# Climate modeling at GFDL: challenges for the next cycle

V. Balaji
Princeton University

Petascale Computing in the Geosciences
San Diego Supercomputing Center
5 April 2006

# From a recent issue of Nature...

”Milestones in Scientific Computing”, from Nature (23 March 2006)

\>>

**1976** At Los Alamos, Seymour Cray installs the first Cray supercomputer, which can process large amounts of data at fast speeds.

1980s>>

**1983** Danny Hillis develops the Connection Machine, the first supercomputer to feature parallel processing. It is used for artificial intelligence and fluid-flow simulations.

**1985** After receiving reports of a lack of high-end computing resources for academics, the US National Science Foundation establishes five national supercomputing centres.

**1989** Tim Berners-Lee of the particle-physics laboratory CERN in Geneva develops the World Wide Web — to help physicists around the globe to collaborate on research.

1990s>>

**1990** The widely used bioinformatics program Basic Local Alignment Search Tool (BLAST) is developed, enabling quick database searches for specific sequences of amino acids or base pairs.

**1996** George Woltman combines disparate databases and launches the Great Internet Mersenne Prime Search. It has found nine of the largest known Mersenne prime numbers (of the form $2^n-1$), including one that is 9,152,052 digits long.

\>>

\>> PERSONAL COMPUTERS... >> INTERNET...

\>>

**1996** Craig Venter develops the shotgun technique, which uses computers to piece together large fragments of DNA code and hastens the sequencing of the entire human genome.

**1998** The first working quantum computers based on nuclear magnetic resonance are developed.

21st CENTURY>>

**2001** The National Virtual Observatory project gets under way in the United States, developing methods for mining huge astronomical data sets.

**2001** The US National Institutes of Health launches the Biomedical Informatics Research Network (BIRN), a grid of supercomputers designed to let multiple institutions share data.

**2002** The Earth Simulator supercomputer comes online in Japan, performing more than 35 trillion calculations each second in its quest to model planetary processes.

**2005** The IBM Blue Gene family of computers is expanded to include Blue Brain, an effort to model neural behaviour in the neocortex — the most complex part of the brain.

**2007** CERN's Large Hadron Collider in Switzerland, the world's largest particle accelerator, is slated to come online. The flood of data it delivers will demand more processing power than ever before.

**Jacqueline Ruttimann**

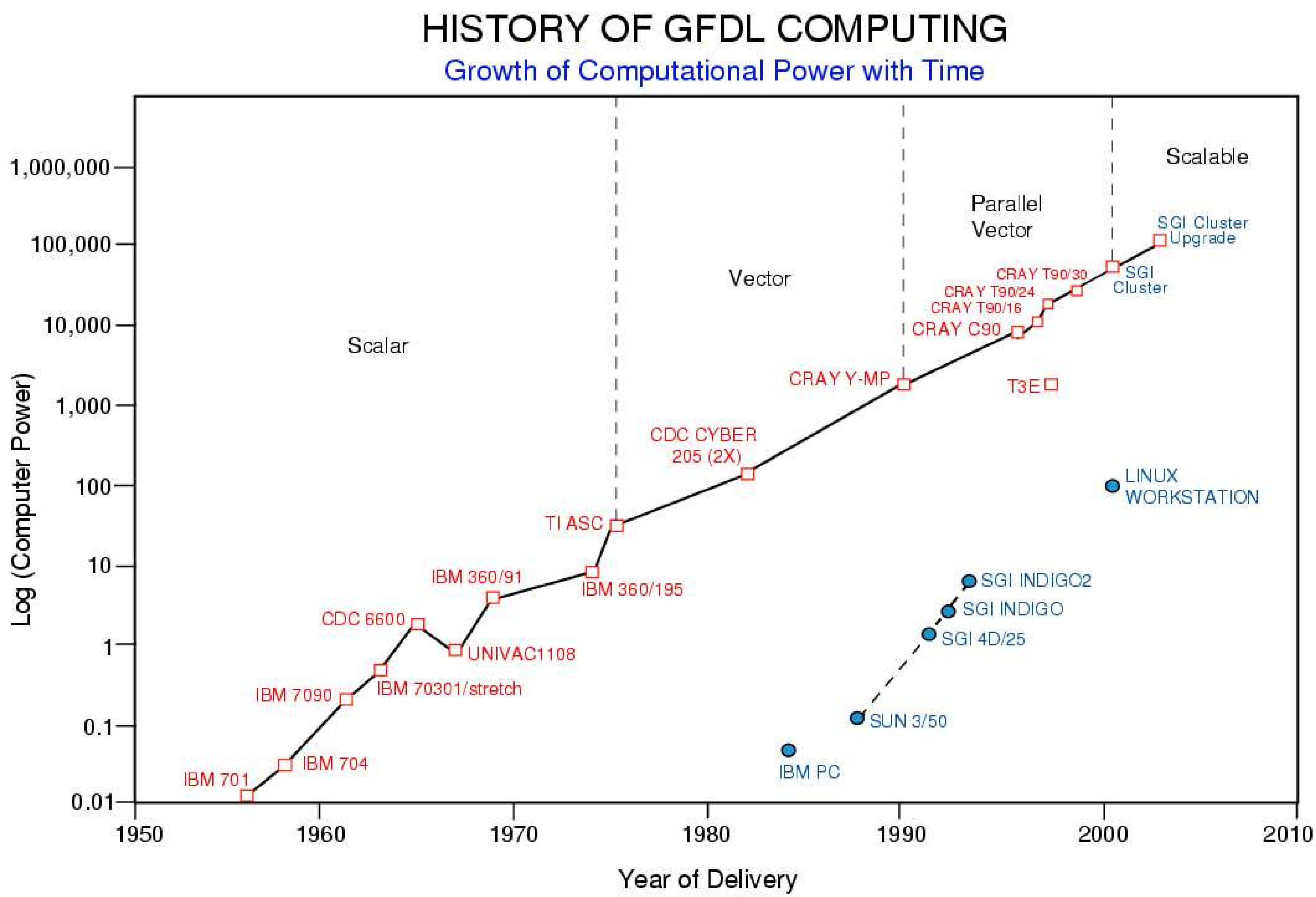
\>> IMPLICIT COMPUTING...

Among the milestones listed are:

- 1946 ”ENIAC, ... the first electronic digital computer”
- 1972 ”.. the first hand-held scientific calculator”
- 1989 ”Tim Berners-Lee ... develops the World Wide Web”
- ...
- 1969 Results of the first coupled ocean-atmosphere general circulation model are published by Syukuro Manabe and Kirk Bryan, paving the way for later climate simulations that become a powerful tool in research on global warming.

**http://www.nature.com/nature/journal/v440/n7083/full/440399a.html**

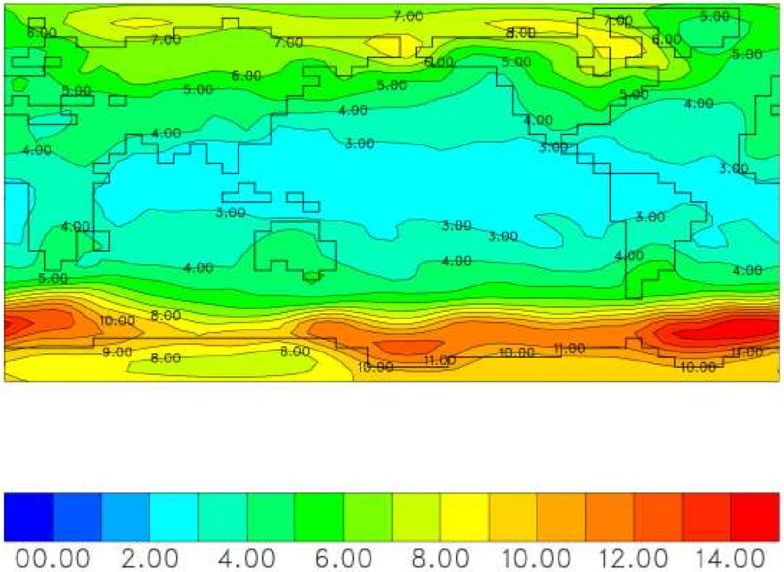
# History of GFDL computing



# Mid 70s: TI ASC era

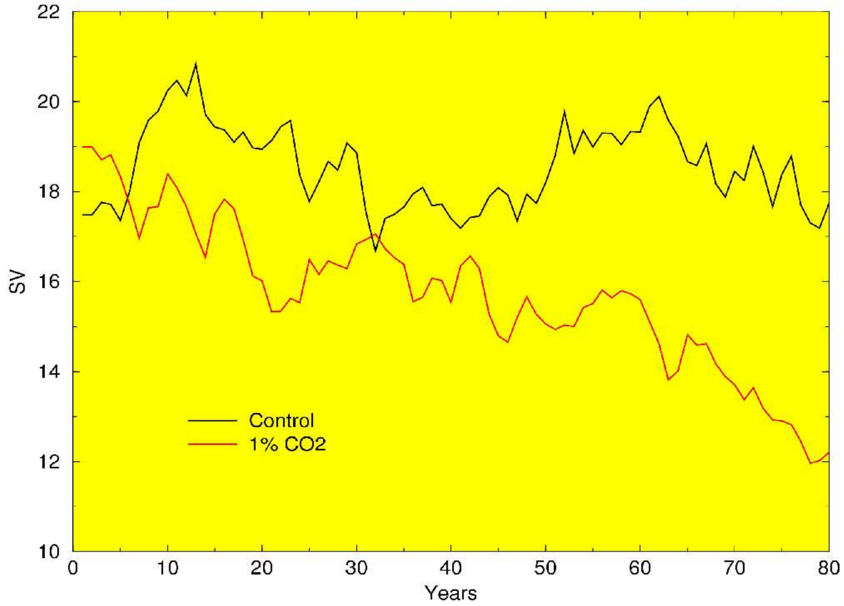
(History of computation and science at GFDL courtesy Ron Stouffer).

- Model: R15 atmosphere coupled to mixed-layer ocean.
- Speed: 2 model years per day.
- Run length 20 years.
- Key results: polar amplification under $2\times CO_2$, land warms faster.



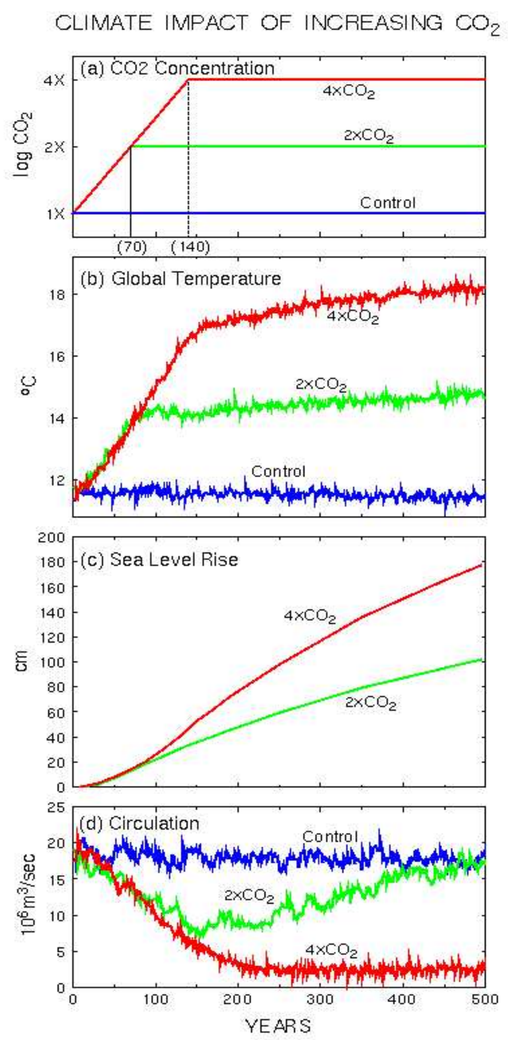
# Early 80s: Cyber-205

- Model: R15 atmosphere coupled to MOM ocean.
- Speed: 3 years/day.
- Runlength: 100 years.
- Key results: weakening of overturning circulation.



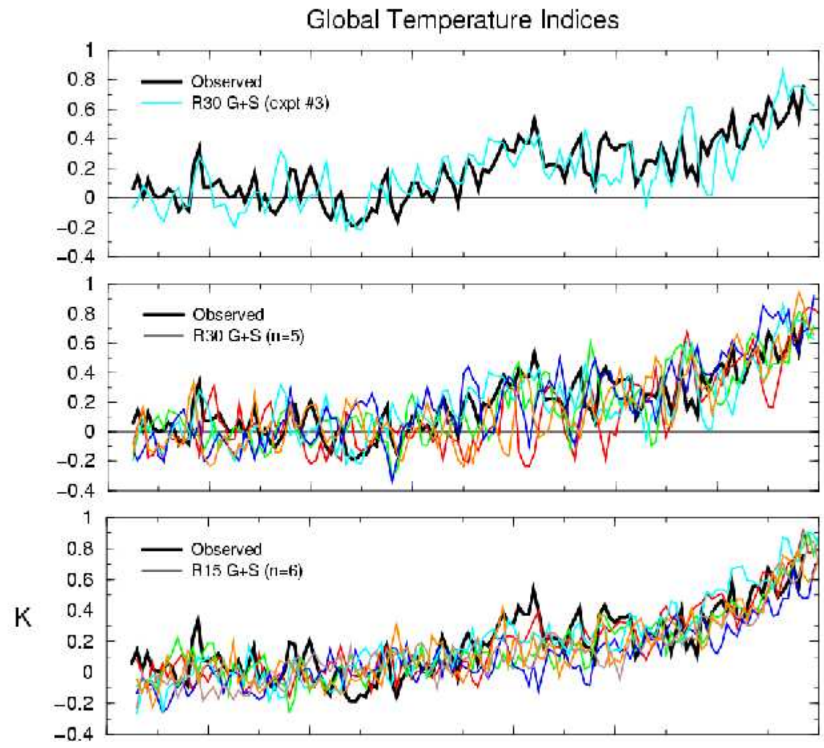
# Late 80s: Cray Y-MP

- Model: R15 atmosphere coupled to MOM ocean.
- Speed: 16 years/day.
- Runlength: 1000 years.
- Key results: stabilization has very long time scales.



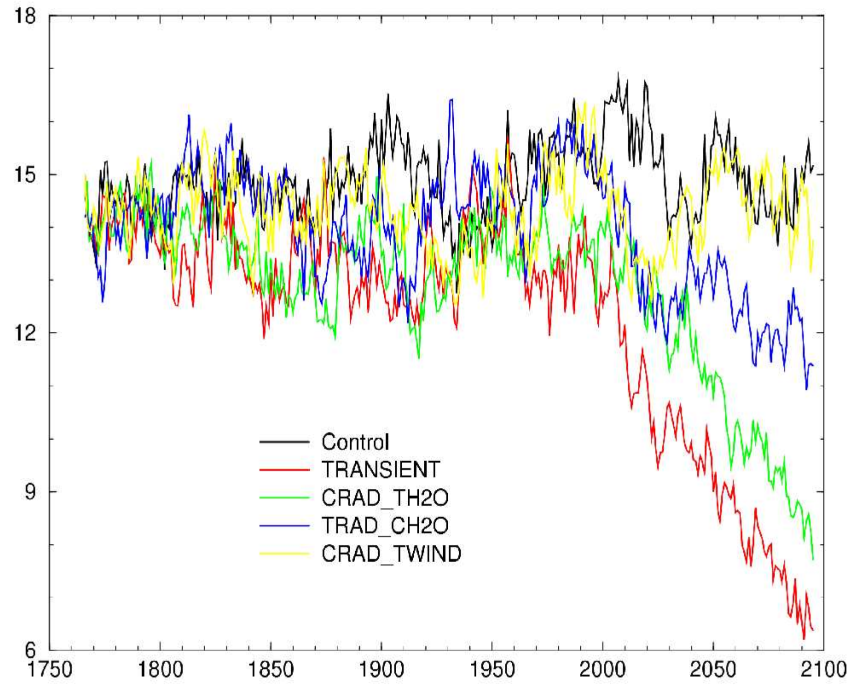
# Early 90s: Cray C90

- Model: R30 atmosphere coupled to MOM ocean.
- Speed: 3 years/day.
- Runlength: 100 years.
- Key results: detection and attribution of climate change.



# Late 90s: Cray T90

- Model: R30 atmosphere coupled to MOM ocean.
- Speed: 6 years/day.
- Runlength: 1000 years.
- Key results: water fluxes are main cause of THC weakening.



# This decade: SGI Origin and Altix

- Model: CM2.0 and CM2.1 models: 2°atmosphere, 1°ocean.
- Speed: 6 years/day.
- Runlength: several thousand years.
- Key results: attribution of regional climate change.

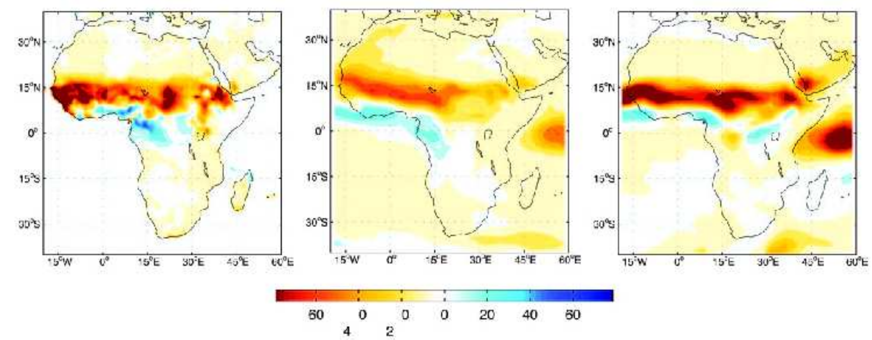
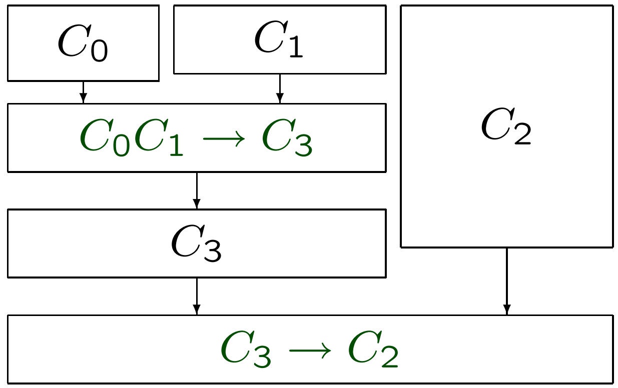


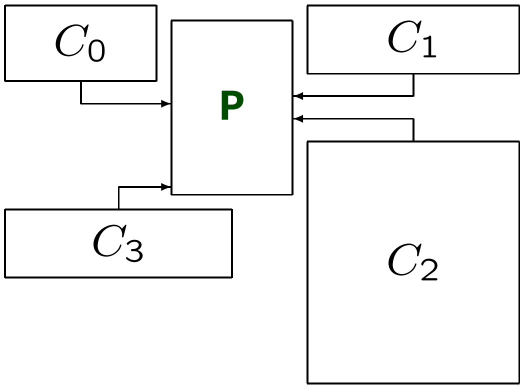
**Fig. 2.** Observed and modeled rainfall trends. (*Left*) The linear trend from 1950 to 2000 in the observed (CRU) July–August–September rainfall over land, in mm/month per 50 years. Blue areas correspond to a trend toward wetter conditions, and brown areas toward a drier climate. (*Center*) The linear trend for the eight-member ensemble mean of CM2 but plotted over both land and ocean. (*Right*) Linear trend for an ensemble mean of 10 simulations with the atmospheric/land component of CM2.0 running over observed sea surface boundary conditions.

# Recent changes in methodology

- Future projections of climate are performed at many sites, and a key goal of current research is to reduce the uncertainty of these projections by understanding the differences in the output from different models. This ***comparative study of climate simulations*** (e.g IPCC, ENSEMBLES, APE) across many models has spawned efforts to build uniform access to output datasets from major climate models, as well as modeling frameworks that will promote uniform access to the models themselves.

- As hardware and software complexity increase, we seek to encapsulate scalable data-sharing layers within an ***infrastructure***. Components of the physical climate system are now also code components, with coupling embedded in a standardized ***superstructure***. This has led to the emergence of Earth system modeling ***frameworks***, of which ESMF and PRISM are leading examples.

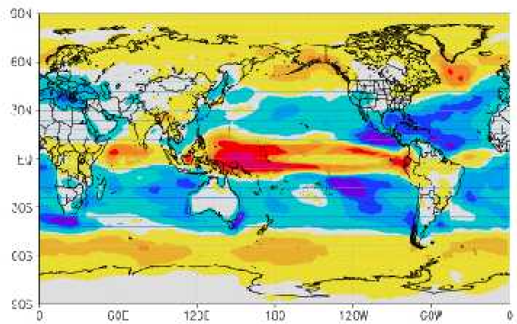


# The IPCC AR4 archive at PCMDI

The IPCC data archive at PCMDI is a truly remarkable resource for the comparative study of models. Since it came online in early 2005, it has been a resource for ∼300 scientific papers aimed at providing consensus and uncertainty estimates of climate change, from ∼20 state-of-the-art climate models worldwide.

| Model | Modeling Center |
|---|---|
| BCCR BCM2 | Bjerknes Centre for Climate Research |
| CCCMA CGCM3 | Canadian Centre for Climate Modeling & Analysis |
| CNRM CM3 | Centre National de Recherches Meteorologiques |
| CSIRO MK3 | CSIRO Atmospheric Research |
| GFDL CM2_0 | Geophysical Fluid Dynamics Laboratory |
| GFDL CM2_1 | Geophysical Fluid Dynamics Laboratory |
| GISS AOM | Goddard Institute for Space Studies |
| GISS EH | Goddard Institute for Space Studies |
| GISS ER | Goddard Institute for Space Studies |
| IAP FGOALS1 | Institute for Atmospheric Physics |
| INM CM3 | Institute for Numerical Mathematics |
| IPSL CM4 | Institut Pierre Simon Laplace |
| MIROC HIRES | Center for Climate System Research |
| MIROC MEDRES | Center for Climate System Research |
| MIUB ECHO | Meteorological Institute University of Bonn |
| MPI ECHAM5 | Max Planck Institute for Meteorology |
| MRI CGCM2 | Meteorological Research Institute |
| NCAR CCSM3 | National Center for Atmospheric Research |
| NCAR PCM1 | National Center for Atmospheric Research |
| UKMO HADCM3 | Hadley Centre for Climate Prediction |



This figure, from Held and Soden (2005), is a composite analysis across the entire IPCC archive.

Computational load at GFDL:

- 5500 model years run.
- Occupied half of available compute cycles at GFDL for half a year (roughly equivalent to 1000 Altix processors).
- 200 Tb internal archive; 40 Tb archived at GFDL data portal; 4 Tb archived at PCMDI data portal.

I would argue that the IPCC experiment is ***already*** petascale!

# The routine use of Earth System models in research and operations

Let's declare that 2000-2010 (the "noughties") is the decade of the coming-of-age of Earth system models.

**Operational forecasting** model-based seasonal and inter-annual forecasts delivered to the public;

**Decision support** models routinely run for decision support on climate policy by governments, for energy strategy by industry and government, as input to pricing models by the insurance industry, etc.

**Fundamental research** the use of models to develop a predictive understanding of the earth system and to provide a sound underpinning for all applications above.

This requires a radical shift in the way we do modeling: from the current dependence on a nucleus of very specialized researchers to make it a more accessible general purpose toolkit. This requires ***an infrastructure for moving the building, running and analysis of models and model output data from the "heroic" mode to the routine mode***.

# From heroic to routine in other fields

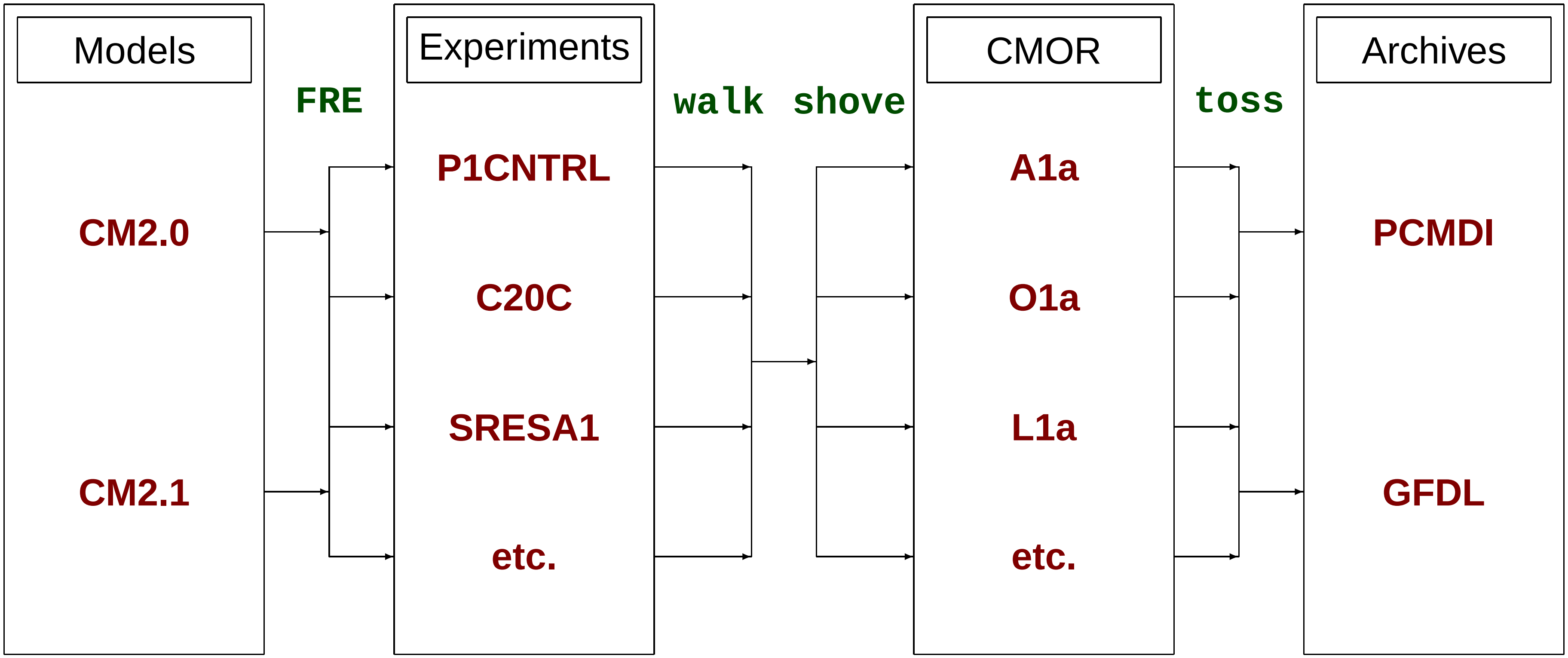
The ***polymerase chain reaction*** was awarded a Nobel prize not long ago. Later, you could get a PhD for developing PCR in different contexts. Now you order online and receive samples through the mail...

***Transgenic implants*** in different organisms are another example... below, you see a service provided by a lab at Princeton University which will develop and store transgenic mice and other organisms.



What will the transition from heroic to routine look like in our field?

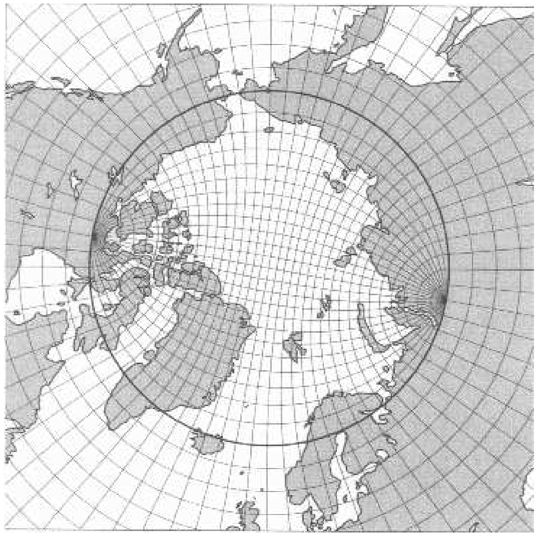
# The IPCC data pipeline at GFDL

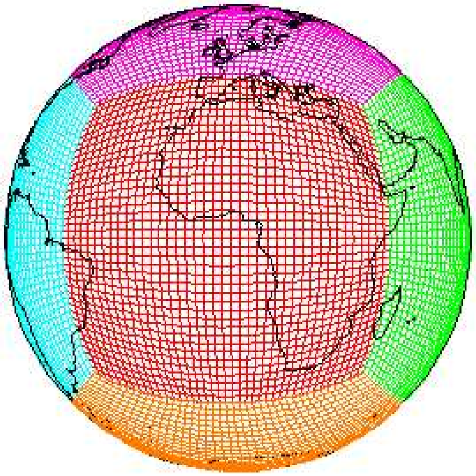



The process was time- and data-intensive, with multiple access episodes for the same datasets. Clearly it would be ideal if FRE already produced compliant data.

# Current problems with CMOR-compliant data

- A principal difficulty is CMOR's restricted view of model grids: only simple latitude-longitude grids are permitted. This is because the current crop of visualization and analysis tools cannot easily translate data among different grids. Shown at right are the ***tripolar grid*** (Murray 1996, Griffies et al 2004) used by MOM4 for GFDL's current IPCC model CM2. Below is the ***cubed sphere*** (Rancic and Purser 1990) planned for the Finite-Volume atmosphere dynamical core for the next-generation GFDL models AM3 and CM3. If there were a ***grid metadata standard***, regridding operations could potentially be applied by the end-user using standard-compliant tools.
- The model descriptions demanded by CMOR do not contain enough information about the models, and are added after the fact. If there were a ***model metadata standard*** such as NMM in force, comprehensive model descriptions could be automatically produced. The end-user could better diagnose specific differences between different models in an archive.

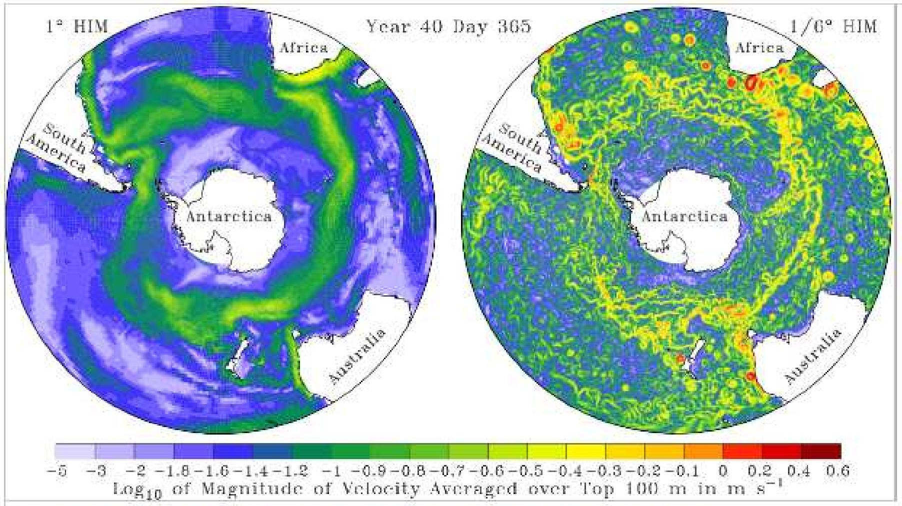




# Can an experiment like IPCC be run at higher resolution?

Possible key challenges for the next IPCC:

- Robust estimates of regional climate change.
- Interactive carbon dynamics: inclusion of land-use change, ocean carbon uptake, marine and terrestrial biospheric response to climate change.
- Increased resolution in the atmosphere (even before we get to cloud-resolving scales) will lead to better characterization of storm track changes and hurricane intensity projections in a changed climate. Target: 1°or 0.5°model for IPCC AR5.
- Increased resolution in the ocean is even more critical: key mechanisms of ocean mass and energy transport are currently unresolved. Targets: 0.25°("eddy-permitting") models next time around, 0.0825°("eddy-resolving") still out of reach.



Shaded is the speed of the surface currents in the ocean model at two resolutions.

# Petascale methodologies

As much emphasis must be placed on methodologies to facilitate scientific analysis of multi-model ensembles on distributed data archives, as on the computational technology itself.

Some current efforts:

**ESC** Earth System Curator, funded by NSF. Partners GFDL, NCAR, PCMDI, Georgia Tech. Will be used to promote the existence of a model and grid metadata standard, and build a prototype relational database housing these metadata. Will build tools for model configuration and compatibility checking based on automatic harvesting of metadata from code.

**MAPS** Modeling, Analysis and Prediction System? funded by NASA, partners NASA/GSFC, GFDL, MIT. Proposes to build a configuration layer for a subset of coupled models based on PRISM config files, and conformant with grid and metadata standards. Will attempt to promote a “standard coupling architecture” and develop a standard for exchange grids for ESMF.

**GO-ESSP and CF** should be the medium of exchange for standard-building. CF is seeking funding and WGCM backing to become a mandated activity. GO-ESSP is the ideal medium for the actual technical work of standard-building.

**IPCC!** PCMDI and other data centres should be core participants.

With a complete metadata hierarchy defined, one can envisage the convergence of modeling and data frameworks into a single environment: a model ***curator***.

# Scenario 1: dynamically generated data catalogues

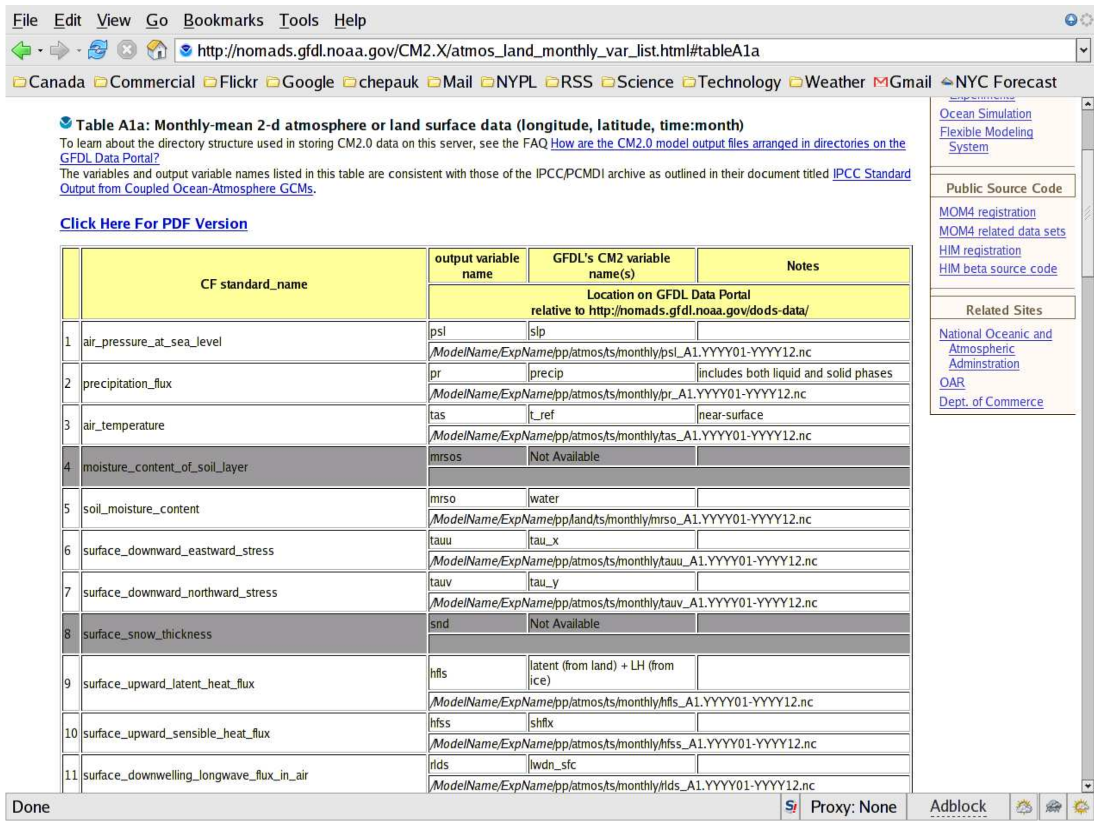



**Table A1a: Monthly-mean 2-d atmosphere or land surface data (longitude, latitude, time:month)**

To learn about the directory structure used in storing CM2.0 data on this server, see the FAQ How are the CM2.0 model output files arranged in directories on the GFDL Data Portal?

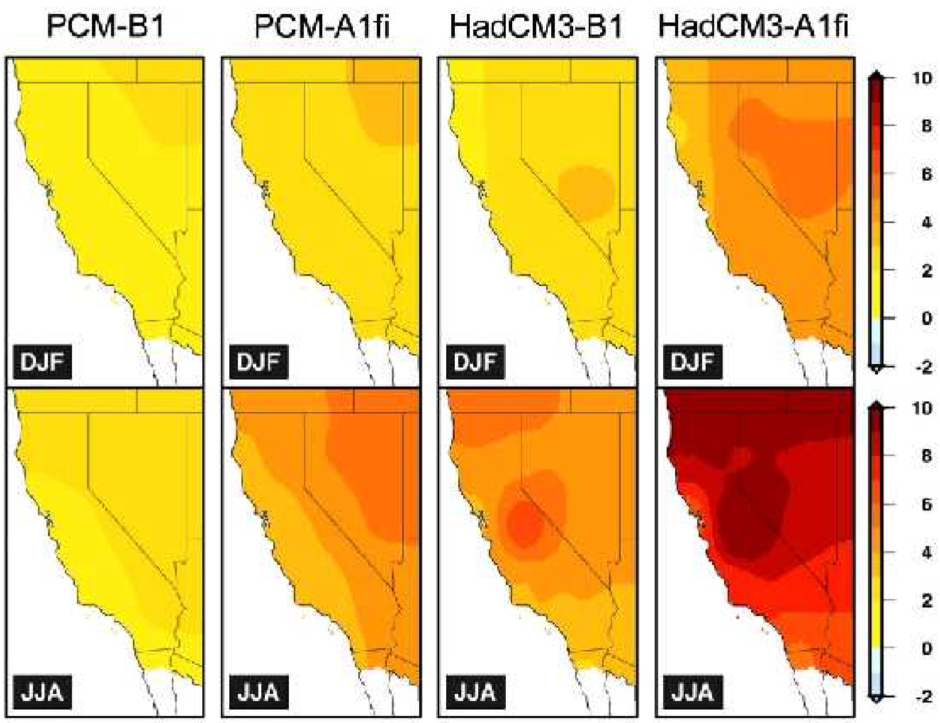
The variables and output variable names listed in this table are consistent with those of the IPCC/PCMDI archive as outlined in their document titled IPCC Standard Output from Coupled Ocean-Atmosphere GCMs.

**Click Here For PDF Version**

| | CF standard_name | output variable name | GFDL's CM2 variable name(s) | Notes | Location on GFDL Data Portal relative to http://nomads.gfdl.noaa.gov/dods-data/ |
|---|---|---|---|---|---|
| 1 | air_pressure_at_sea_level | psl | slp | | /ModelName/ExpName/pp/atmos/ts/monthly/psl_A1.YYYY01-YYYY12.nc |
| 2 | precipitation_flux | pr | precip | includes both liquid and solid phases | /ModelName/ExpName/pp/atmos/ts/monthly/pr_A1.YYYY01-YYYY12.nc |
| 3 | air_temperature | tas | t_ref | near-surface | /ModelName/ExpName/pp/atmos/ts/monthly/tas_A1.YYYY01-YYYY12.nc |
| 4 | moisture_content_of_soil_layer | mrsos | Not Available | | |
| 5 | soil_moisture_content | mrso | water | | /ModelName/ExpName/pp/land/ts/monthly/mrso_A1.YYYY01-YYYY12.nc |
| 6 | surface_downward_eastward_stress | tauu | tau_x | | /ModelName/ExpName/pp/atmos/ts/monthly/tauu_A1.YYYY01-YYYY12.nc |
| 7 | surface_downward_northward_stress | tauv | tau_y | | /ModelName/ExpName/pp/atmos/ts/monthly/tauv_A1.YYYY01-YYYY12.nc |
| 8 | surface_snow_thickness | snd | Not Available | | |
| 9 | surface_upward_latent_heat_flux | hfls | latent (from land) + LH (from ice) | | /ModelName/ExpName/pp/atmos/ts/monthly/hfls_A1.YYYY01-YYYY12.nc |
| 10 | surface_upward_sensible_heat_flux | hfss | shflx | | /ModelName/ExpName/pp/atmos/ts/monthly/hfss_A1.YYYY01-YYYY12.nc |
| 11 | surface_downwelling_longwave_flux_in_air | rlds | lwdn_sfc | | /ModelName/ExpName/pp/atmos/ts/monthly/rlds_A1.YYYY01-YYYY12.nc |

Already in use at PCMDI, DDC, GFDL Curator, elsewhere: metadata requires extension.

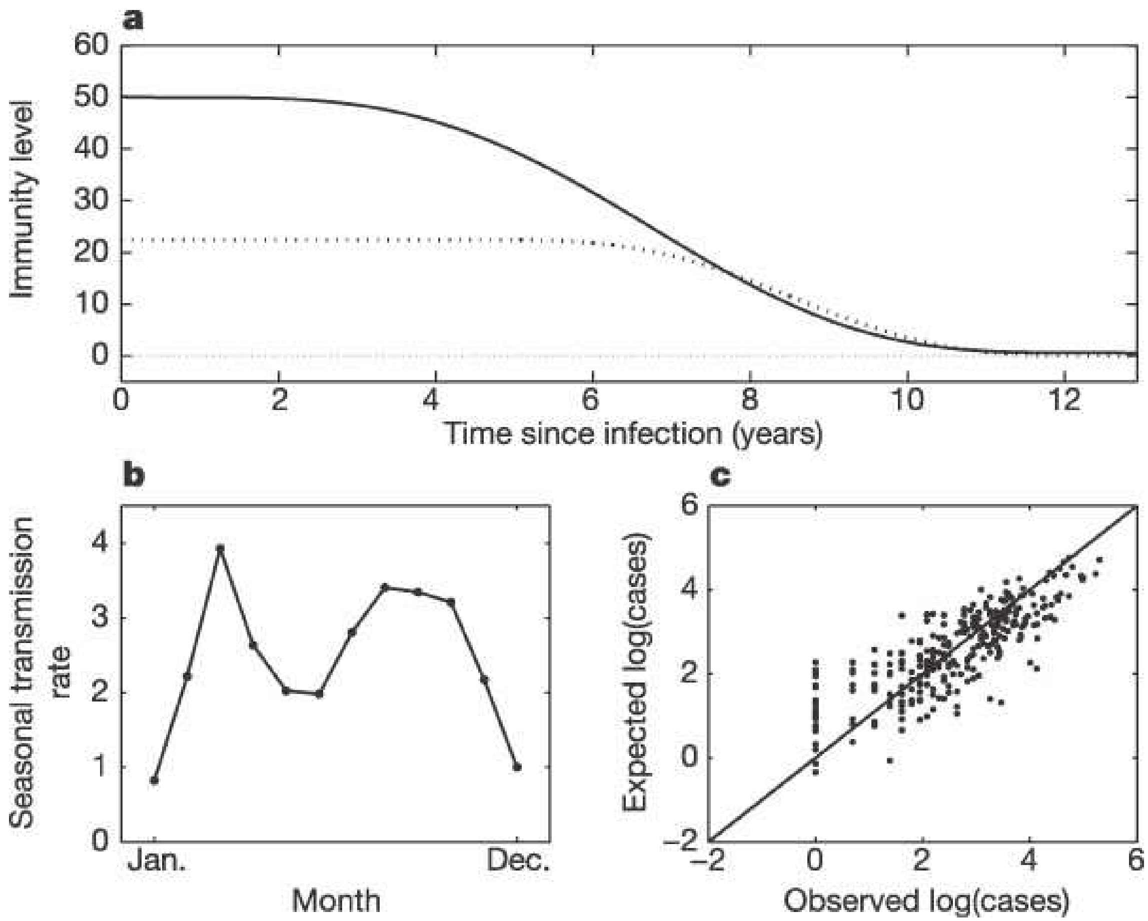
# Scenario 2: statistical downscaling of climate change projections




Hayhoe et al, ***PNAS***, 2004: ***Emissions pathways, climate change, and impacts on California.***

Uses daily data for “heat degree days” and other derived quantities. Requires data beyond that provided by IPCC AR4 SOPs (1960-2000).

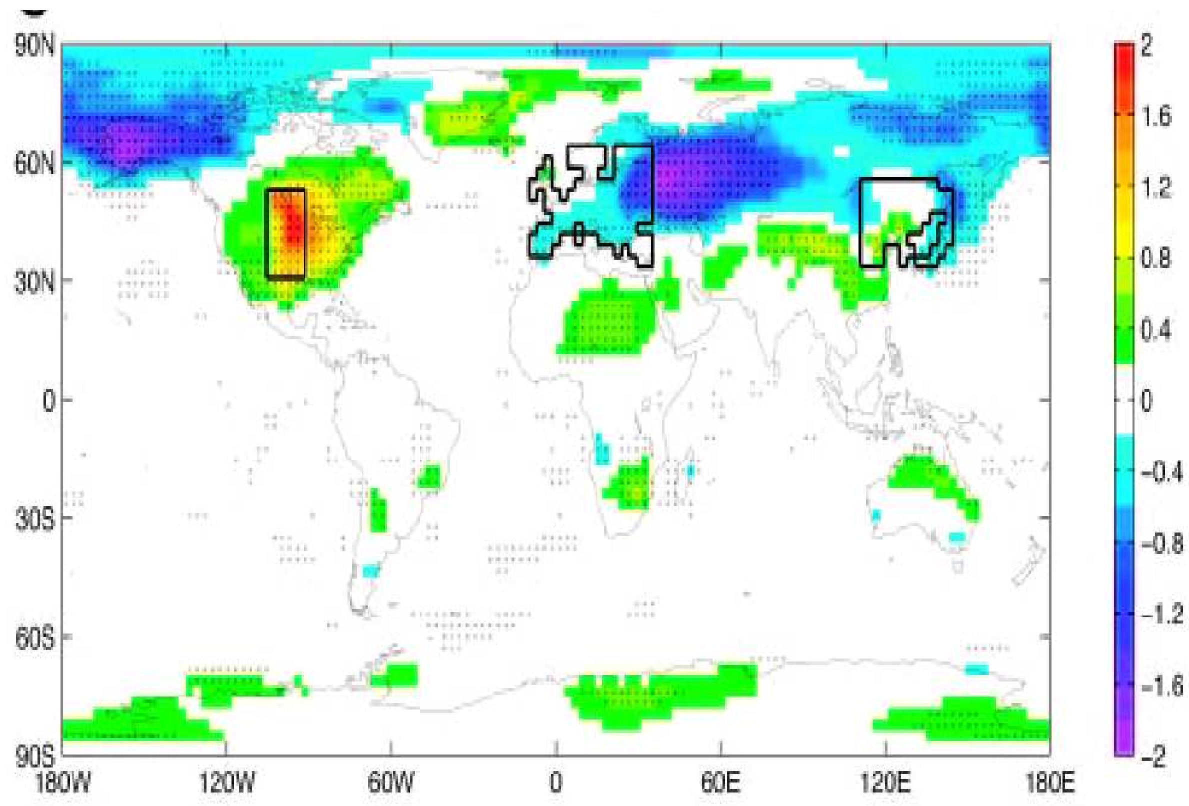
# Scenario 3: disease vectors in a changing climate




Koelle et al, ***Nature***, 2005: ***Refractory periods and climate forcing in cholera dynamics.***

Requires monthly forcing data, no feedback.

# Scenario 4: alternate energy sources




Keith et al, ***PNAS***, 2005: ***The influence of large-scale wind power on global climate.***

Feedback on atmospheric timescales: but does not require model to be retuned.

# Taking stock halfway through the noughties

- Earth system models are evolving into powerful tools for advancing our understanding, and well on their way to being operational tools in support of policy and industrial strategy.

- The principal research path for consensus and uncertainty estimates of climate change is the comparative study of models.

- The building of appropriate standards has been identified as a key element in uniting modeling and data communities.

- This requires convergence and cross-fertilization between model and data frameworks: by developing a clear understanding of the architecture of Earth system models, PRISM and ESMF also point the way to a metadata hierarchy to be used in building curators.

- Leadership in standards will come from custodians of international multi-model data archives well connected to data consumers, and will be embedded in the modeling frameworks.

- Research is needed into hierarchical data storage, use of pattern recognition and feature detection for data reduction, remote data analysis and visualization.