

Experiences with the ESGF data portal at CCCma

Slava Kharin

Canadian Centre for Climate Modelling and Analysis
Environment Canada, Victoria, BC, Canada
WGCM Infrastructure Panel (WIP) member

“Typical” task

A “typical” user of model output from a multi-model ensemble:

- download data for one variable for a given frequency (say, daily) for a given experiment for **ALL** models (and **ALL** realizations);
- repeat the above for different variables, and different experiments, if needed.

Pre-ESGF era (CMIP3)

Download all daily files for one variable (tasmax) for historical experiment (20c3m) for **all** models/**all** ensemble members:

```
wget --user USER --password PWD --accept ".nc"\\  
  --recursive --continue \\  
ftp://ftp-esg.ucllnl.org/20c3m/atm/da/tasmax
```

Is it possible with an ESGF data portal?

Can the same simplicity be achieved (approximated) with an ESGF data portal?

A desired ESGF batch command line downloading tool:

```
‘esgf_command_line_tool’ --user USER --password PWD \\  
--variable variable_name \\  
--frequency frequency \\  
--experiment experiment_name \\  
[--realm ... --rip ... --model ...] \\  
--version/date
```

A new ESGF functionality (compared to CMIP3 `wget`) is version control. This is crucial for 'reproducibility' of results that are based on analysis of the archive contents downloaded on a particular date.

- The OpenID authentication is too complicated and restrictive.
- Search interface is too slow.
- Batch wget scripts require a proprietary version of Java, which is non-standard on most Linux desktops.
- The entire ESG foundation depends upon stability at PCMDI. Staffing and network issues at PCMDI still bring down the entire foundation from time to time.
- ESGF system is complex and delicate, depending on too many software packages. Unless you're one of the developers, it is unclear when upgrades to ESGF nodes are required and when the software is stable enough to upgrade without fear of trouble (not a “turnkey” system)
- Replicated data means that there is no feedback to data producers on how much and where their data is being used.

Thank you!