Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

The GeoMIP Perspective on Interactions with ESGF

BEN KRAVITZ

Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory

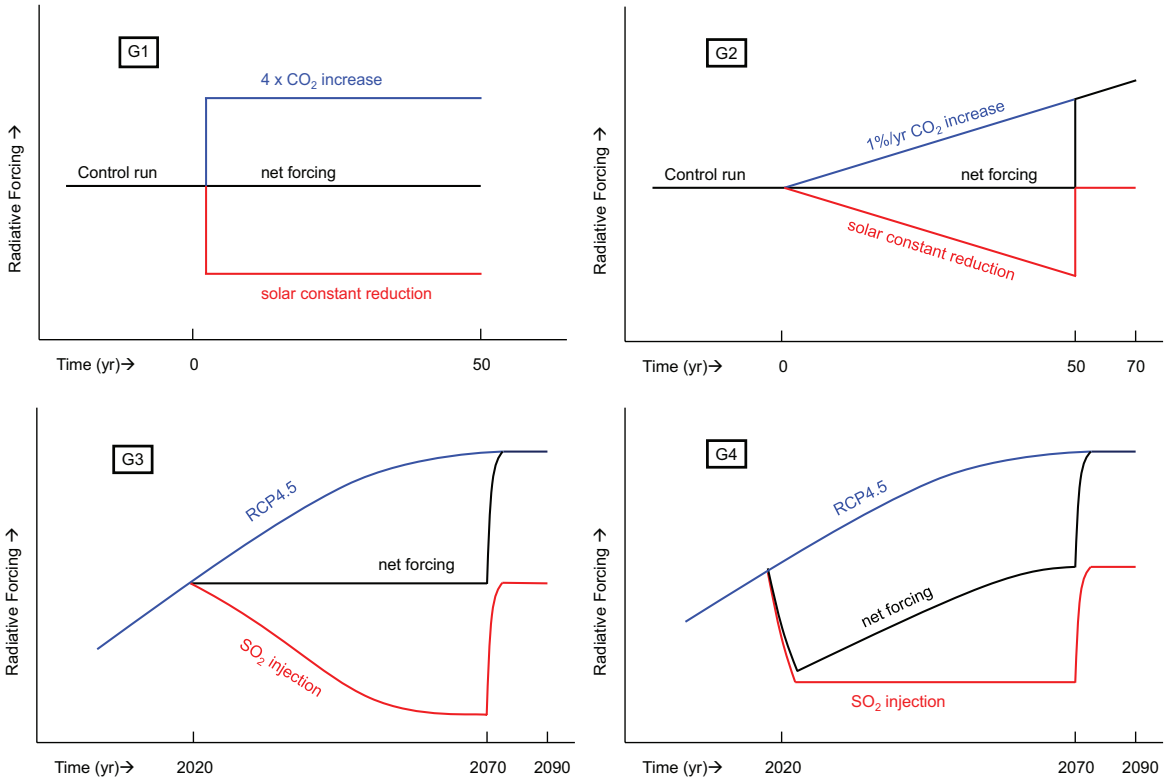
U.S. DEPARTMENT OF ENERGY

What is GeoMIP?

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- The Geoengineering Model Intercomparison Project
- Four core experiments (Kravitz et al., 2011) and three new ones (Kravitz et al., 2013)

GeoMIP Experiments G1-G4 for
SRM with Stratospheric Aerosols



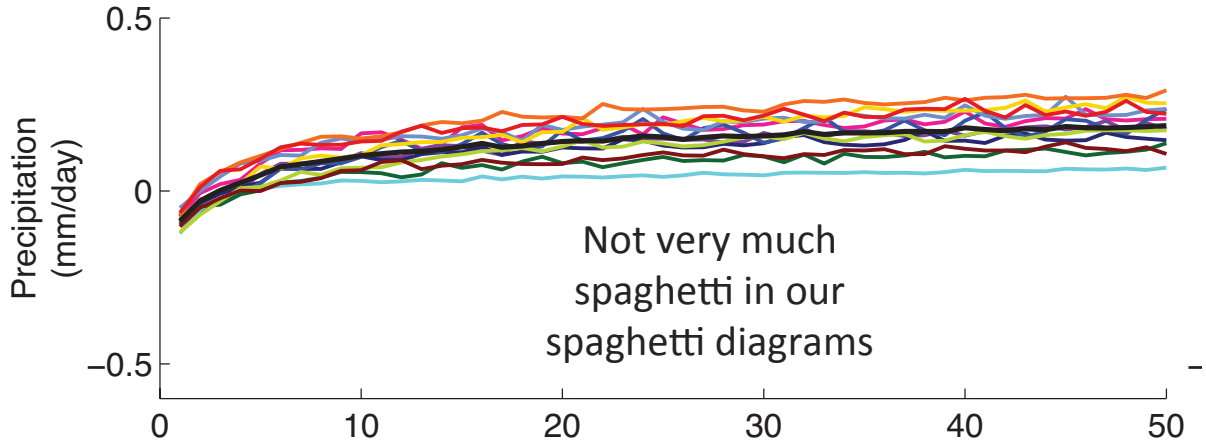
Insolation reduction to offset higher CO_2

Stratospheric sulfate aerosols

A little about GeoMIP's infrastructure requirements

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- GeoMIP is relatively small: ~12 models, each outputting 1-3 TB per round of experiments



- Experiments are built on a handful of standard CMIP5 experiments, so downloading requirements are minimal. I've written several papers using output stored on my laptop.

- Most participating modeling centers already have data nodes (those that have established data nodes have done so for other reasons, using GeoMIP as a catalyst)

Why is GeoMIP using ESGF?

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Infrastructure (ESGF is designed to hold an obscene amount of data, so what's another few terabytes?)
- Legitimacy (ESGF is where all of the cool kids hang out)
- Devoted staff who run the infrastructure (it takes a lot of work to provide output to individual requestors)
- It's free (for the end user)

ESGF was designed by the climate modeling community to meet the requirements of the climate modeling community. It's hard to think of a better system.

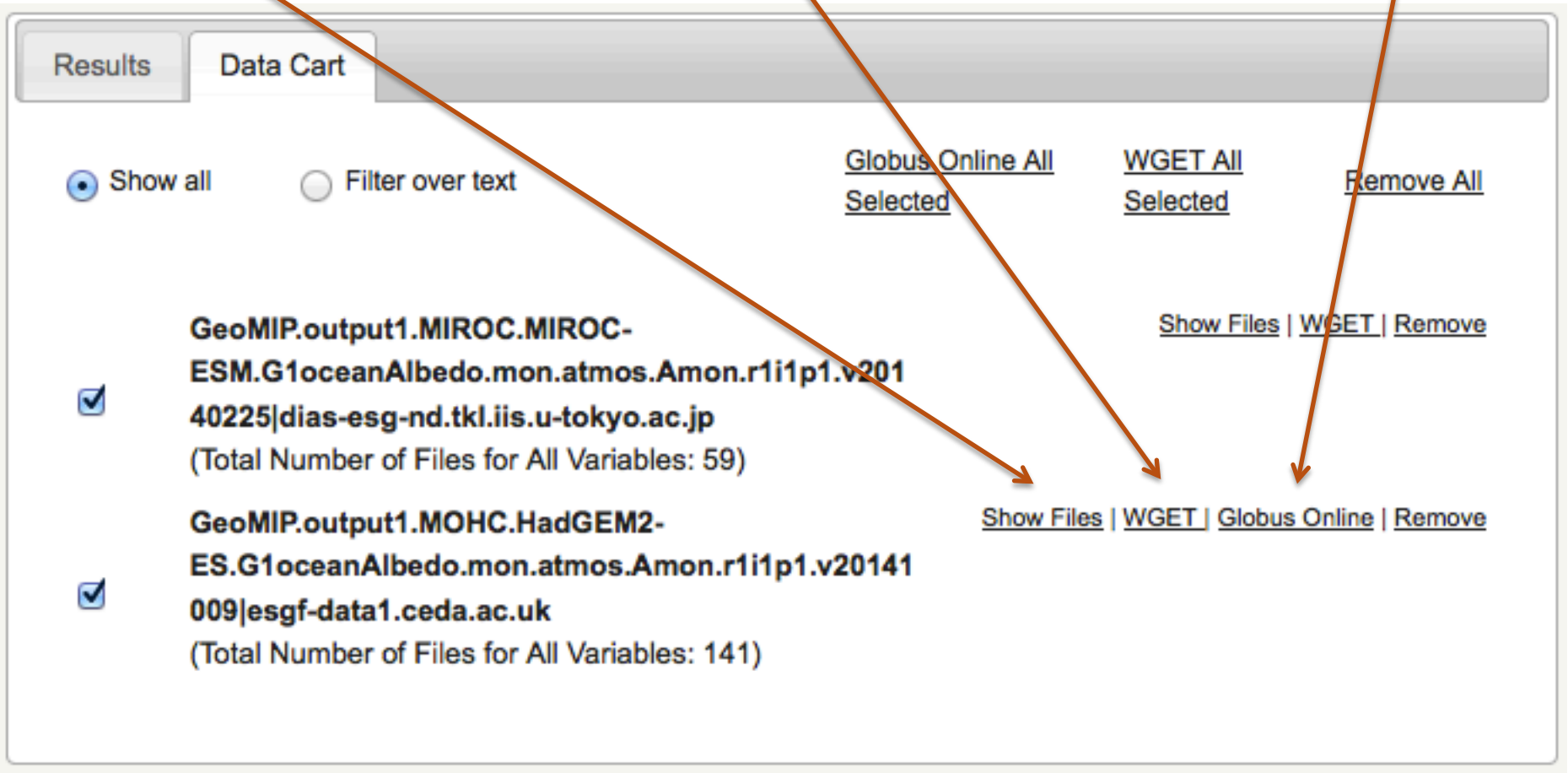
Downloading

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

http
(clicking one file at a time)

wget

Globus Online
(sometimes)



Downloading
(my own personal experience)

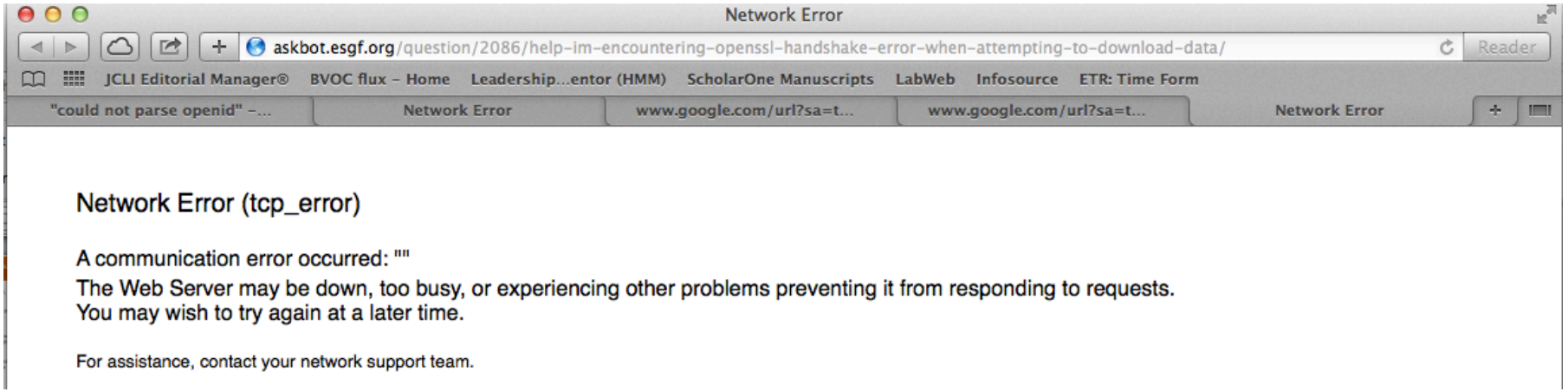
- wget: Getting this to work reliably seems to require a staff member who is willing to figure out all of the issues with certificates (and even then, it doesn't work for some nodes). Individual users (like myself) often don't have this, so doing analysis on lots of model output may be limited to large centers.

- Globus Online:
  - Not all data nodes have Globus Online capability
  - Even if they do, it doesn't always work
  - Designed for downloading ALL of the files, so not ideal for small requests (like mine)

- http (clicking one file at a time): This always works. It seems tedious, but GeoMIP's requirements are small enough that this is often a viable option.

Getting help…

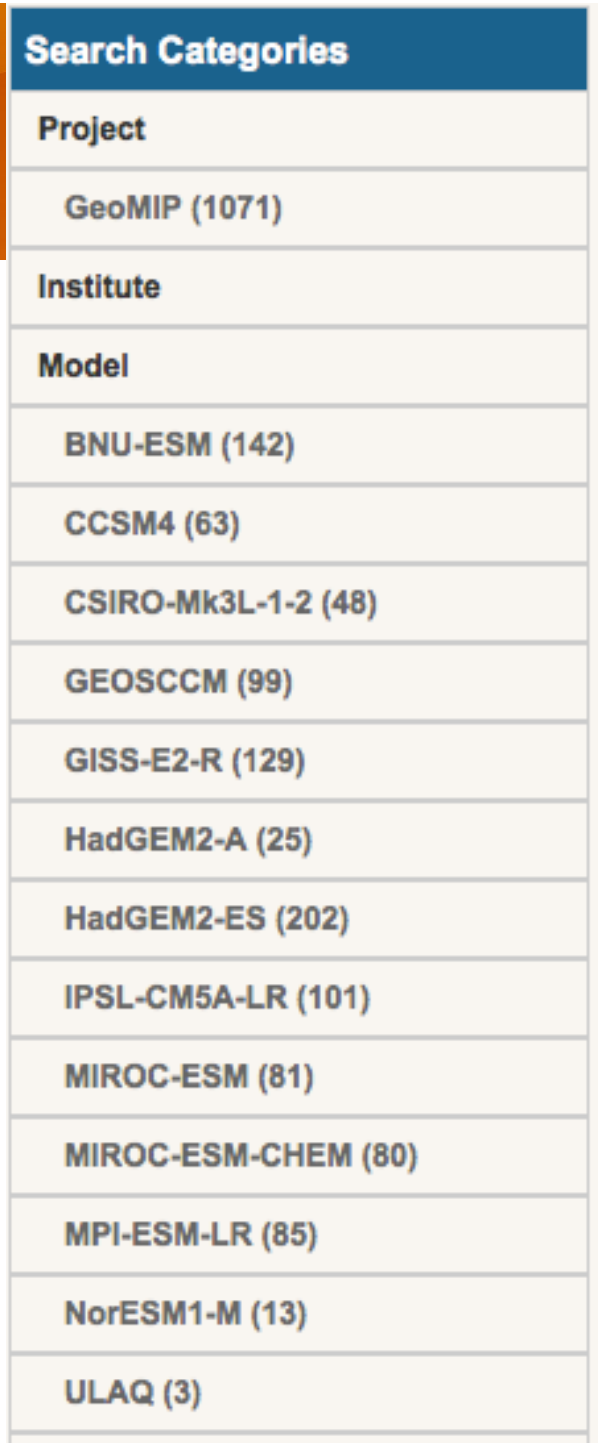
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*



Maybe I just have bad luck with askbot….

Hosting data

- Most modeling centers will simply add the GeoMIP output to their respective ESGF nodes.
- I don't have an ESGF node, so I've asked people to host my data for me (PCMDI in particular has been very nice about this). Relying on the kindness of strangers can only work for so long.
- Recently at PNNL, the IT folks have been trying to establish a data node (for many reasons). This is taking a lot of work – I'm not sure everyone has the capability of establishing a data node. This is also proving expensive – an FTE week at a DOE lab is not a negligible cost.



Establishing a data node at PNNL (notes from our IT specialists)

- ESGF is very complex, and there are lots of different options. The learning curve is very steep.
- Documentation for ESGF is scattered across many different places. It's difficult to know what the correct steps to follow are. Also, the rationale behind some of the steps is unclear, so it's hard to gain understanding from the instructions. "Even after installing, it isn't clear whether we're missing some important capability or configuration option."
- Hardcoded URLs in the install script.
- We are unaware of documentation on how to work with certificates.
- Upgrades can break the server. This has happened multiple times at PNNL.
- The server requires administration, which isn't free. Publishing files requires superuser privileges. The security diligence is also higher, as the server resides outside our firewall.

Conclusions

- ESGF is a great framework, and it's basically the only reason we've managed to be so successful. (Apologies for all of the negativity in my previous slides.)
- Some nodes work better than others, perhaps depending upon how much time and energy is spent setting everything up. This requires a lot of knowhow and resources (time, money, effort) that are dedicated to maintaining the nodes.
- A few suggestions: (next slide)

A few suggestions from my naïve perspective as an end user

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Documentation for both setting up the node and for using ESGF
  - We're a small operation, so we don't have a lot of manpower to do troubleshooting
  - Working with certificates is difficult
  - The PCMDI staff is great and always very helpful, but they've got their own work to do.
- Support
  - Dedicated support is expensive and possibly not viable
  - Askbot or user community (problems with website?)
- Suggestions from our IT people to benefit data hosts, end users, and the folks at PCMDI:
  - Having a zip file with precompiled versions of the ESGF software would guarantee that all necessary software is present and compatible.
  - Release of virtual machines (so superuser privileges aren't an issue)