CMIP6 and ESGF reciprocation: Case study using GFDL preparation to CMIP6

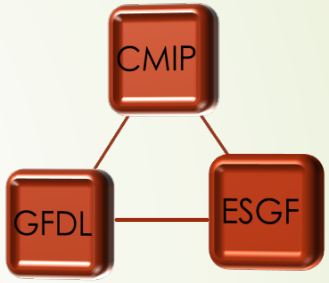
Sergey Nikonov[1,4], V.Balaji[1,4], Aparna Radhakrishnan[2,4], Hans Vahlenkamp[3,4]

[1] Princeton University, NJ
[2] Engility, NJ
[3] UCAR, CO
[4] GFDL, NJ

7th Annual ESGF F2F Conference.
Dec 4-8, 2017. San Francisco, CA.

Outline

It's simple – A Relationship Triangle



- CMIP (and CMIP6) as motivator of ESGF
- CMIP6 is a next challenge rising GFDL publishing infrastructure to higher level
- ESGF & GFDL - mutual influence
- GFDL' CMIP6 publishing preparation as example of these relationships

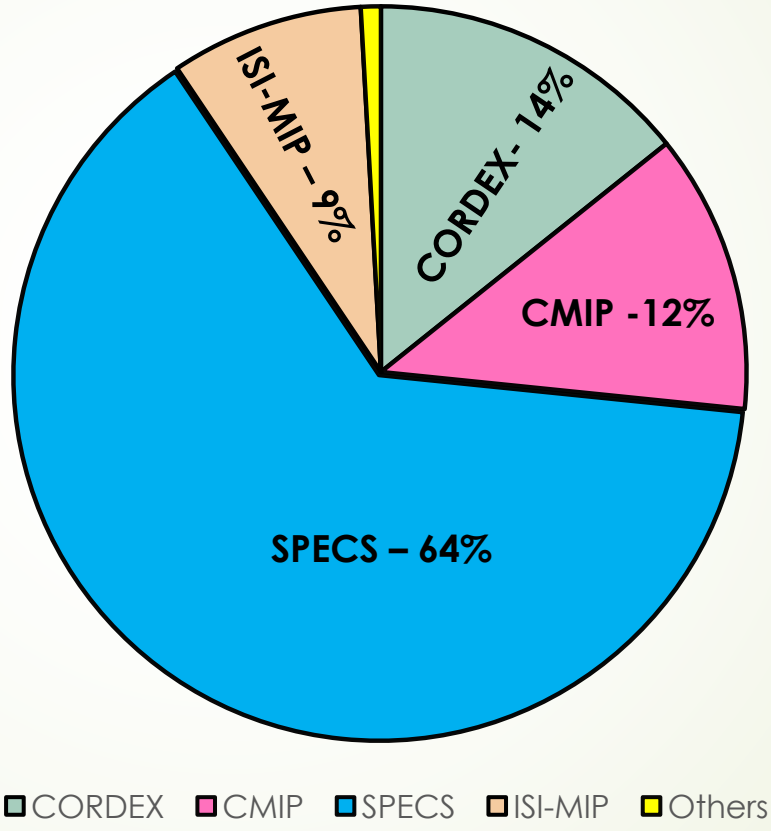
CMIP6 as Driver of ESGF

- We know that ESG started as data discovery/navigation/access system for CMIP5.
- Metadata standards were developed from CMIP and then expanded/elaborated/specified to other projects.
- CMIP diversity overcomes all other projects
- CMIP is still a main source of inspiration and ideas
- Now ESGF is so widely popular that CMIP is no longer the most prevalent content. But CMIP is a base for many derivative projects (downscaling, seasonal to decadal forecast). CMIP is a scientific fundament of any climate change research.

ESGF universal vehicle or set of project-oriented modes?

- CMIP is the one of main motivators for ESGF - analysis of datasets numbers on Nodes: 1 – SPECS, 2 - CORDEX, 3 - CMIP;
- CMIP6 will definitely be a leader in published dataset volume
- There is already distinguished specific in publishing CMIP6 and other projects (search facets, QC, publishing procedure, ESDOC, etc.)
- Different projects require different analysis web services
- Different projects need different facets in interface, sometimes not overlapping.
- Universalization of system can convert it to heavy all-in-one machine which is difficult to develop, maintain and use.
- Another pitfall can happen – dramatic influx of users and overloading servers. Splitting to several specialized servers can mitigate this potential problem.

Distribution of dataset numbers over projects



GFDL Infrastructure Evolution Driven by CMIP: CMIP3

- Newly developed CMOR package by PCMDI
- Main stress in CMORizing: archive bottleneck caused by CMOR runs
- Too much manual job
- No QC standards tool at all in workflow; people relied on their own practices and experiences
- Data delivery from GFDL HPC archive to GFDL Data Portal by FTP, and FedExing external HDs to PCMDI archive. FedEx was actually faster!

GFDL Infrastructure Evolution Driven by CMIP: CMIP5

- DB Curator was established with all metadata used in modeling, QC & publishing process
- Main stress in manually populating DB with variables' metadata
- Rudimental QC (just statistics moments, max/min, missed values); other analyses are scientist's choice
- frametar (analog of CMOR) based on Curator DB
- Copying data from local tape archive to external rotating storage by GridFTP. Transfer speed substantially increased versus FTP.
- Data delivery via ESGF and GFDL Data Portal located on separate servers mitigated "clogging potential" and mutually backed up each other

GFDL Infrastructure Evolution Driven by CMIP: CMIP6

- GFDL models migrated to CMIP variables names standards
- Metadata DB is populated automatically from dreqPy XML (thanks to Martin Juckes)
- QC is redeveloped cardinally. Several levels of QC are provided to scientists: one for deep scientific analysis and another for fast review of variable fields.
  - real time control of critical variables in a course of simulation;
  - detailed physical analysis is being done with ESMValTool, ILAMB, PMP during simulation and other GFDL analysis tools incorporated in dedicated web based CMIP6 Curator tracker;
  - MDBI (Model Development Database Interface) is equipped with LAS;
  - Global statistical characteristics and outlier are checked in MDBI before pushing variable to data portal
- Direct writing fremetarized (CMORized) data onto portal disk
- Passing each file through Prepare provides the 2nd control of metadata (1st – in femetar)
- Data delivery via ESGF only

GFDL Infrastructure Evolution

- **CMIP3**

  Main stress and bottleneck was CMOR and archive access. 90% of manual work. No any QC standards. Slow data transfer

- **CMIP5**

  DB populating with metadata and QC was the main agenda

- **CMIP6**

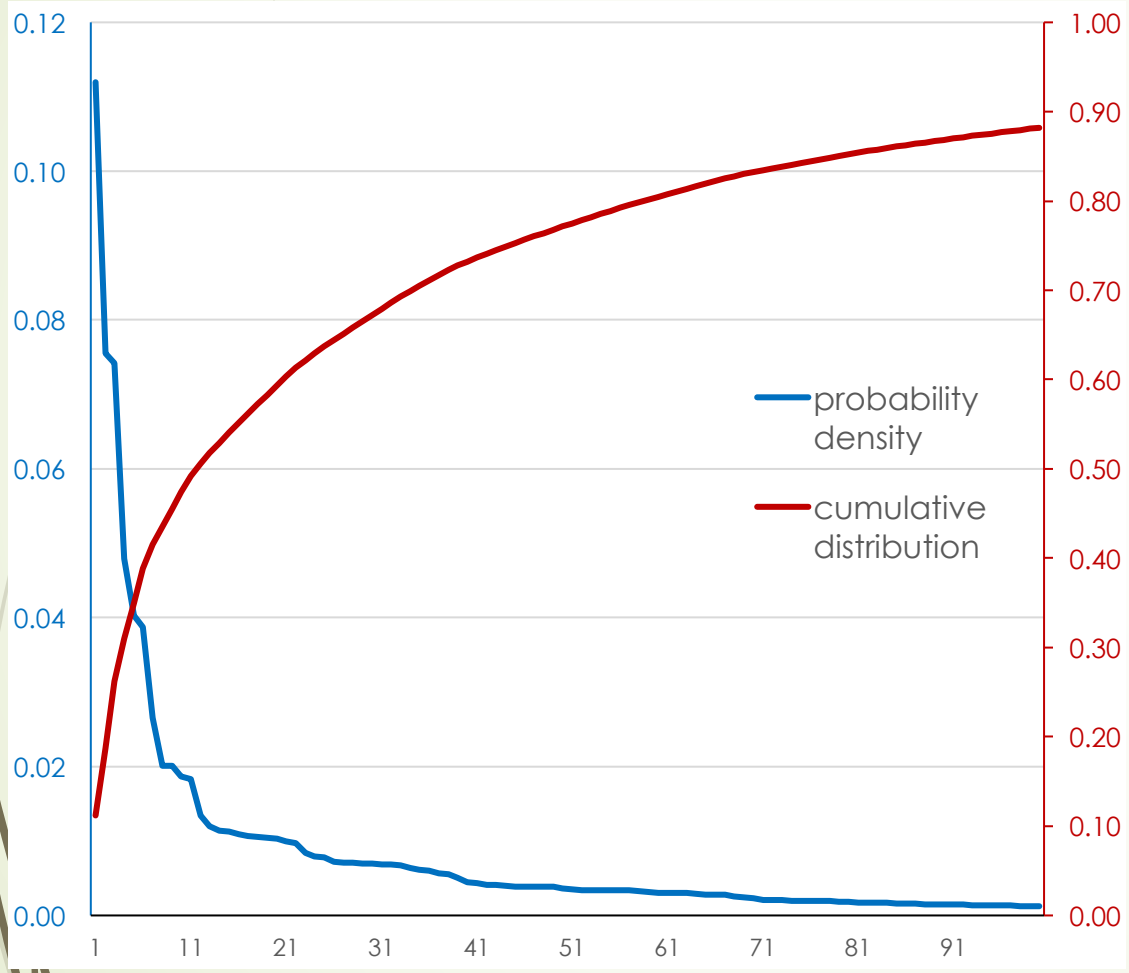
  We expecting storage limitation and data transfer will be the most popular topics

- **CMIP7**

  Rating variables by popularity should affect workflow and data access accordingly. Using CMIP6 statistics will reveal the probability of downloading different variables and can help in experiments design. Ranking them according to current popularity and may be mirroring them to fast dedicated servers.

Variables Download Probability

(courtesy CMIP5 statistic data of PCMDI and DKRZ)



air_temperature	11%
eastward_wind	8%
northward_wind	7%
precipitation_flux	5%
specific_humidity	4%
geopotential_height	4%
sea_water_potential_temperature	3%
surface_air_pressure	2%
relative_humidity	2%
air_pressure_at_sea_level	2%
surface_temperature	2%
surface_upwelling_longwave_flux_in_air	1%

New in Organization of CMIP6

- CMIP6 Data Request strict formalization with centralized XML DB is an important for such diversified cycle as CMIP6. At GFDL it allows populating and updating Curator DB – core of publishing workflow within an hour.
- Martin Juckes' dreqPy library is extremely useful. Now we can get programmatic request for info about all aspects of variables metadata including such important thing as CMIP table, dimension, frequency, cell measures / methods etc.
- CMOR & PrePARE based on CMIP6 Data Request suit and CVs set up new high standards of error proof in publishing workflow.

How CMIP6 can affect ESGF?

- CMIP6 content will increase substantially on ESGF Nodes
- It will raise interest from a large number of scientists around the world
- Differentiation for MIPs in CMIP6 will attract a diverse audience and will require a more specialized web services analytics
- Grown popularity attracts new participants to Federation with their new projects and ideas. As positive feedback it makes ESGF Suite more powerful, diversified, and robust.

GFDL Started CMIP6 Production

- GFDL scientists started CMIP6 official runs: CMIP AMIP & piControl
- AMIP is ready for publishing, piControl is running
- ESDOC is not started yet.
- Citation support: Citation Information gathering stage for AMIP run.
- Publication support: DOIs, access to model source codes, analysis via Jupyter notebooks and GitHub is being done for documentation and publishing article
- Publishing on ESGF is in a progress. 1st test is successful.

GFDL Connection to ESGF

- GFDL has long history of relationships with PCMDI starting from beginning of century - early versions of CMOR.
- GFDL has been running ESGF Node since the 1st version in 2009.
- That time we had GFDL developed CMIP5 site as a ESGF backup. Now GFDL considers ESGF Node as the main and only means for data delivery of big projects like IPCC.
- According to GFDL-Google collaboration project ESGF will be deployed via Docker image on cloud and CMIP data will be hold and served from there.
- Ideally, would be move other ESGF Nodes there. It will allow to avoid replication.

CMIP7: what to expect and how to influence

- All positive/negative CMIP6 experience are motivators for future development. We should carefully record and analyze it. From my CMIP5 experience some details were lost and we could move faster if kept them in records.
- It's time to jump to new era - using cutting edge technologies such as machine learning, data mining, clusterization, etc.):
  - QC is the 1st and most urgently needed candidate
  - data squeezing using categorization also can be tried
  - Dynamic ranking variables based on download probability and make a search according to rank (like a Google raising the hottest item in a result list)
  - filtering robots (or too avid customers) overloading Nodes by downloading all data
- Clouds! GFDL starts project with Google for moving simulations and ESGF Node to clouds. Ingress and egress are free!

Some Obvious Conclusions

- CMIP is undoubtedly the most influential project within ESGF
- ESGF started from CMIP needs
- ESGF became a serious player to influence climate science projects design
- CMIP cycles are engine moving ESGF to the next level of science and technology (Clouds, Big Data, Machine Learning, Data Mining)
- GFDL has good lessons from CMIP cycles improving infrastructure and from ESGF feeding on new ideas and technologies