

# CMIP5 early assessment

IGIM

Ron Stouffer with lots of help

October 2012

# CMIP

- Climate “Science” side
  - Experiments runs
  - Variables saved, QC’ed, made public
  - METAFOR
- Data serving (computer science) side
  - Security
  - Servers interacting with each other
  - Visualization
  - Bandwidth and other hardware issues
  - Analysis on servers

# CMIP5 – Good things

- **Amazingly complex, extraordinary increase in scope compared to CMIP3 and relative to *any* other database in world.**
  - It must be counted a success that the experiments were run/data archived.
  - CMIP3=40TB, CMIP5=1.5 PB+, at least 40X bigger! And growing....
  - Number of experiments also up by some big X factor.
- **Distributed data management system was a first!**
  - Amazing accomplishment.
    - Complexity not well appreciated by users.
  - Software effort brought together people working on many different aspects of the software.
- **The systems set in place for CMIP are extensible.**
  - ESGF architecture is designed to be scalable and can be extended to meet future data needs
  - All other MIPs are and should be encouraged to embrace and extend the model, including ESGF nodes, DRS (vbl names/units) structure and CIM metadata (METAFOR).

# CMIP5 – Things that need work

- **Infrastructure funding specifically targeting CMIP5 was insufficient, resulting in delayed achievement of some of the ambitious goals.**
  - ESGF was funded as a “research” project to develop a system that would serve communities broader than CMIP. (CMIP was one of 4 “use” cases.)
  - Efforts to make the system work operationally for CMIP were initially underfunded.
- **Governance model was informal with poorly understood procedures for decision making**
  - ESGF governance needed to be in place earlier
  - Uncertainty and disagreements on the timing of upgrading the system meant a delay in integration of the METAFOR model documentation effort with ESGF
- **Capabilities that were important to scientists attempting to meet IPCC-dictated deadlines were not deployed in time:**
  - Access to CIM metadata
  - Data citation mechanism using DOIs
  - Replication of data to improve accessibility
  - Quality assurance checks on model output
- **We are all downloading gobs of data instead of leaving it on the servers.**

# CMIP5 – Things that need work

- **Modeling Groups were very late making data public**
  - In Feb 2012 there was very little carbon variable data available from ESMs where atm pCO<sub>2</sub> was predicted. Better now?
  - Tension between making data public and writing papers on results
    - Especially a issue for “new” runs – like the ESMs this time
  - Most physical variables available in Feb 2012...which was very late
    - WGCM agreed to have the data public by Jan 2011 (!) in Paris (~Sept 2008)
  - Modelers always want to delay until the last possible moment to get the “latest” into the model
- **METAFOR**
  - Lots of effort to get people to fill out forms
  - Very little feedback to modeling groups to date
  - Publish and QC took lots of time
  - If we plan to use this again – there needs some feedback now and some encouragement for the groups. The usefulness is not clear to me as it exists. Large potential, but unrealized.

# CMIP6(7,8,..) - Action Items

- **Encourage the formalization of ESGF with international inter-agency agreements, orchestrated by WCRP.**
  - WCRP and NRC endorsement should be capitalized on by agencies to increase base funding for a global data infrastructure.
  - Governance proposal being developed
  - **Infrastructure cannot be financed with soft money! See recent NAS report**
- **Conduct survey not only of users but also of data providers.**
  - This was missed in CMIP3.
- **Data providers -- aka modeling centers – are key to making this all happen.**
  - No overarching group that covers “science” side, software side
  - Role of WCRP Data Council unclear and could easily hinder things
- **All MIPS should follow the lead and standards set by CMIP5.**
  - With that understanding, CMIP could be divided into smaller, more focused and manageable sets of experiments, which would be less disruptive to the scientific life at the centers.

# Summary

## Computer Science Side

- CMIP5 distributed data base is a remarkable accomplishment
  - World's most complex distributed database
    - GOOGLE, CERN LHC databases probably larger
  - Working fairly well at present
- Governance and funding models need changed for future CMIPs
  - This needs to begin NOW.
- All other MIPs should use CMIP standards
  - Other “standard” development should be discouraged
  - Modeling groups need to be central to process

# Summary (con't)

## Computer science side

- WGCM use PCMDI as its main connection to ESGF
  - Relationship has worked well in all CMIPs
  - WGCM support of PCMDI helped organize software folks
  - PCMDI a leader in serving data

# Future CMIPs

## Climate Science Side

- Time slice experiments need to be better defined
- Standardize 20<sup>th</sup> C forcing fields
  - Aerosols – natural and human-made – conc and emissions
  - Solar
  - Volcanoes (?)
- Easier to perform MIPs
  - 1 giant MIP (ala CMIP)
  - Many smaller MIPs
- Tuning issues
  - Using estimates of historical forcing changes in tuning
  - Metrics - Circular testing?
    - Tune to present day and then evaluate present day simulation

# CMIP data serving and availability

## Users

- Modeling groups late in making data public
- CMIP5 is arguably the world's largest, most complex database – **now working**
  - Users need to be familiar with models, exps and vrbls
    - Need for assistance for some groups of users
    - The number of experiments/variables/models is daunting
  - Need for downscaling model data – space and time
    - Some of the high resolution (25km) better other CMIPs but only partly addresses problem



# CMIP data serving and availability

## Data Providers

- Modeling groups late in making data public
  - Trade-off between getting in latest stuff and timeline
  - Groups underestimated effort to CMORize data (again!)
- CMIP5 is arguably the world's largest, most complex database – **now working**
  - Software and governance issues hindered getting database functional
  - 1.5 pb of data available (40X CMIP3) and growing!