

Efficient, robust and timely analysis of Earth System Models

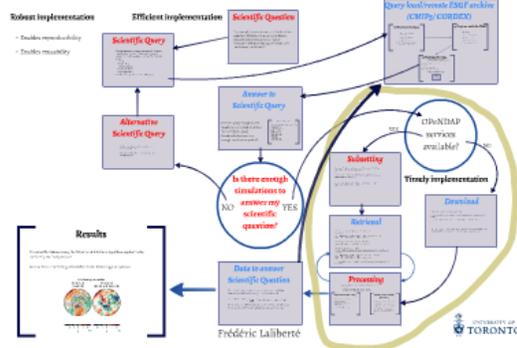
Lessons from four years of experience using advanced features of the ESGF

Frédéric Laliberté
February 25, 2015, GO-ESSP

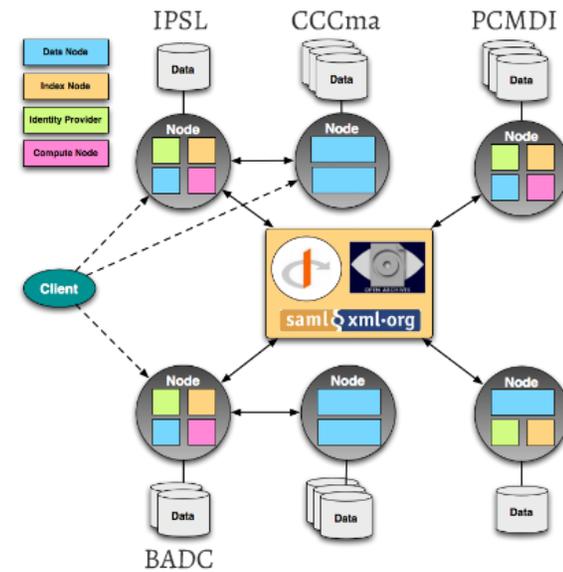


M. Juckes (BADC), S. Derrill (IPSL), Paul J. Kushner (U of T)

Climate Diagnostic Benchmark (CDB) Workflow

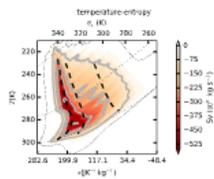


ESGF Federation Design



More Results

- Want to allow even more complex climate diagnostics.
- For example, Laliberté et al. (2015, Science) which uses six-hourly data.



Summary

- Need to be able to "ask" using complex queries.
- Need to "validate" availability of discovered data.
- Require a format to share this information for reproducibility and efficient collaboration.
- Ideally, this format would contain all the meta-data from the netCDF header.
- Need a retrieval tool that can use the collected information.
- Ideally, the retrieval tool would seamlessly perform simple remote operations (subsetting).

Outlook

- If there was a minimum "standard" to pass the information, some of these features could be provided by a WPS.
- A "standard" format could include checksum information at the array level, improving reproducibility even more. For example, have the checksum for every time slice.
- More permanent URLs would greatly improve the stability of this approach.

Efficient, robust and timely analysis of Earth System Models

Lessons from four years of experience using advanced
features of the ESGF

Frédéric Laliberté

February 25, 2015, GO-ESSP



UNIVERSITY OF
TORONTO

M. Jukes (BADC), S. Denvil (IPSL), Paul J. Kushner (U of T)

Scientific Question

"Are atmospheric motions associated with cyclonic systems in NH winter more energetic (kinetic, potential and/or thermal) in 21st century simulations with greenhouse gas forcing than in 20th century simulations? "

Scientific Query

- Translate question into a data requirement. Can I compare enough simulations for DJF of 1981-1999 (experiment historical) and 2081-2099 (experiment rcp85) that have six-hourly:
 - winds
 - temperature
 - specific humidity
 - sea-level pressure
 - surface pressure?
- Translate data requirement into an ESGF query.

Query local/remote ESGF archive (CMIP5/ CORDEX)

1) Discover the data

a. Tell the CDB where to look for data

List index nodes

b. Discover what is broadcast

Use the "ask" command

```
$ cdb_query_CMIP5 ask cyclone_DJF_hdr cyclone_DJF_pointers.nc
```

Returns a hierarchical netCDF4 file with pointers to the data and the associated metadata.

2) Register with the ESGF

a. Obtain an OpenID

Online and register for datasets

b. Obtain your certificates

Use BADC WPS myproxy service:

```
$ cdb_query_CMIP5 certificates flaliberte badc
```

3) Validate the data

Use the "validate" command

```
$ cdb_query_CMIP5 validate cyclone_DJF_pointers.nc cyclone_DJF_pointers.validate.nc
```

This command queries the ESGF nodes and the local file system to ensure that the discovered data is:

- available
- contains all the months and years requested

It can take up to several hours to complete!
To speed-up the process, you can query using more processes.

Returns a hierarchical netCDF4 file with pointers to the data and the associated metadata.

Down data nodes

```
#!/bin/sh
# This script is used to download the data
# from the ESGF nodes to the local file system
# It is a wrapper around the cdb_query_CMIP5
# command.
# The user should provide the path to the
# netCDF4 file in the first argument.
# Usage: ./download_data.sh <netCDF4 file>
```

1) Discover the data

a. Tell the CDB where to look for data

List index nodes

b. Discover what is broadcast

Use the "ask" command

```
$ cdb_query_CMIP5 ask cyclone_DJF.hdr cyclone_DJF_pointers.nc
```

Returns a hierarchical netCDF4 file with pointers to the data and the associated metadata.

2) Register with the ESGF

a. Obtain an OpenID

Online and register for datasets

b. Obtain your certificates

Use BADC WPS myproxy service:

```
$ cdb_query_CMIP5 certificates flaliberte badc
```


Down data nodes

The most common error message with "validate" is the following:

The url http://cmip3.dkrz.de/thredds/fileServer/cmip5/output1/BCC/bcc-csm1-1/historical/6hr/atmos/6hrPlev/r1i1p1/v1/psl/psl_6hrPlev_bcc-csm1-1_historical_r1i1p1_198001010000-201212311800.nc could not be opened. Copy and paste this url in a browser and try downloading the file. If it works, you can stop the download and retry using `cdb_query`. If it still does not work it is likely that your certificates are either not available or out of date.

Usually it appears when users fail to renew their certificates.

Sometimes, it appears when a data node is down. The data node is the first part of the URL: <http://cmip3.dkrz.de>

This error can be circumvented by excluding the offending data node from "validate":

```
$ cdb_query_CMIP5 validate --Xdata_node=http://cmip3.dkrz.de cyclone_DJF_pointers.nc  
cyclone_DJF_pointers.validate.nc
```

```

$ ncdump -h cyclone_DJF_pointers.validate.nc | more
netcdf cyclone_DJF_pointers.validate {

// global attributes:
 :data_node_list = "[\"http://esgf-data1.ceda.ac.uk\", \"http://albedo2.dkrz.de\", \"http://cmip3.dkrz.de\", \"http://vesg.ips
1.fr\", \"http://cmip-dn1.badc.rl.ac.uk\"]" ;
 :file_type_list = "[\"HTTPServer\", \"local_file\"]" ;
 :months_list = "[1, 2, 10, 11, 12]" ;
 :search_list = "[\"http://esgf-index1.ceda.ac.uk/esg-search/\", \"http://esgf-node.ips1.fr/esg-search/\", \"http://pcmdi9.lln
1.gov/esg-search/\", \"http://esgf-data.dkrz.de/esg-search/\", \"http://esgdata.gfdl.noaa.gov/esg-search/\", \"http://esg-datanode.jpl.nasa.g
ov/esg-search/\", \"/in/CMIP5\"]" ;
 :experiment_list = "{\"rcp45\": \"2081,2099\", \"historical\": \"1981,1999\", \"rcp85\": \"2081,2099\"}" ;
 :variable_list = "{\"va\": [\"6hr\", \"atmos\", \"6hrLev\"], \"ps1\": [\"6hr\", \"atmos\", \"6hrPlev\"], \"ps\": [\"6hr\", \"
atmos\", \"6hrLev\"], \"orog\": [\"fx\", \"atmos\", \"fx\"], \"hus\": [\"6hr\", \"atmos\", \"6hrLev\"], \"ua\": [\"6hr\", \"atmos\", \"6hrLev
\"], \"ta\": [\"6hr\", \"atmos\", \"6hrLev\"]}" ;

group: NOAA-GFDL {

// group attributes:
 :level_name = "institute" ;

group: GFDL-ESM2G {

// group attributes:
 :level_name = "model" ;

group: rcp45 {

// group attributes:
 :level_name = "experiment" ;

group: \6hr {

// group attributes:
 :level_name = "time_frequency" ;

group: atmos {

// group attributes:
 :level_name = "realm" ;

group: \6hrLev {

// group attributes:
 :level_name = "cmor_table" ;

group: r11p1 {

// group attributes:
 :level_name = "ensemble" ;

group: va {
 dimensions:
  time = UNLIMITED ; // (11476 currently)
  lev = 24 ;
  bnds = 2 ;
  lat = 90 ;
  lon = 144 ;
 variables:
  double time(time) ;
  time:calendar = "noleap" ;
  time:units = "days since 2006-01-01 00:00:00" ;
  double lev(lev) ;

```



```

lon = 144 ;
variables:
double time(time) ;
  time:calendar = "noleap" ;
  time:units = "days since 2006-01-01 00:00:00" ;
double lev(lev) ;
  lev:long_name = "hybrid sigma pressure coordinate" ;
  lev:units = "1" ;
  lev:positive = "down" ;
  lev:axis = "Z" ;
  lev:bounds = "lev_bnds" ;
  lev:formula = "p(n,k,j,i) = a(k)*p0 + b(k)*ps(n,j,i)" ;
  lev:formula_terms = "p0: p0 a: a b: b ps: ps" ;
  lev:standard_name = "atmosphere_hybrid_sigma_pressure_coordinate" ;
double bnds(bnds) ;
  bnds:long_name = "vertex number" ;
  bnds:cartesian_axis = "N" ;
double lev_bnds(lev, bnds) ;
  lev_bnds:formula = "p(n,k,j,i) = a(k)*p0 + b(k)*ps(n,j,i)" ;
  lev_bnds:formula_terms = "p0: p0 a: a_bnds b: b_bnds ps: ps" ;
  lev_bnds:standard_name = "atmosphere_hybrid_sigma_pressure_coordinate" ;
  lev_bnds:units = "1" ;
double lat(lat) ;
  lat:long_name = "latitude" ;
  lat:units = "degrees_north" ;
  lat:standard_name = "latitude" ;
  lat:axis = "Y" ;
  lat:bounds = "lat_bnds" ;
double lat_bnds(lat, bnds) ;
double lon(lon) ;
  lon:long_name = "longitude" ;
  lon:units = "degrees_east" ;
  lon:standard_name = "longitude" ;
  lon:axis = "X" ;
  lon:bounds = "lon_bnds" ;
double lon_bnds(lon, bnds) ;
float va(time, lev, lat, lon) ;
  va:_FillValue = 1.e+20f ;
  va:cell_methods = "time: point" ;
  va:long_name = "Northward Wind" ;
  va:missing_value = 1.e+20f ;
  va:original_name = "vcomp" ;
  va:units = "m s-1" ;
  va:valid_range = -330.f, 350.f ;
  va:standard_name = "northward_wind" ;
  va:original_units = "m/sec" ;
  va:cell_measures = "area: areacella" ;
  va:associated_files = "baseURL: http://cmip-pcmdi.llnl.gov/CMIP5/dataLocation areacella: areacella_fx_GFDL-ES
M2G_rcp45_r0i0p0.nc" ;
double a(lev) ;
  a:bounds = "a_bnds" ;
  a:long_name = "vertical coordinate formula term: a(k)" ;
double b(lev) ;
  b:bounds = "b_bnds" ;
  b:long_name = "vertical coordinate formula term: b(k)" ;
float p0 ;
  p0:long_name = "reference pressure for hybrid sigma coordinate" ;
  p0:units = "Pa" ;
double a_bnds(lev, bnds) ;
  a_bnds:long_name = "vertical coordinate formula term: a(k+1/2)" ;
double b_bnds(lev, bnds) ;
  b_bnds:long_name = "vertical coordinate formula term: b(k+1/2)" ;
// group attributes:

```



```
a_bnds:long_name = "vertical coordinate formula term: a(k+1/2)";
double b_bnds(lev, bnds);
b_bnds:long_name = "vertical coordinate formula term: b(k+1/2)";
```

```
// group attributes:
```

```
:level_name = "var" ;
:title = "NOAA GFDL GFDL-ESM2G, RCP4.5 (run 1) experiment output for CMIP5 AR5" ;
:institute_id = "NOAA GFDL" ;
:source = "GFDL-ESM2G 2010 ocean: TOPAZ (TOPAZ1p2,Tripolar360x210L63); atmosphere: AM2 (AM2p14,M45L24); sea i
```

```
ce: SIS (SISp2,Tripolar360x210L63); land: LM3 (LM3p7_cESM,M45)" ;
```

```
:contact = "gfdl.climate.model.info@noaa.gov" ;
```

```
:project_id = "CMIP5" ;
```

```
:table_id = "Table 6hrLev (31 Jan 2011)" ;
```

```
:experiment_id = "rcp45" ;
```

```
:realization = 1 ;
```

```
:modeling_realm = "atmos" ;
```

```
:tracking_id = "71f7ac73-fa4a-4593-81e9-5dc0a1440aca" ;
```

```
:Conventions = "CF-1.4" ;
```

```
:references = "The GFDL Data Portal (http://nomads.gfdl.noaa.gov/) provides access to NOAA/GFDL's publicly a
```

```
vailable model input and output data sets. From this web site one can view and download data sets and documentation, including those related to the GFDL coupled models experiments run for the IPCC's 5th Assessment Report and the US CCSP." ;
```

```
:comment = "GFDL experiment name = ESM2G-HC2_2006-2100_all_rcp45_XC2. PCMDI experiment name = RCP4.5 (run1).
```

Initial conditions for this experiment were taken from 1 January 2006 of the parent experiment, ESM2G-C2_all_historical_HC2 (historical). Several forcing agents varied during the 95 year duration of the RCP4.5 experiment based upon the MINICAM integrated assessment model for the 21st century. The time-varying forcing agents include the well-mixed greenhouse gases (CO2, CH4, N2O, halons), tropospheric and stratospheric O3, model-derived aerosol concentrations (sulfate, black and organic carbon, sea salt and dust), and land use transitions. Volcanic aerosols were zero and solar irradiance varied seasonally based upon late 20th century averages but with no interannual variation. The direct effect of tropospheric aerosols is calculated by the model, but not the indirect effects."

```
:gfdl_experiment_name = "ESM2G-HC2_2006-2100_all_rcp45_XC2" ;
```

```
:creation_date = "2012-01-11T16:02:46Z" ;
```

```
:model_id = "GFDL-ESM2G" ;
```

```
:branch_time = "52925" ;
```

```
:experiment = "RCP4.5" ;
```

```
:forcing = "GHG,SD,Oz,LU,SS,BC,MD,OC (GHG includes CO2, CH4, N2O, CFC11, CFC12, HCFC22, CFC113)" ;
```

```
:frequency = "6hr" ;
```

```
:initialization_method = 1 ;
```

```
:parent_experiment_id = "historical" ;
```

```
:physics_version = 1 ;
```

```
:product = "output1" ;
```

```
:institution = "NOAA GFDL(201 Forrestal Rd, Princeton, NJ, 08540)" ;
```

```
:history = "File was processed by fremetar (GFDL analog of CMOR). TripleID: [exper_id_CjJqLuIBEN,realiz_id_Or
```

```
NaY4lwmd,run_id_1EJwv7KIgx]" ;
```

```
:parent_experiment_rip = "r1i1p1" ;
```

```
:DODS_EXTRA.Unlimited_Dimension = "time" ;
```

```
:netcdf_soft_links_version = "1.0" ;
```

```
group: soft_links {
```

```
  dimensions:
```

```
    path = 5 ;
```

```
    indices = 2 ;
```

```
  variables:
```

```
    string path(path) ;
```

```
    uint version(path) ;
```

```
    uint path_id(path) ;
```

```
    string data_node(path) ;
```

```
    string file_type(path) ;
```

```
    string checksum(path) ;
```

```
    string indices(indices) ;
```

```
    uint va(time, indices) ;
```

```
    va:_FillValue = 4294967295U ;
```

```
  } // group soft_links
```

```
} // group va
```



Prezi

Answer to Scientific Query

- Once the query completes, list models and ensemble members that satisfy the query.
- Decide whether there are enough simulations available.

Use the `list_fields` command

```
$ cdb_query_CMIP5 list_fields -f institute -f model -f ensemble  
--Xensemble=r0i0p0 cyclone_DJF_pointers.validate.nc  
BCC,BCC-CSM1.1,r1i1p1  
CCCMA,CanESM2,r1i1p1  
CSIRO-BOM,ACCESS1.0,r1i1p1  
CSIRO-BOM,ACCESS1.3,r1i1p1  
CSIRO-QCCCE,CSIRO-Mk3.6.0,r1i1p1  
IPSL,IPSL-CM5A-LR,r1i1p1  
IPSL,IPSL-CM5A-LR,r2i1p1  
IPSL,IPSL-CM5A-LR,r3i1p1  
IPSL,IPSL-CM5A-LR,r4i1p1  
IPSL,IPSL-CM5A-MR,r1i1p1  
LASG-CESS,FGOALS-g2,r1i1p1  
MIROC,MIROC-ESM,r1i1p1  
MIROC,MIROC-ESM-CHEM,r1i1p1  
MIROC,MIROC5,r1i1p1  
MOHC,HadGEM2-ES,r2i1p1  
MRI,MRI-CGCM3,r1i1p1  
NCAR,CCSM4,r6i1p1  
NOAA-GFDL,GFDL-CM3,r1i1p1  
NOAA-GFDL,GFDL-ESM2G,r1i1p1  
NOAA-GFDL,GFDL-ESM2M,r1i1p1
```

Use the list_fields command

```
$ cdb_query_CMIP5 list_fields -f institute -f model -f ensemble
--Xensemble=r0i0p0 cyclone_DJF_pointers.validate.nc
BCC,BCC-CSM1.1,r1i1p1
CCCMA,CanESM2,r1i1p1
CSIRO-BOM,ACCESS1.0,r1i1p1
CSIRO-BOM,ACCESS1.3,r1i1p1
CSIRO-QCCCE,CSIRO-Mk3.6.0,r1i1p1
IPSL,IPSL-CM5A-LR,r1i1p1
IPSL,IPSL-CM5A-LR,r2i1p1
IPSL,IPSL-CM5A-LR,r3i1p1
IPSL,IPSL-CM5A-LR,r4i1p1
IPSL,IPSL-CM5A-MR,r1i1p1
LASG-CESS,FGOALS-g2,r1i1p1
MIROC,MIROC-ESM,r1i1p1
MIROC,MIROC-ESM-CHEM,r1i1p1
MIROC,MIROC5,r1i1p1
MOHC,HadGEM2-ES,r2i1p1
MRI,MRI-CGCM3,r1i1p1
NCAR,CCSM4,r6i1p1
NOAA-GFDL,GFDL-CM3,r1i1p1
NOAA-GFDL,GFDL-ESM2G,r1i1p1
NOAA-GFDL,GFDL-ESM2M,r1i1p1
```


Alternative Scientific Query

Could my scientific question be answered with a different scientific query?

How many simulations have the same variables but at the daily frequency?

Efficient implementation

Scientific Question

"Are atmospheric motions associated with cyclonic systems in NH winter more energetic (kinetic, potential and/or thermal) in 21st century simulations with greenhouse gas forcing than in 20th century simulations?"

Scientific Query

- Translate question into a data requirement. Can I compare enough simulations for DJF of 1961-999 (experiment historical) and 2020-3000 (experiment rcp85) that have six-hourly
 - winds
 - temperature
 - specific humidity
 - sea-level pressure
 - surface pressure?
- Translate data requirement into an ESGF query.

Alternative Scientific Query

Could my scientific question be answered with a different scientific query?
Have any simulations have the same variables but at the daily frequency?

Answer to Scientific Query

- Once the query completes, list models and ensemble members that satisfy the query.
- Decide whether there are enough simulations available.

Is there enough simulations to answer my scientific question?

NO YES

Data to answer Scientific Question

The retrieved/processed data can then be converted to a CMIP5/ CORDEX DRS local archive using the 'convert' command.

```

$ cd /opt/ESGF/...
$ convert --input-dir /opt/ESGF/... --output-dir /opt/ESGF/...

```

The directory out/CMIP5 can then be used in another scientific query. There the query will find six-hourly column-integrated energies.

(CMIP5/ CORDEX)



OPeNDAP services available?

YES

NO

Timely implementation

Subsetting

The result of 'validate' can be incorporated for any netCDF file. In particular, it is safe to subset dimensions.

Here, we want only the Northern Hemisphere extratropical, so we select latitudes between 30N and 60N:

```

$ echo -e 'lat:20,50,90 & longitude:0,360 & pressure:1000,100000 & time:0,1000000' > file.nc

```

Retrieval

We can now retrieve the data from the archive using 'download':

```

$ cd /opt/ESGF/...
$ curl -O http://esgf-data.ornl.gov/.../file.nc

```

This command essentially converts the netCDF file to a binary format. The retrieved data is stored in the local archive.

Processing

The retrieved/processed data is processed using a complex procedure:

```

$ compute_energy.sh -i /opt/ESGF/... -o /opt/ESGF/...

```

Asynchronous processing

Simultaneous retrieval/processing

Download

The data discovered can be downloaded 'in situ':

```

$ cd /opt/ESGF/...
$ curl -O http://esgf-data.ornl.gov/.../file.nc

```

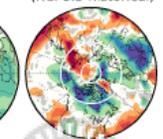
NOTE: It checks the checksum of the downloaded file against the checksum database of the local archive to ensure the data was downloaded using the intended

Results

and Schierz algorithm applied to the

from climatology) in cyclones:

energy anomalies response (RCP8.5-historical)



Download

The data discovered can be downloaded "as is":

```
$ cdb_query_CMIP5 download_raw cyclone_DJF_pointers.validate.nc out/  
CMIP5/
```

This command essentially creates a local copy of the files available in the archive.

NOTE: It checks the checksum of the downloaded files against the broadcast checksums. If the broadcast checksum is wrong, the data can't be downloaded using this method.

Subsetting

The result of "validate" can be manipulated like any netCDF4 files. In particular, it is safe to subset dimensions.

Here, we want only the Northern Hemisphere stormtracks, so we select latitudes between 20N and 90N:

```
$ ncks -d lat,20.0,90.0 cyclone_DJF_pointers.validate.nc  
cyclone_DJF_pointers.validate.NH.nc
```

Are there enough
calculations to
answer my
scientific
question?

YES

How to answer
scientific Question

Retrieved data can then be converted to a CMIP5/
archive using the "convert" command:

```
cyclone_DJF_pointers.validate.NH.retrieved.energies.nc \
  out/CMIP5/
```

CMIP5 can then be used in another scientific query.
and six-hourly column-integrated energies.

Subsetting

The result of "validate" can be manipulated like any netCDF4 files. In particular, it is safe to subset dimensions.

Here, we want only the Northern Hemisphere stormtracks, so we select latitudes between 20N and 90N:

```
$ ncks -d lat,20.0,90.0 cyclone_DJF_pointers.validate.nc
cyclone_DJF_pointers.validate.NH.nc
```

Retrieval

We can now retrieve the data from the archive using "download":

```
$ cdb_query CMIP5 download cyclone_DJF_pointers.validate.NH.nc
cyclone_DJF_pointers.validate.NH.retrieved.nc
```

This command essentially converts the soft links in cyclone_DJF_pointers.validate.NH.nc into "hard" data.

The subsetting is done by the server.

It can be split into several smaller queries by requesting the data year by year, for example:

```
$ cdb_query CMIP5 download --year=1981 --experiment=historical
cyclone_DJF_pointers.validate.NH.nc
cyclone_DJF_pointers.validate.NH.retrieved.1981.nc
```

The retrieval is done asynchronously. **The more the data is scattered among data nodes, the faster it is!**

Processing

- The six-hourly model-level data is processed using a complex procedure.
- ```
"compute_energies.sh in.nc out.nc"
```

### Asynchronous processing

The processing is ready for use for each node.

- Environmentally portable.**

The CRB includes a script manager that can keep scripts running through the transition to recovery after the processing is completed.

```
#!/bin/sh
set -e
...

```

### Simultaneous retrieval/processing

- The processing and the retrieval can be done simultaneously.

Retrieve your file.

Clear both sides to request and data is processed, copy the processing script for this and data respectively.

For an example, we can retrieve an other file that we could process, using the procedure. So here when using 12 iterations, the processing script will take a time longer than the retrieval script. With respect to time, **the retrieval is slower**.

## Timely implementation

### Download

The data discovered can be downloaded 'as is':

```
$ cdb_query CMIP5 download_raw cyclone_DJF_pointers.validate.nc out/
CMIP5/
```

This command essentially creates a local copy of the files available in the archive.

**NOTE:** It checks the checksum of the downloaded files against the broadcast checksums. If the broadcast checksum is wrong, the data can't be downloaded using this method.

# Retrieval

We can now retrieve the data from the archive using "download":

```
$ cdb_query_CMIP5 download cyclone_DJF_pointers.validate.NH.nc
cyclone_DJF_pointers.validate.NH.retrieved.nc
```

This command essentially converts the soft links in `cyclone_DJF_pointers.validate.NH.nc` into "hard" data.

The subsetting is done by the **server**.

It can be split into several smaller queries by requesting the data year by year, for example:

```
$ cdb_query_CMIP5 download --year=1981 --experiment=historical
cyclone_DJF_pointers.validate.NH.nc
cyclone_DJF_pointers.validate.NH.retrieved.1981.nc
```

The retrieval is done asynchronously. **The more the data is scattered among data nodes, the faster it is!**

Are there enough  
calculations to  
answer my  
scientific  
question?

YES

How to answer  
scientific Question

Retrieved data can then be converted to a CMIP5/  
archive using the "convert" command:

```
cyclone_DJF_pointers.validate.NH.retrieved.energies.nc \
 out/CMIP5/
```

CMIP5 can then be used in another scientific query.  
and six-hourly column-integrated energies.

## Subsetting

The result of "validate" can be manipulated like any netCDF4 files. In particular, it is safe to subset dimensions.

Here, we want only the Northern Hemisphere stormtracks, so we select latitudes between 20N and 90N:

```
$ ncks -d lat,20.0,90.0 cyclone_DJF_pointers.validate.nc
cyclone_DJF_pointers.validate.NH.nc
```

## Retrieval

We can now retrieve the data from the archive using "download":

```
$ cdb_query CMIP5 download cyclone_DJF_pointers.validate.NH.nc
cyclone_DJF_pointers.validate.NH.retrieved.nc
```

This command essentially converts the soft links in cyclone\_DJF\_pointers.validate.NH.nc into "hard" data.

The subsetting is done by the server.

It can be split into several smaller queries by requesting the data year by year, for example:

```
$ cdb_query CMIP5 download --year=1981 --experiment=historical
cyclone_DJF_pointers.validate.NH.nc
cyclone_DJF_pointers.validate.NH.retrieved.1981.nc
```

The retrieval is done asynchronously. **The more the data is scattered among data nodes, the faster it is!**

## Processing

- The six-hourly model-level data is processed using a complex procedure.
- ```
"compute_energies.sh in.nc out.nc"
```

Asynchronous processing

The processing is ready for use for each node.

- Environmentally portable.**

The CRB includes a script manager that can keep scripts running through the transition to recovery after the processing is completed.

```
#!/bin/sh
set -e
...

```

Simultaneous retrieval/processing

- The processing and the retrieval can be done simultaneously.

Retrieve your file.

Clear both sides to request and data is processed, copy the processing script for this and data respectively.

For an example, we can retrieve on other faster data we could process, using the procedure. So here when using 12 connections, the processing script will take a time longer than the retrieval script. With respect to time, **the retrieval is slower**.

Timely implementation

Download

The data discovered can be downloaded 'as is':

```
$ cdb_query CMIP5 download_raw cyclone_DJF_pointers.validate.nc out/
CMIP5/
```

This command essentially creates a local copy of the files available in the archive.

NOTE: It checks the checksum of the downloaded files against the broadcast checksums. If the broadcast checksum is wrong, the data can't be downloaded using this method.

Processing

- The six-hourly model-level data is processed using a complex procedure.

```
"compute_energies.sh in.nc out.nc"
```

Asynchronous processing

The processing is exactly the same for each simulation.

-> **Embarrassingly parallelizable.**

The CDB includes a simple manager than can loop asynchronously through the simulations to carry out the processing (using 10 processes):

```
$ cdb_query_CMIP5 apply --num_procs=10 \  
  'compute_energies.sh' \  
  cyclone_DJF_pointers.validate.NH.retrieved.nc \  
  cyclone_DJF_pointers.validate.NH.retrieved.energies.nc
```

Simultaneous retrieval/ processing

The processing and the retrieval can be done simultaneously:

- Retrieve year 1981.
- Start retrieving 1982 and simultaneously start processing 1981.
- Once both 1982 is retrieved and 1981 is processed, carry the previous step for 1983 and 1982, respectively.

For our example, we can retrieve 30 times faster than we could process, using one processor. So even when using 15 processors, the processing step still takes 2 times longer than the retrieval step. With respect to time, *the retrieval is therefore free.*

Asynchronous processing

The processing is exactly the same for each simulation.

-> **Embarrassingly parallelizable.**

The CDB includes a simple manager than can loop asynchronously through the simulations to carry out the processing (using 10 processes):

```
$ cdb_query_CMIP5 apply --num_procs=10 \  
    'compute_energies.sh' \  
    cyclone_DJF_pointers.validate.NH.retrieved.nc \  
    cyclone_DJF_pointers.validate.NH.retrieved.energies.nc
```

Simultaneous retrieval/ processing

The processing and the retrieval can be done simultaneously:

- Retrieve year 1981.
- Start retrieving 1982 and simultaneously start processing 1981.
- Once both 1982 is retrieved and 1981 is processed, carry the previous step for 1983 and 1982, respectively.

For our example, we can retrieve 30 times faster than we could process, using one processor. So even when using 15 processors, the processing step still takes 2 times longer than the retrieval step. With respect to time, *the retrieval is therefore free.*

Climate Diagnostic Benchmark (CDB) Workflow

Robust implementation

- Enables reproducibility
- Enables reusability

Efficient implementation

Scientific Question

'Are atmospheric motions associated with cyclonic systems in NH winter more energetic (kinetic, potential and/or thermal) in 21st century simulations with greenhouse gas forcing than in 20th century simulations?'

Scientific Query

- Translate question into a data requirement. Can I compare enough simulations for EOP of 100-1000 experiments historical and 20th-2099 experiment replicates that have circulation?
- metadata
- timeperiodic
- specific latitude
- sea level pressure
- surface pressure?
- Translate data requirement into an ESGF query.

Alternative Scientific Query

Could we answer the question by answering with a different scientific query?

How many simulations have the characteristics for an efficient implementation?

Answer to Scientific Query

- Once the query completes, list models and ensemble members that satisfy the query.
- Decide whether there are enough simulations available.

Is there enough simulations to answer my scientific question?

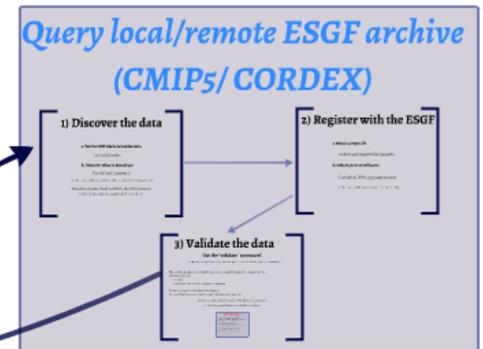
NO YES

Data to answer Scientific Question

The relevant processed data can then be converted to a CMIP5/ CORDEX data local archive using the 'convert' command.

• Use daily ODP output. Use daily ODP data for variables of interest available on 1-31-2010.

The directory use CMIP5 can then be used in another scientific query. There the query will find six hourly columns integrated energies.



Subsetting

The results of a subset can be uploaded to the any ESGF Plus. In particular, it is useful to subset dimensions.

There are many ways to subset dimensions. For example, we can subset latitude between 30N and 50N.

• Use the 'subset' command. The results of a subset can be uploaded to the any ESGF Plus. In particular, it is useful to subset dimensions.

Retrieval

Once we have the data from the archive using 'subset', we can retrieve it.

• Use the 'retrieve' command. The results of a subset can be uploaded to the any ESGF Plus. In particular, it is useful to subset dimensions.

Processing

The relevant processed data can then be converted to a CMIP5/ CORDEX data local archive using the 'convert' command.

• Use daily ODP output. Use daily ODP data for variables of interest available on 1-31-2010.

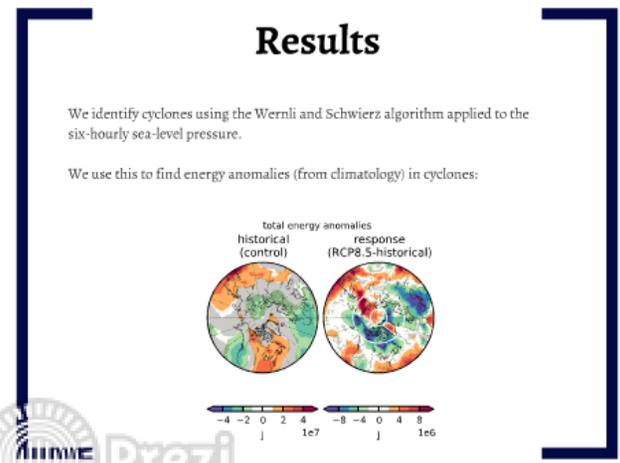
The directory use CMIP5 can then be used in another scientific query. There the query will find six hourly columns integrated energies.

Timely implementation

Download

The data downloaded can be downloaded to a local archive.

• Use the 'download' command. The results of a subset can be uploaded to the any ESGF Plus. In particular, it is useful to subset dimensions.



Frédéric Laliberté



Data to answer Scientific Question

The retrieved/processed data can then be converted to a CMIP5/
CORDEX DRS local archive using the "convert" command:

```
$ cdb_query_CMIP5 convert \  
    cyclone_DJF_pointers.validate.NH.retrieved.energies.nc \  
    out/CMIP5/
```

The directory out/CMIP5 can then be used in another scientific query.
There the query will find six-hourly column-integrated energies.

Climate Diagnostic Benchmark (CDB) Workflow

Robust implementation

- Enables reproducibility
- Enables reusability

Efficient implementation

Scientific Question

'Are atmospheric motions associated with cyclonic systems in NH winter more energetic (kinetic, potential and/or thermal) in 21st century simulations with greenhouse gas forcing than in 20th century simulations?'

Scientific Query

- Translate question into a data requirement. Can I compare enough simulations for EOP of 100-1000 experiments historical and 20th-2099 experiment replicates that have circularity?
 - winds
 - temperature
 - specific humidity
 - sea level pressure
 - surface pressure?
- Translate data requirement into an ESGF query.

Alternative Scientific Query

Could we answer the question by answering with a different scientific query?

How many simulations have the characteristics for an 'efficient' implementation?

Answer to Scientific Query

- Once the query completes, list models and ensemble members that satisfy the query.
- Decide whether there are enough simulations available.

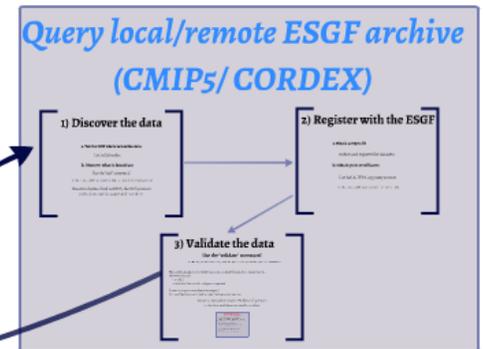
Is there enough simulations to answer my scientific question?

NO YES

Data to answer Scientific Question

The relevant processed data can then be converted to a CMIP5/ CORDEX data local archive using the 'convert' command.

• Use daily ODP output. Use daily ODP data for kinetic, potential and/or thermal energies. The directory use CMIP5 can then be used in another scientific query. There the query will find six hourly columns (integrated energies).



Subsetting

The results of a subset query are displayed in the query results page. In particular, it is useful to subset dimensions.

There are many ways to subset dimensions (temporal, spatial, etc.) and the syntax is described in the OPeNDAP manual.

• Use the 'subset' command to subset the data.

Retrieval

Once you have the data from the archive using 'subset', you can retrieve the data using 'retrieve'.

• Use the 'retrieve' command to retrieve the data.

The data is retrieved in a format that is compatible with the OPeNDAP client.

• Use the 'retrieve' command to retrieve the data.

Processing

The retrieved data can be processed using the 'process' command.

• Use the 'process' command to process the data.

The processed data can then be converted to a CMIP5/ CORDEX data local archive using the 'convert' command.

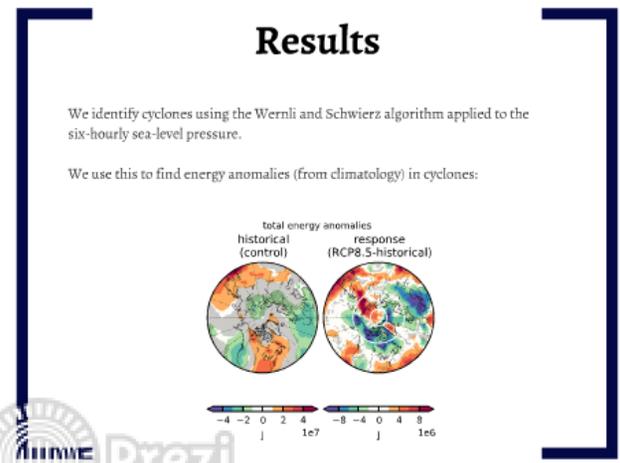
Timely implementation

Download

The data downloaded can be downloaded to a local archive.

• Use the 'download' command to download the data.

The downloaded data can then be converted to a CMIP5/ CORDEX data local archive using the 'convert' command.



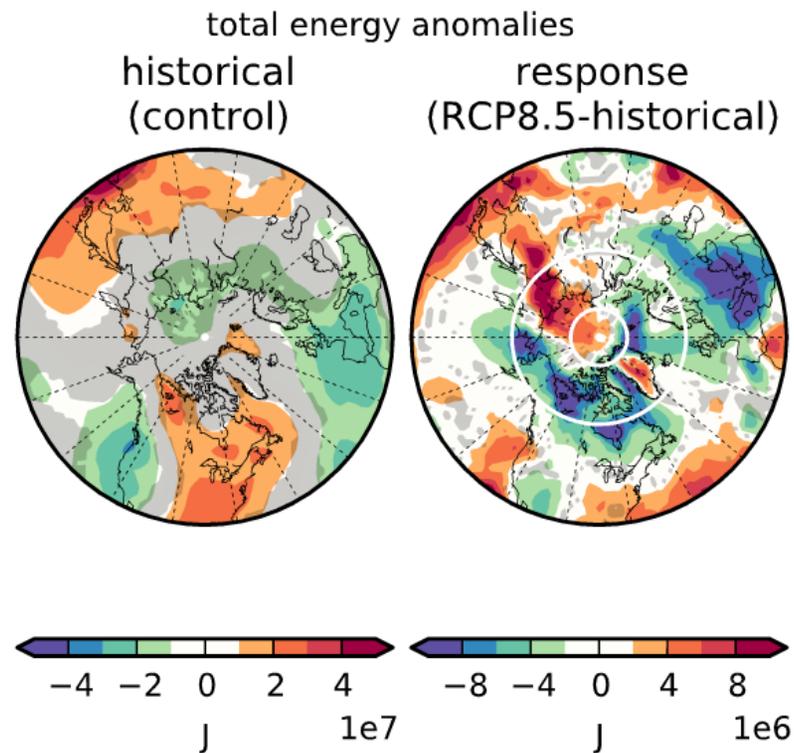
Frédéric Laliberté



Results

We identify cyclones using the Wernli and Schwierz algorithm applied to the six-hourly sea-level pressure.

We use this to find energy anomalies (from climatology) in cyclones:



Climate Diagnostic Benchmark (CDB) Workflow

Robust implementation

- Enables reproducibility
- Enables reusability

Efficient implementation

Scientific Question

'Are atmospheric motions associated with cyclonic systems in NH winter more energetic (kinetic, potential and/or thermal) in 21st century simulations with greenhouse gas forcing than in 20th century simulations?'

Scientific Query

- Translate question into a data requirement. Can I compare enough simulations for EOF of 100-1000 dependent historical and 20th-2099 dependent eqs that have circularity?
 - winds
 - temperature
 - specific humidity
 - sea level pressure
 - surface pressure?
- Translate data requirement into an ESGF query.

Alternative Scientific Query

Could we answer the question by answering with a different scientific query?

How many simulations have the characteristics for an 'efficient' implementation?

Answer to Scientific Query

- Once the query completes, list models and ensemble members that satisfy the query.
- Decide whether there are enough simulations available.

Is there enough simulations to answer my scientific question?

NO YES

Data to answer Scientific Question

The retrieved/processed data can then be converted to a CMIP5/ CORDEX CDB local archive using the 'convert' command.

• Use daily OPR output from the CDB archive to calculate monthly and 5-day OPR.

The directory use CMIP5 can then be used in another scientific query. There the query will find six hourly columns (integrated energies).

Query local/remote ESGF archive (CMIP5/ CORDEX)

- 1) Discover the data
- 2) Register with the ESGF
- 3) Validate the data

OPeNDAP services available?

YES NO

Subsetting

The results of a subset can be uploaded to the any ESGF Plus. In particular, it is useful to subset dimensions.

There are many ways to subset dimensions (temporal, spatial, or other) between local and sub.

• Use the 'subset' command to subset the data.

Retrieval

Once we have the data from the archive using 'retrieve'.

• Use query 'retrieve' to get the data. The data will be stored in the local archive.

The data is retrieved and converted to the local archive.

The data is retrieved and converted to the local archive.

Processing

The retrieved/processed data can then be converted to a CMIP5/ CORDEX CDB local archive using the 'convert' command.

• Use daily OPR output from the CDB archive to calculate monthly and 5-day OPR.

The directory use CMIP5 can then be used in another scientific query. There the query will find six hourly columns (integrated energies).

Timely implementation

Download

The data downloaded can be downloaded to a local archive.

• Use query 'download' to get the data. The data will be stored in the local archive.

The data is downloaded and converted to the local archive.

Results

We identify cyclones using the Wernli and Schwierz algorithm applied to the six-hourly sea-level pressure.

We use this to find energy anomalies (from climatology) in cyclones:

total energy anomalies
historical (control) response (RCP8.5-historical)

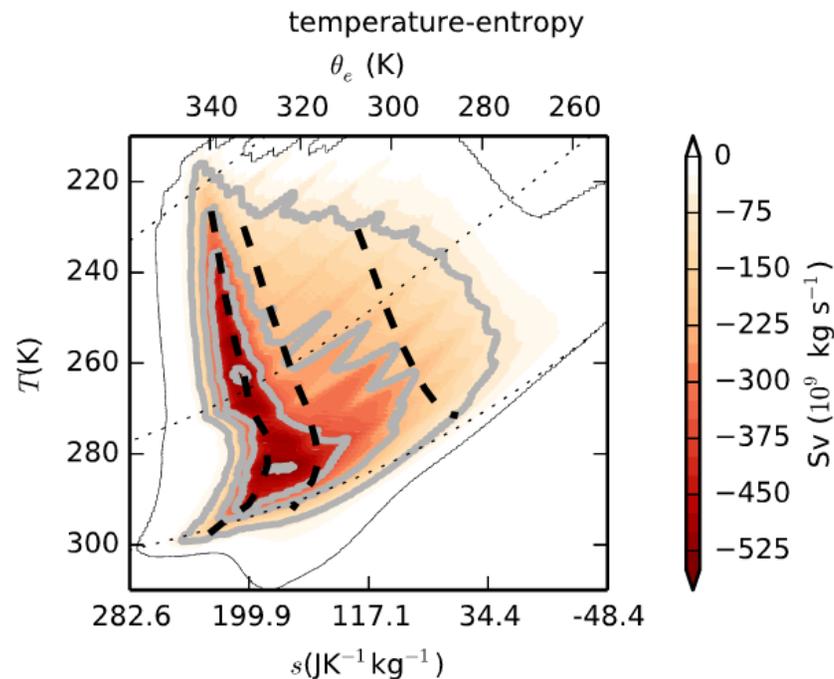
Legend: -4 -2 0 2 4 -8 -4 0 4 8
1e7 1e6

Frédéric Laliberté



More Results

- Want to allow even more complex climate diagnostics.
- For example, Laliberte et al. (2015, Science) which uses six-hourly data.



Summary

- Need to be able to "ask" using complex queries.
- Need to "validate" availability of discovered data.
- Require a format to share this information for reproducibility and efficient collaboration.
- Ideally, this format would contain all the meta-data from the netCDF header.
- Need a retrieval tool that can use the collected information.
- Ideally, the retrieval tool would seamlessly perform simple remote operations (subsetting).

Outlook

- If there was a minimum "standard" to pass the information, some of these features could be provided by a WPS.
- A "standard" format could include checksum information at the array level, improving reproducibility even more. For example, have the checksum for every time slice.
- More permanent URLs would greatly improve the stability of this approach.