 4096 Aug 20 19:06 reproducibility
drwxr-xr-x 3 jovyan jovyan 4096 Aug 20 19:06 warming-levels
jovyan@jupyter-santandermetgroup-2dbinder-2datlas-2dgjqf2v5o:~/Atlas$
```

The bottom status bar shows the system tray with icons for Simple, 1, 1, and Mem: 208.73 / 8192.00 MB. The user is `jovyan@jupyter-santandermetgroup-2dbinder-2datlas-2dgjqf2v5o: ~/Atlas`.

Jupyter Rstudio UI (jupyter.org)

The screenshot displays a web-based RStudio interface. At the top, a browser window shows the URL `https://hub-binder.mybinder.ovh/user/santandermetgro-binder-example-wr`. Below the browser, the RStudio application is visible with a menu bar (File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help) and a toolbar. The main workspace is divided into several panes:

- Source Editor:** Contains a script named `test.R` with the following R code:

```
1 curve(sin, -pi, pi, col="blue")  
2
```
- Console:** Shows the execution of the R code:

```
R 4.1.3 ~/  
> curve(sin, -pi, pi, col="blue")  
>
```
- Environment:** Displays "Environment is empty".
- Plots:** Shows a plot of a sine wave, `sin(x)`, with the x-axis ranging from -3 to 3 and the y-axis from -1.0 to 1.0. The curve is blue.

JupyterHub


- Multi-user version of Jupyter
- Centralized deployment
(no installation by the user)
- Can be deployed next to the data




kubernetes

MyBinder (mybinder.org)



- MyBinder is a cloud service providing an interactive computing environment in your browser 
- It uses binder to create an image of your environment from a variety of specification files:

```
requirements.txt (pip)      environment.yml (conda)
Install.R              (R)      postBuild
```

- It can load the contents of a code repository (GitHub, Zenodo, ...) 
- Changes do NOT persist across sessions

The Cloud

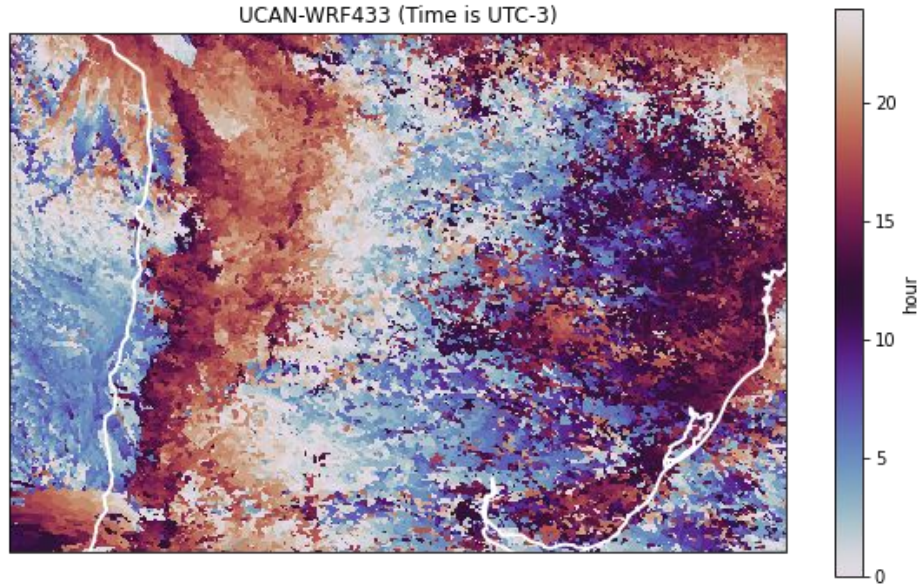
Pangeo (enable Big Data geoscience research)

<https://pangeo.io/cloud.html> <https://gallery.pangeo.io>

NA-CORDEX data on the Amazon cloud:

<https://github.com/NCAR/na-cordex-aws>

Practical exercise



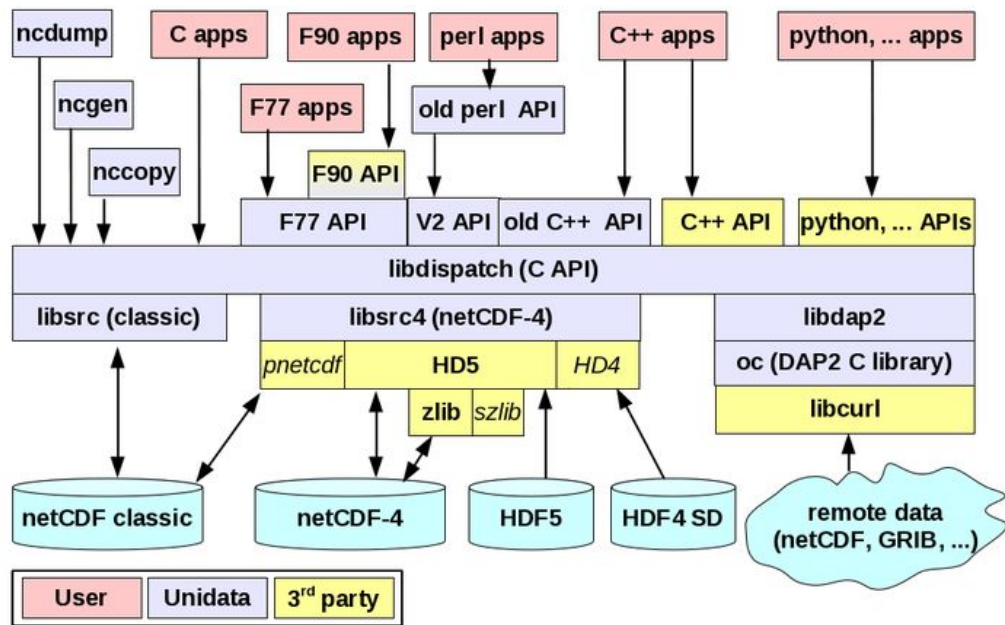
OPeNDAP test: https://dapds00.nci.org.au/thredds/...tas_Amon_ACCESS-CM2_historical_r1i1p1f1_gn_185001-201412.nc.das

Suppl. material



NetCDF: storage formats

Software libraries and machine-independent data ~~format~~ **model** for array-oriented scientific data.



- **NetCDF-Java** is an independent implementation, not shown here
- C-based 3rd-party netCDF APIs for other languages include Python, Ruby, Perl, Fortran-2003, MATLAB, IDL, and R
- 3rd party libraries are optional (HDF5, HDF4, zlib, szlib, PnetCDF, libcurl), depending on what features are needed and how netCDF is configured

[https://docs.unidata.ucar.edu/ ... netcdf_format](https://docs.unidata.ucar.edu/...netcdf_format)

R tools

<https://arrow.apache.org/docs/r>

<https://github.com/tidyverse/multidplyr>

<https://cran.r-project.org/web/packages/future/index.html>

<https://cran.r-project.org/web/views/HighPerformanceComputing.html> (see Large memory and out-of-memory data)

“One possibility to overcome the output avalanche is to merely store the simulation setup, initial conditions and restart files, and rerun the simulation on demand when needed to perform a specific analysis. A more sophisticated scheme would restart the simulation in parallel from a series of restart files. This, in principle, enables us to arbitrarily trade off storage for computation.”

Kilometer-Scale Climate Models

Prospects and Challenges

Christoph Schär, Oliver Fuhrer, Andrea Arteaga, Nikolina Ban, Christophe Charpilloz, Salvatore Di Girolamo, Laureline Hentgen, Torsten Hoefler, Xavier Lapillonne, David Leutwyler, Katherine Osterried, Davide Panerai, Stefan Rüdiger, Linda Schlemmer, Thomas C. Schulze

BAMS
Article

MARCH 2020

Remote access

OPeNDAP (opendap.org)

Open-source Project for a Network Data Access Protocol (DAP)

DAP2 is a discipline-neutral means of requesting and providing data across the World Wide Web (HTTP).

The NetCDF-C library has a built-in DAP2 client

Drawbacks:

- Slow for large requests (it is not magic, it's remote)

```
$ curl https://remotetest.unidata.ucar.edu/dts/test.01.dds
```

```
Dataset {  
  Byte b;  
  Int32 i32;  
  UInt32 ui32;  
  Int16 i16;  
  UInt16 ui16;  
  Float32 f32;  
  Float64 f64;  
  String s;  
  Url u;
```

Remote access

OPeNDAP (opendap.org)

Open-source Project for a Network Data Access Protocol (DAP)

DAP2 is a discipline-neutral means of requesting and providing data across the World Wide Web (HTTP).

The NetCDF-C library has a built-in DAP2 client

Drawbacks:

- Slow for large requests (it is not magic, it's remote)

```
$ curl https://remotetest.unidata.ucar.edu/dts/test.01.das
```

```
Attributes {  
  Facility {  
    String PrincipleInvestigator "Mark Abbott", "Ph.D";  
    String DataCenter "COAS Environmental Computer Facility";  
    String DrifterType "MetOcean WOCE/OCM";  
  }  
  b {  
    String Description "A test byte";  
    String units "unknown";  
  }  
  i32 {  
    String Description "A 32 bit test server int";  
    String units "unknown";  
  }  
}
```

Data access: transfer speed limiting factors

Bandwidth



Latency



R/W speed

