

Data Storage and Distribution: Lessons from the CMIP3

Karl E. Taylor

Program for Climate Model Diagnosis and Intercomparison

Presented to the
**WCRP Workshop on
Evaluating and Improving Regional Climate Projections**

Toulouse, France

13 February 2009

Outline

- What was done in *CMIP3*
- What was done right?
- What problems were encountered?
- How can the process be improved?

What was done?

- WCRP's WGCM, representing the major modeling centers,
 - Conceived and established CMIP3 following its much more limited earlier CMIP efforts.
 - Set up oversight by the CMIP panel
 - Decided on the set of experiments to be performed.
 - Asked PCMDI to support CMIP3 (in particular, to collect and serve CMIP3 output to "WG1 scientists")
 - Dictated terms of use: for non-commercial purposes only. (exceptions made for U.S. data).
- Modeling groups
 - post-processed their model output, transforming it to facilitate analysis by others.
 - Sent the output to a central archive (PCMDI)

- Model output (36 Tbytes) collected from the modelling groups by PCMDI was made available to others via ftp and the web.
- A subset of output was transferred to the IPCC's DDC to help serve WG2.
- Model output (~500 Tb) was accessed by scientists (~2000) all over the world who analyzed it.
 - Now over 500 publications, which relied on the CMIP3 data archive, have been self-reported on the PCMDI website.

What contributed to the success of CMIP3?

- The WGCM requirements, as made precise and explicit by PCMDI, were met by all the modeling groups.
 - Participation was seen as mandatory (because of the expected importance of CMIP3 to the IPCC's AR4).
 - This was a huge effort on the part of the individual modeling groups.
- A research group (PCMDI) had a mission and funding that permitted dedicated support of the CMIP3 project
 - Scientific input on experiment design
 - Collection and support of data archive (including 40 Tb RAID and server)
 - Software for rewriting model output (CMOR) and for browsing and serving the data (ESG)
 - Website development

What contributed to the success of CMIP3 data storage and distribution?

- The requirements were well-defined:
 - Simulations
 - Model output (list, time-periods, units, etc.)
 - Data formats, metadata, and structure
- Software (CMOR) was developed to facilitate adherence to the standards
- Quality control checks performed by CMOR trapped mistakes in model output so they could be corrected prior to transfer to PCMDI.
- PCMDI checked that sample model output met all requirements before accepting the full contribution.

What contributed to the success of CMIP3 data storage and distribution?

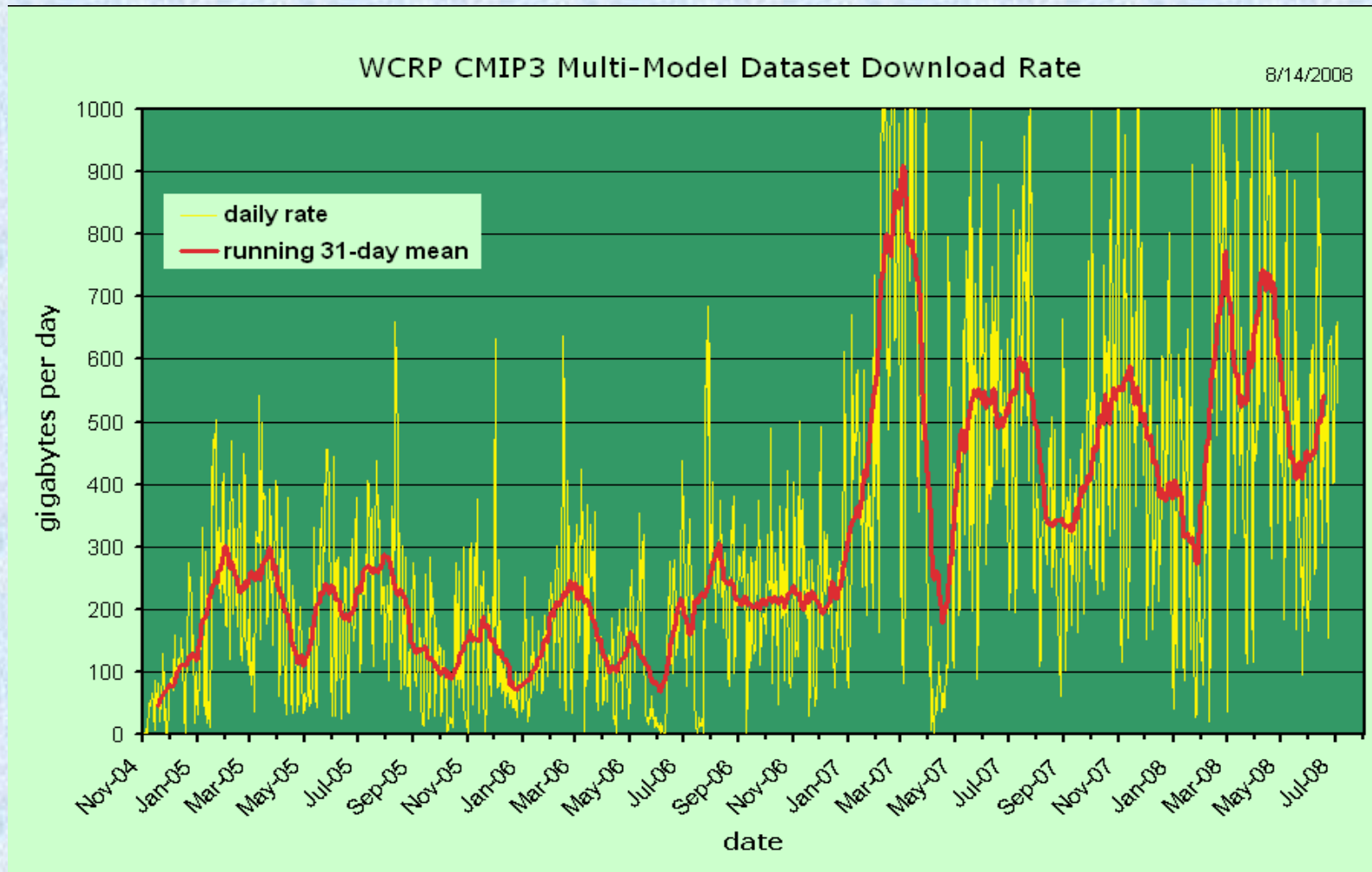
- An errata page was maintained to alert users of identified problems with the dataset.
- Extensive CMIP3 website was developed to provide information to both the modeling groups performing the experiments and to the users.

The payoff for rewriting data in conformance to strict conventions:

Output from multiple models:

- Could be downloaded from a single site.
- Could be analyzed in a fairly automated way, since all the output conformed to specific conventions.
- Has been scrutinized more comprehensively than ever before.
- Continues to be of value to a rather broad range of scientists.

Interest in CMIP3 has increased since the AR4 Cumulative download total exceeds 260 Tbytes



What problems were encountered?

- Unfamiliar new requirements for processing output confronted modeling groups
 - CF-conforming files with very specific metadata required
 - CMOR software facilitated conformance, but was new and the earliest groups encountered bugs.
- Requirement to map output to a Cartesian latitude-longitude grid resulted in some groups not contributing ocean output.
- When errors were found in model output, correction required intervention of several individuals.

What mistakes were made?

- We placed too much confidence in our RAID system, and a convenient backup to it was not initially in place.
- We purchased 1 Tb disks transfer data from modeling groups to PCMDI, which were not "top of the line" and turned out to be flaky.
- We failed to communicate sufficiently with the climate effects and "impacts" scientists, so some model output they would have found valuable was not saved.

How can the process be improved?

- Improve convenience and "user-friendliness" of data serving in several ways:
 - implement capability to extract subsets of data and perform simple server- side calculations (e.g., obtain a single pressure level, a climatological mean, a zonal mean)
 - refine somewhat confusing registration procedure
 - improve catalog search capability
 - ingest model documentation and expt. details into a searchable database
 - Improve errata notification

How can the process be improved?

(cont.)

- Make earlier plans; don't underestimate the work.
- Move toward a distributed database (i.e., avoid transferring output from modeling centers to a single repository)
 - Requires reliable software that mimics the look of current single site archive (e.g., ESG, OPeNDAP extension).
 - May not completely supplant a central repository since some groups can't serve data.
 - Avoids bandwidth and logistical problems transmitting data to a single archive.
 - Allows groups to immediately correct their output when errors are found.
 - Minimizes single point of failure issues.

Summary of lessons learned in previous MIP's

- Don't call for an overly-elaborate set of experiments. [We hope CMIP5 hasn't violated this rule.]
- Articulate clearly the science (or other) objectives.
- Precisely define the expt. design and the output required.
- Require some model documentation prior to accepting model output for distribution.

