Managing Data for Climate Model Intercomparison: The User Perspective

Reto Knutti
Institute for Atmospheric and Climate Science
ETH Zurich, Switzerland
reto.knutti@env.ethz.ch

Symptoms of hitting a wall

- Uncertainties in projections across models do not decrease
- Criteria for a good model are unclear
- Ensembles of models are hard to understand
- Results are of limited value for end users

- Models are slow and produce too much data
- Download and analysis of data is painful

Motivation

A not so unusual example

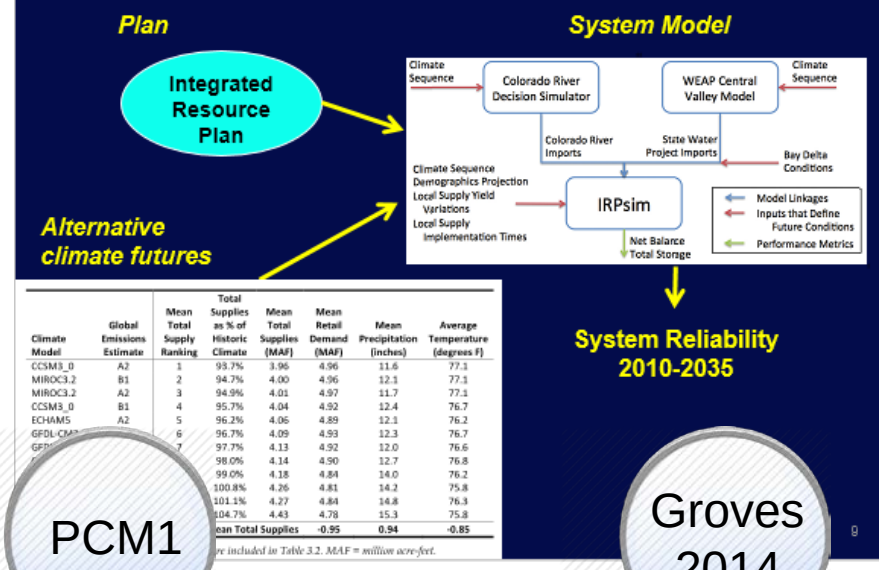
What is a Good Decision?

No universal criterion exists, but good decisions tend to emerge from processes in which people are:

- Explicit about their goals
- Consider a range of alternative options
- Consider tradeoffs
- Use best available science to understand the potential consequences of their actions
- Contemplate the decision from a wide range of views and vantages
- Follow agreed-upon rules and norms that enhance the legitimacy of the process and its outcomes

RAND IPCC AR5 WGII Chap 2 54

Combine Climate Projections and Model of MWD System to Consider IRP's Performance



Climate Model	Global Emissions Estimate	Mean Total Supply Ranking	Total Supplies as % of Historic Climate	Mean Total Supplies (MAF)	Mean Retail Demand (MAF)	Mean Precipitation (inches)	Average Temperature (degrees F)
CCSM3_0	A2	1	93.7%	3.96	4.96	11.6	77.1
MIROC3.2	B1	2	94.7%	4.00	4.96	12.1	77.1
MIROC3.2	A2	3	94.9%	4.01	4.97	11.7	77.1
CCSM3_0	B1	4	95.7%	4.04	4.92	12.4	76.7
ECHAM5	A2	5	96.2%	4.06	4.89	12.1	76.2
GFDL-CM2		6	96.7%	4.09	4.93	12.3	76.7
GFD		7	97.7%	4.13	4.92	12.0	76.6
			98.0%	4.14	4.90	12.7	76.8
			99.0%	4.18	4.84	14.0	76.2
			100.8%	4.26	4.81	14.2	75.8
			101.1%	4.27	4.84	14.8	76.3
			104.7%	4.43	4.78	15.3	75.8
			ean Total Supplies		-0.95	0.94	-0.85

re included in Table 3.2. MAF = million acre-feet.

PCM1

Groves 2014

9

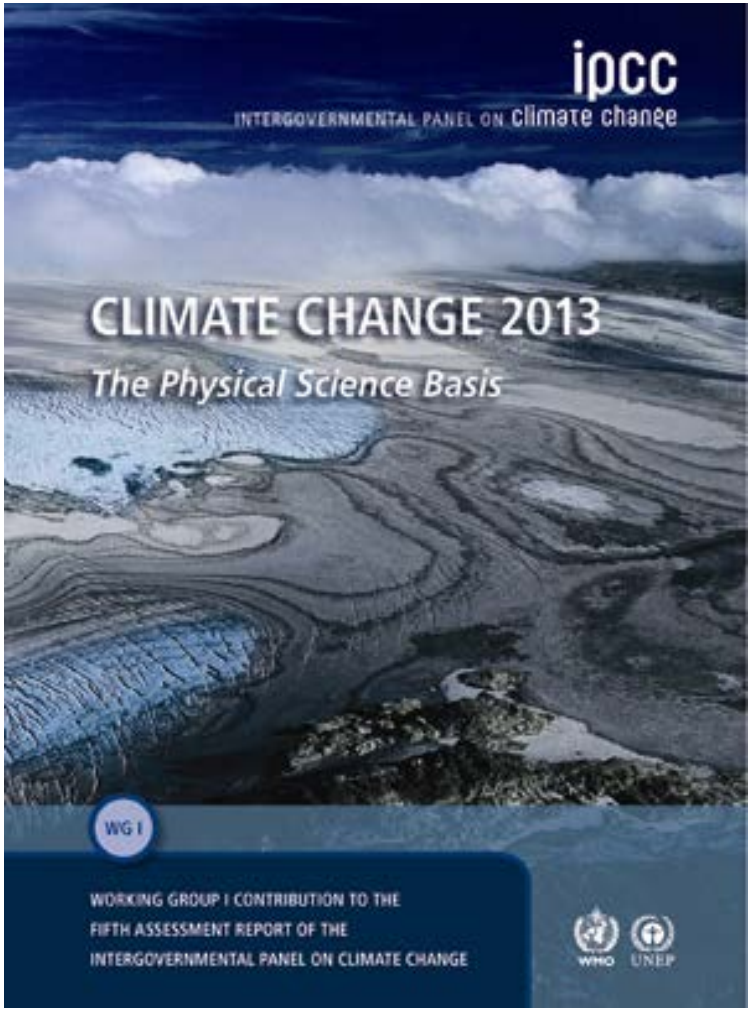
Slides courtesy of Rob Lempert

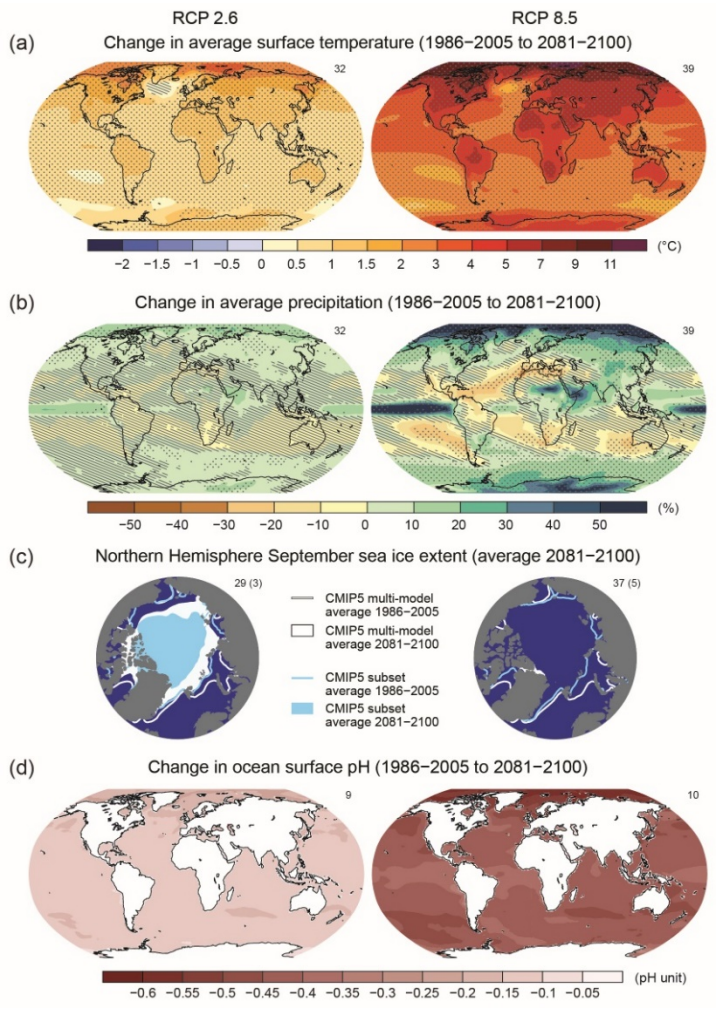
Challenges wrt model intercomparisons
faced in IPCC and other projects

- Sheer amount of data in CMIP5: ~ 3 Petabyte distributed across centers → Storage and bandwidth problem
- Dimensionality: lat x lon x height x time x hourly/daily/monthly x variable x mean/extreme/… x model x model version x ensemble member x scenario
- Model simulations are always delayed… only weeks to produce results
- Data quality: 1) technical sense (completeness, units, format), 2) scientific sense
- Evolving database rather than once produced and published
- Traceability, user notification
- Distributed system: performance, coordination, downtime

Multimodel results

therefore require some analysis platform





Analysis platform

The ETH Zurich CMIP5 snapshot

- Need for a single, (reasonably) quality controlled subset of CMIP5 data, immediately available, simple to use, fast, reliable, automated synchronisation to various sites
- ETH Zurich archive: 100 TB, half a million files, simple directory structure
- Single command synchronisation

  Get list of filenames and their corresponding md5 checksum and creation date

  ```
  rsync -vrlpt cmip5user@atmos.ethz.ch::cmip5/filelist.txt .
  ```

  Get monthly mean of maximum surface temperature data from historical runs:

  ```
  rsync -vrlpt --delete
  cmip5user@atmos.ethz.ch::cmip5/historical/Amon/tasmax
  cmip5/historical/Amon/
  ```

- Frozen in March 2013 for IPCC, now permanently archived at DKRZ

Analysis platform

The ETH Zurich CMIP5 snapshot

- Problem: Earth System Grid (ESG) distributed, slow, unreliable: How do we distinguish database error, file error, site down, data withdrawn, data being fixed?
- Workaround: reverse engineering ESG, >20 clients running scripts to search new (and old) data 24/7, lots of scripts trying to intelligently find gaps, errors, overlaps.
- Limitations of our approach: impossible for whole archive, no authentication
- Advantages: users sync quickly, automated, works. Consistent dataset across groups, transparency, traceability.
- General limitations of platforms: Lots of work to manually fix technical problems, No scientific evaluation!
- Files changing every second: When to stop? How do we ensure quality?

Lessons learned

and suggestions for future efforts

- Distributed data makes sense but has been problematic
- Analysis platform needed, mirrored snapshots ok for most,
- Simple file system is enough, scriptable interface to sync
- 100 TB serve the needs of almost all users, grows as needed
- No authentication
- Technical or scientific quality control: by modeling groups, PCMDI, IPCC? Need for a “clean” CMIP subset.
- Constantly evolving data raises technical and scientific issues:
  User notification, error reporting, need for database for verify file status
  Version control (flag vs remove, versions can only increase)
  Unique IDs, consistency of metadata with files on disk
- Think beyond running the model, share efforts across centers
- Exciting data science, or “boring storage”? Funding?