

Convergence of model frameworks and data frameworks

V. Balaji

Princeton University and NOAA/GFDL

GO-ESSP Workshop

Lawrence Livermore National Laboratory

Livermore, CA

19 June 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.

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.

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.

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

PERSONAL COMPUTERS

IMPLICIT COMPUTING

INTERNET

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>

GFDL models in IPCC AR4

- 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.

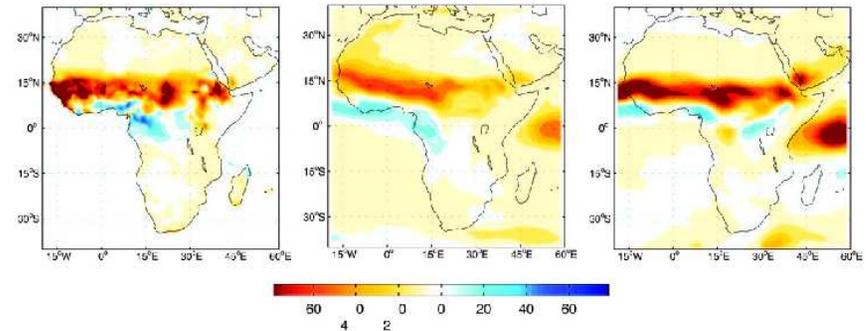
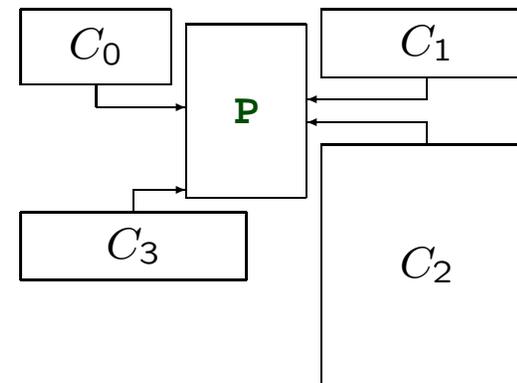
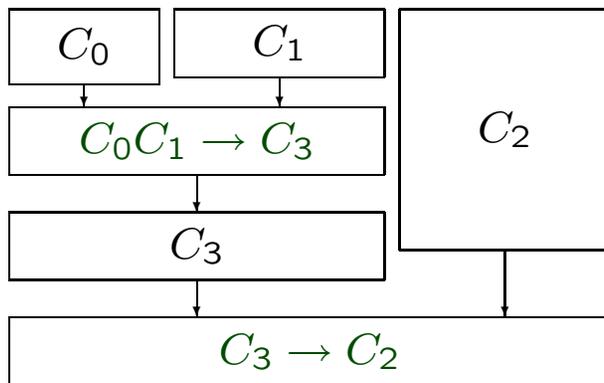


Fig. 2. Observed and modeled rainfall trends. (Left) The linear trend from 1950 to 2000 in the observed (CRU) July–August–September rainfall over land, in mm/month per 50 years. Blue areas correspond to a trend toward wetter conditions, and brown areas toward a drier climate. (Center) The linear trend for the eight-member ensemble mean of CM2.0 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.

Current methodologies in climate research

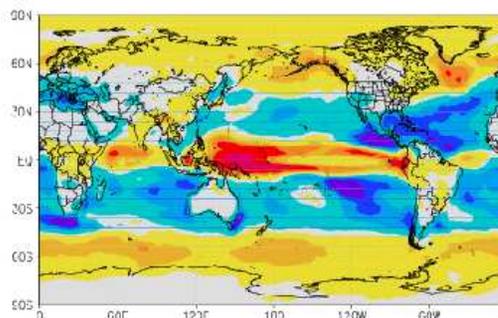
- Future projections of climate are performed at many sites, and a key goal of current research is to reduce the uncertainty of these projections by understanding the differences in the output from different models. This **comparative study of climate simulations** (e.g IPCC, ENSEMBLES, APE) across many 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.

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.



Home	Cryopreservation	
Transgenic Mouse Production		<p>Investigators will provide 5 to 10 fertile males for use in generating embryos to be frozen. The facility will freeze 500 embryos if the males are heterozygous and 300 embryos if the males are homozygous. The embryos will be stored by the facility in liquid nitrogen until requested by the investigator.</p>
Rederivation Service		
Knockout Mouse Production		
Cryopreservation		
Services and Fees		

© 2005 Princeton University Webmaster



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

Good engineering is inaudible

Stages in the development of a modeling experiment:

Conception posing of a scientific question; design of an experiment. An experiment may involve multiple components, with expertise on each component distributed among research teams and institutions. Many current experiments are based on reducing uncertainty by comparative study of diverse models.

Composition Assembly, configuration, and linkage of components into a suitable model.

Orchestration Projection of a model onto available resources; optimization for complex computing architectures; control and scheduling of model runs; archival of output data.

Appreciation Provision of easily ingested model output; analysis tools that “understand” to some degree the meaning of data, and are able to diagnose relationships between output from diverse models.

A current criticism is that the process for moving from “conception” to “appreciation” of an Earth system model involves too many layers of audible engineering to be negotiated by the “audience”.

Routine instead of heroic use of Earth system models requires the engineering to dwindle into the background.

Hence FMS, PRISM and ESMF...

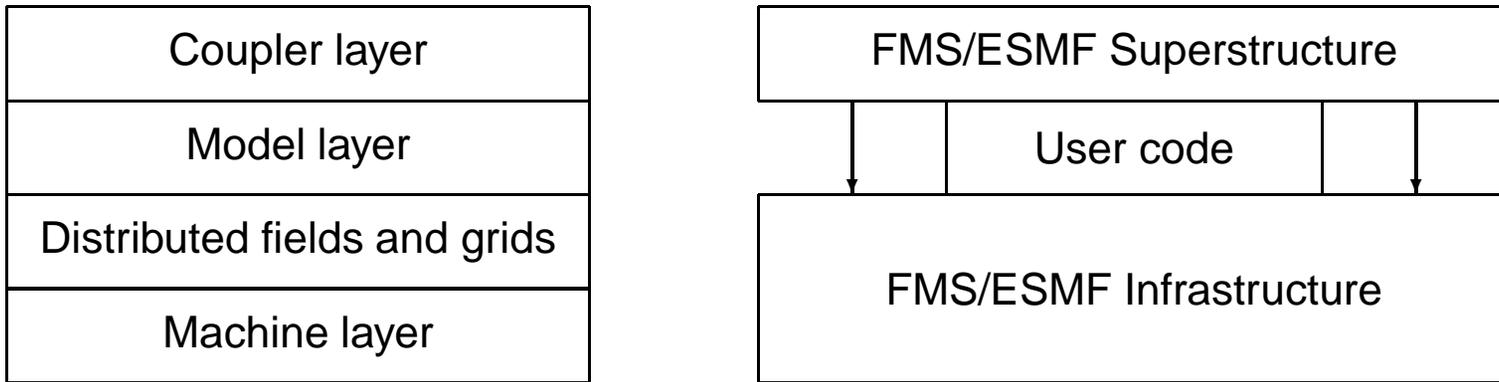
The development of modeling frameworks came in part from the realization that the engineering involved in climate research was “too loud”.

FMS The GFDL Flexible Modeling System (1998). Motivated by the arrival of massively-parallel computing. GFDL then maintained a stable of separate models for climate change research; interannual predictability and ENSO studies; hurricane forecasting and cloud system modeling. Now unified into a number of dynamical cores and physics options within a single framework for running solo and coupled models on parallel hardware.

PRISM Program for Integrated Earth System Modeling (2001). Motivated by the emergence of multi-model ensembles as a research avenue. The goal was a coupler layer with an easy transition path for existing models.

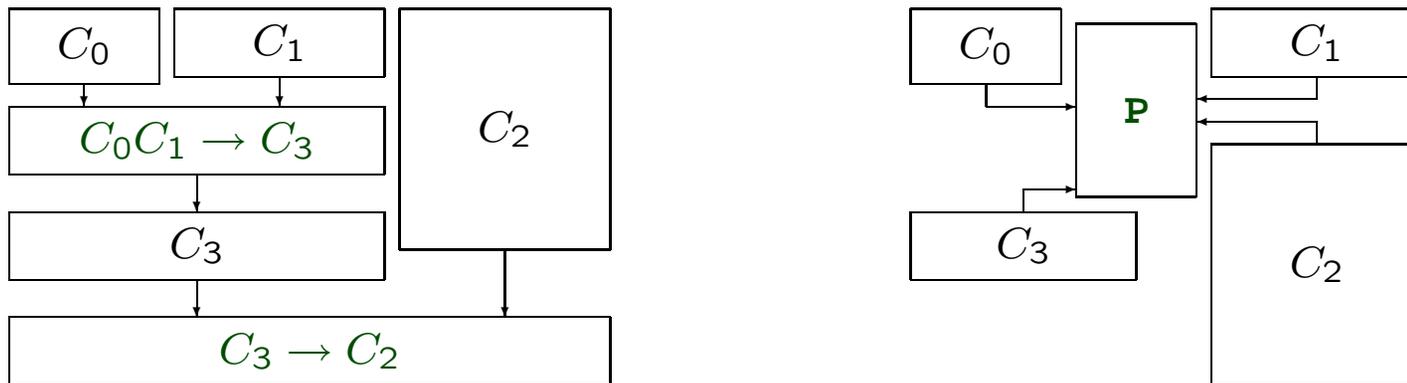
ESMF Earth System Modeling Framework (2002). Motivated by a perception that lack of coordination between major modeling centres was an obstacle to progress. The scope of features provided by ESMF is similar to FMS, but on a much larger scale, encompassing all the major modeling centres. An ultimate goal is “operational climate services”.

FMS/ESMF Overview



- Infrastructure provides simple interface to parallel communication and I/O, captured in a datatypes called **fields** and **grids**.
- FMS superstructure provides a “standard” coupled climate model architecture with implicit coupling between atmosphere, land and ocean surface on independent grids, with an intermediate surface boundary layer component running on an **exchange grid**. Provides serial and concurrent scheduling of components within a single executable. ESMF provides a more general superstructure where **components** exchange data through import and export **states** mediated by **couplers**.

Comparison of coupling models



- In FMS and ESMF, after each independent component run segment, control is returned to the coupler, which runs on the union of all PEs of its child components.
- PRISM uses a client-server model where all components execute concurrently, and the coupler P processes their **PRISM_Put** and **PRISM_Get** requests. Configuration of the coupler is through external files (SMIOC/SCC).

Operational use of model frameworks

The next stage in the evolution of frameworks was the addition of a *runtime environment*.

- Source code maintenance.
- Model configuration, launching and regression testing encapsulated in XML descriptors;
- Relational database for archived model results;
- Standard and custom diagnostic suites;

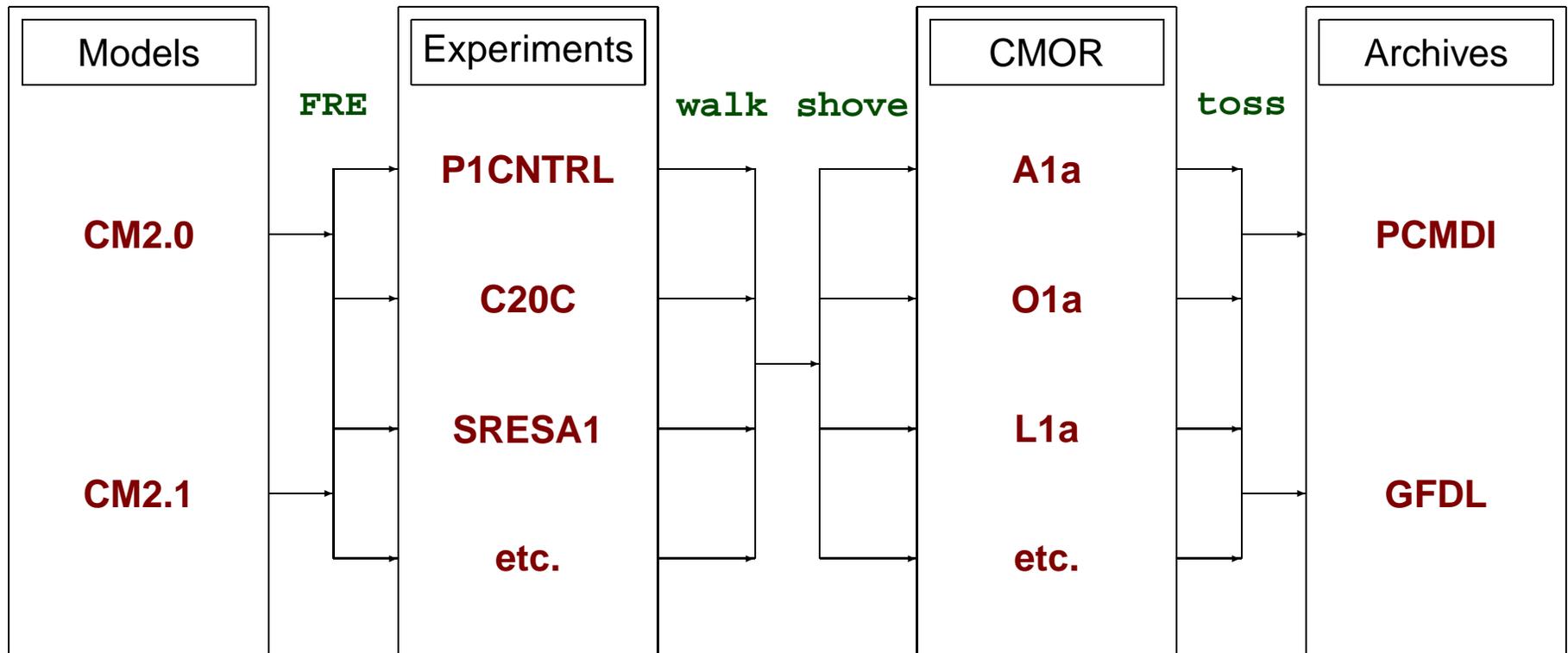
The FMS Runtime Environment (FRE) describes all the steps for configuring and running a model jobstream; archiving, postprocessing and analysis of model results.

fremake, frerun, frepp, frecheck, ...

The Regression Test Suite (RTS) is a set of tests that are run continuously on a set of FMS models to maintain and verify code integrity.

FRE was successfully used at GFDL for the development of climate models targeted for IPCC (CM2.0 and CM2.1) and management of GFDL's IPCC data.

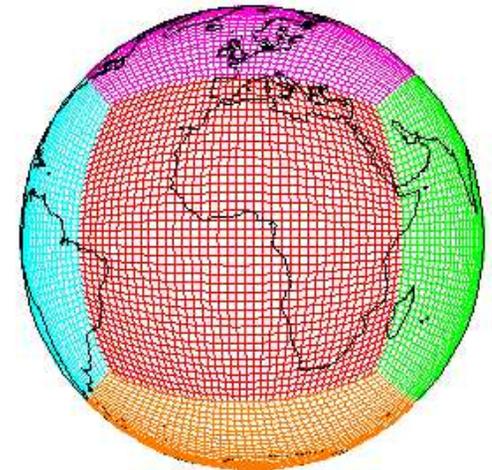
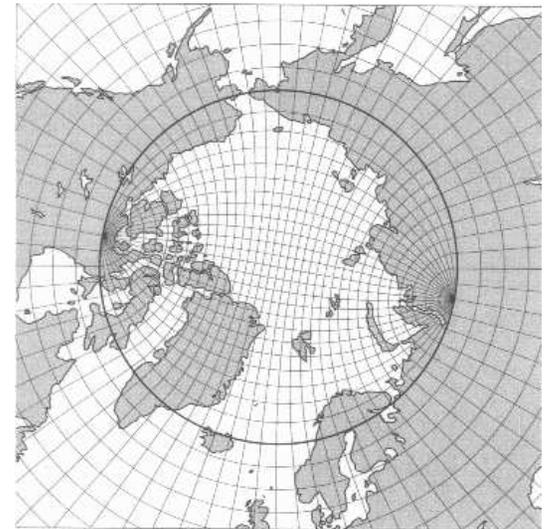
The IPCC data pipeline at GFDL



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, above, are the **tripolar grid** (Murray 1996, Griffies et al 2004) used by MOM4 for GFDL's current IPCC model CM2; and below, 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**, re-gridding operations could potentially be applied by the end-user using standard-compliant tools. A grid metadata standard will be presented tomorrow.
- 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.



Uniting runtime environments with data portals

Runtime environments contain all the necessary information for configuring and running a model. Data portals contain catalogue information for describing the contents of a dataset.

Convergence comes with the crucial insight that the descriptors used for comprehensively specifying a model configuration are needed for a scientifically useful description of the model output data as well. Thus ***the same attributes may be used to specify a model as well as the model output dataset***. thus leading to a ***convergence of models and data***.

Standard-building 2006-2008

Physical fields: standard vocabulary for describing the relevant physical quantities (viz. CF `standard_name`). Variables can contain *gridded* or *point* (station, drifter) data.

Geospatial information: location information: latitude, longitude, elevation. This set of standards unites a much larger community (mobile phones, GIS), in which our community has begun to play a role through OGC. We can provide some useful extensions toward 3D and 4D data.

Grid structure: interrelations between grids, between points and grids. With this information available, it is perhaps possible to perform regridding and subsampling of data by user request, on the archive servers.

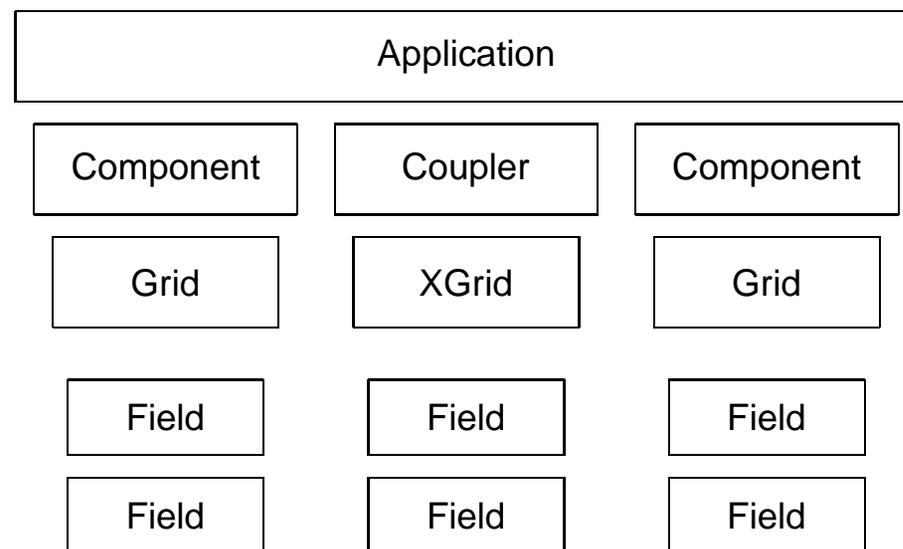
Model metadata: describing data source comprehensively, relatively easy for observations, harder for models but can asymptote toward completeness starting from current PCMDI standard. Two levels of model metadata: components and applications.

Metadata hierarchy

Application metadata experiment, scenario, institution, contact: currently covered by CF/CMOR.

Component metadata physical description of component and its input and configuration parameters. Currently covered by CMOR, but as free-form text.

Coupler metadata inventory of export and import fields, interpolation methods. Currently covered by OASIS4 XML, not exported to model output. Associated with an XGrid: unstructured grid for fractions and masks. May contain a physical component (e.g surface boundary layer).



Grid metadata geospatial information somewhat covered by CF, but bundled with fields; draft proposal for structural metadata in the works, being negotiated within PRISM, ESMF and GO-ESSP communities, will be proposed as a draft CF standard in 2006. To be presented tomorrow.

Field metadata covered by CF/CMOR standard variable name table. Many output fields do not (and should not) have standard names. In general, all metadata categories should allow both standard and bespoke elements.

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

Convergence of model and data frameworks

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. We

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, ENSEMBLES, ... PCMDI and other data centres should be core participants.

Model and data frameworks could converge into a single environment: a model *curator*. A curator contains a catalogue of model output data holdings from an institution; or from a multi-institution modeling campaign. A result from a query can be either a dataset or a model.

Scenario 1: dynamically generated data catalogues

File Edit View Go Bookmarks Tools Help

http://nomads.gfdl.noaa.gov/CM2.X/atmos_land_monthly_var_list.html#tableA1a

Canada Commercial Flickr Google chepauk Mail NYPL RSS Science Technology Weather Gmail NYC Forecast

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](#)

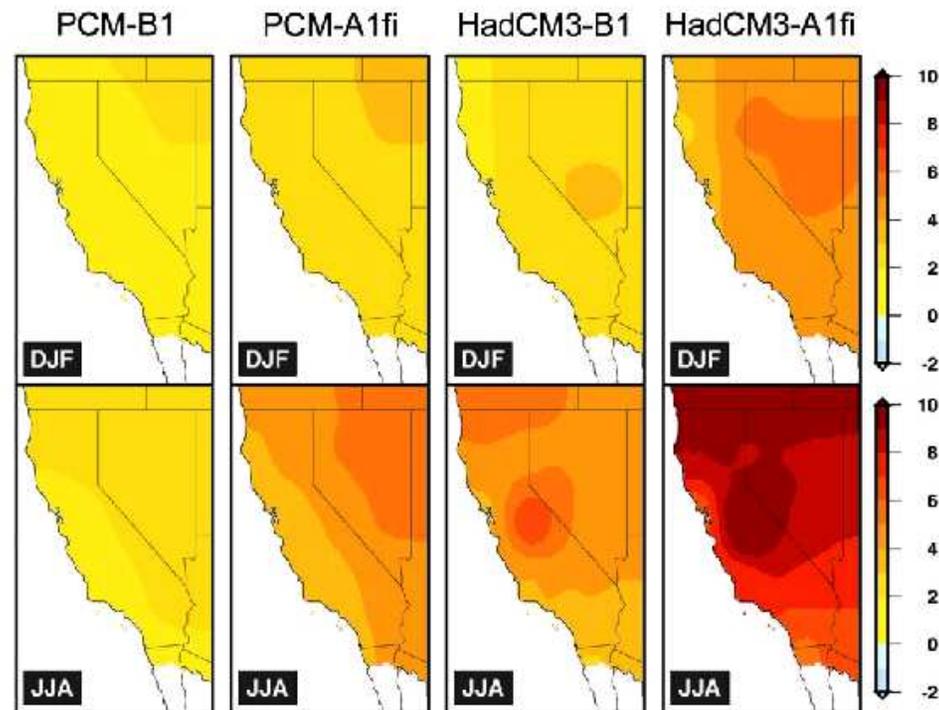
	CF standard_name	output variable name	GFDL's CM2 variable name(s)	Notes
Location on GFDL Data Portal relative to http://nomads.gfdl.noaa.gov/dods-data/				
1	air_pressure_at_sea_level	psl	slp	<i>/ModelName/ExpName/pp/atmos/ts/monthly/psl_A1.YYYY01-YYYY12.nc</i>
2	precipitation_flux	pr	precip	includes both liquid and solid phases <i>/ModelName/ExpName/pp/atmos/ts/monthly/pr_A1.YYYY01-YYYY12.nc</i>
3	air_temperature	tas	t_ref	near-surface <i>/ModelName/ExpName/pp/atmos/ts/monthly/tas_A1.YYYY01-YYYY12.nc</i>
4	moisture_content_of_soil_layer	mrso	Not Available	
5	soil_moisture_content	mrso	water	<i>/ModelName/ExpName/pp/land/ts/monthly/mrso_A1.YYYY01-YYYY12.nc</i>
6	surface_downward_eastward_stress	tauu	tau_x	<i>/ModelName/ExpName/pp/atmos/ts/monthly/tauu_A1.YYYY01-YYYY12.nc</i>
7	surface_downward_northward_stress	tauv	tau_y	<i>/ModelName/ExpName/pp/atmos/ts/monthly/tauv_A1.YYYY01-YYYY12.nc</i>
8	surface_snow_thickness	snd	Not Available	
9	surface_upward_latent_heat_flux	hfls	latent (from land) + LH (from ice)	<i>/ModelName/ExpName/pp/atmos/ts/monthly/hfls_A1.YYYY01-YYYY12.nc</i>
10	surface_upward_sensible_heat_flux	hfls	shflx	<i>/ModelName/ExpName/pp/atmos/ts/monthly/hfls_A1.YYYY01-YYYY12.nc</i>
11	surface_downwelling_longwave_flux_in_air	rlwds	lwdn_sfc	<i>/ModelName/ExpName/pp/atmos/ts/monthly/rlwds_A1.YYYY01-YYYY12.nc</i>

Done Proxy: None Adblock

[Ocean Simulation Flexible Modeling System](#)
[Public Source Code](#)
[MOM4 registration](#)
[MOM4 related data sets](#)
[HIM registration](#)
[HIM beta source code](#)
[Related Sites](#)
[National Oceanic and Atmospheric Administration](#)
[OAR](#)
[Dept. of Commerce](#)

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

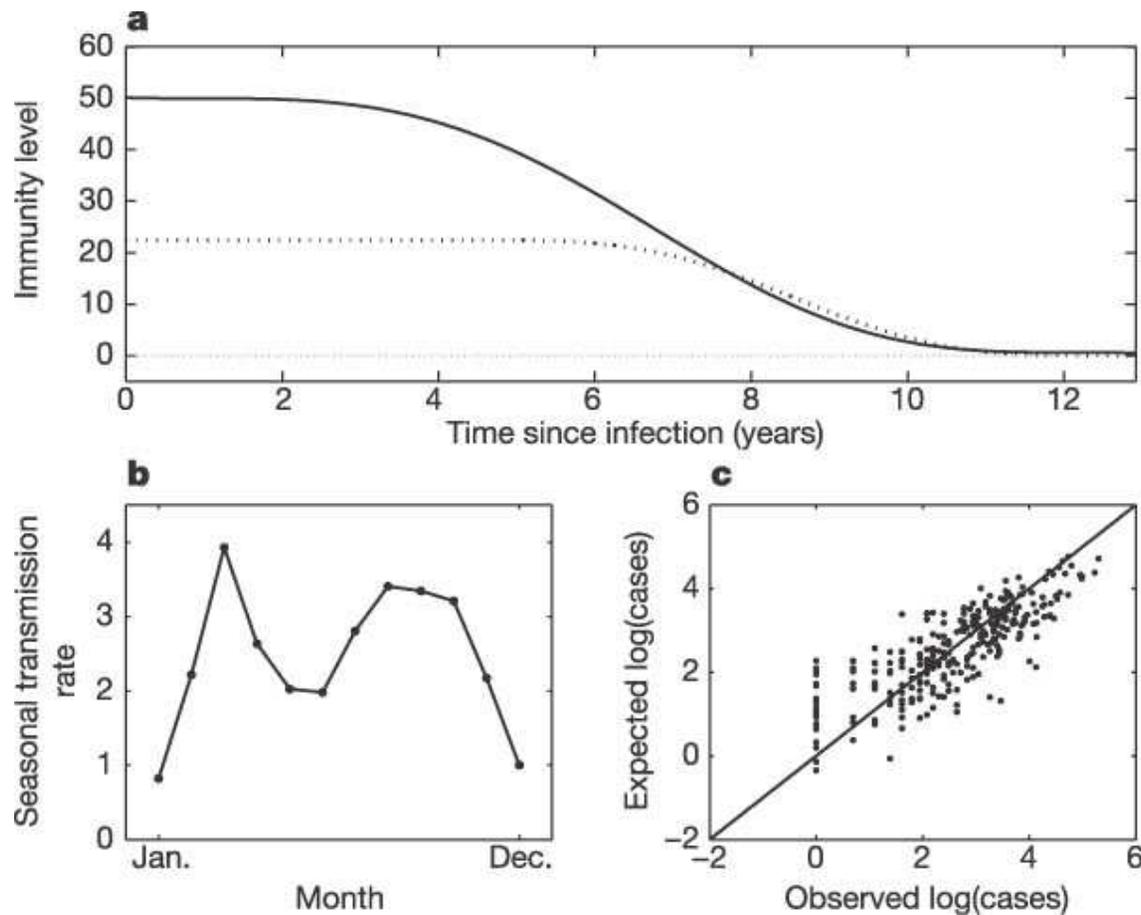
Scenario 2: statistical downscaling of climate change projections



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).

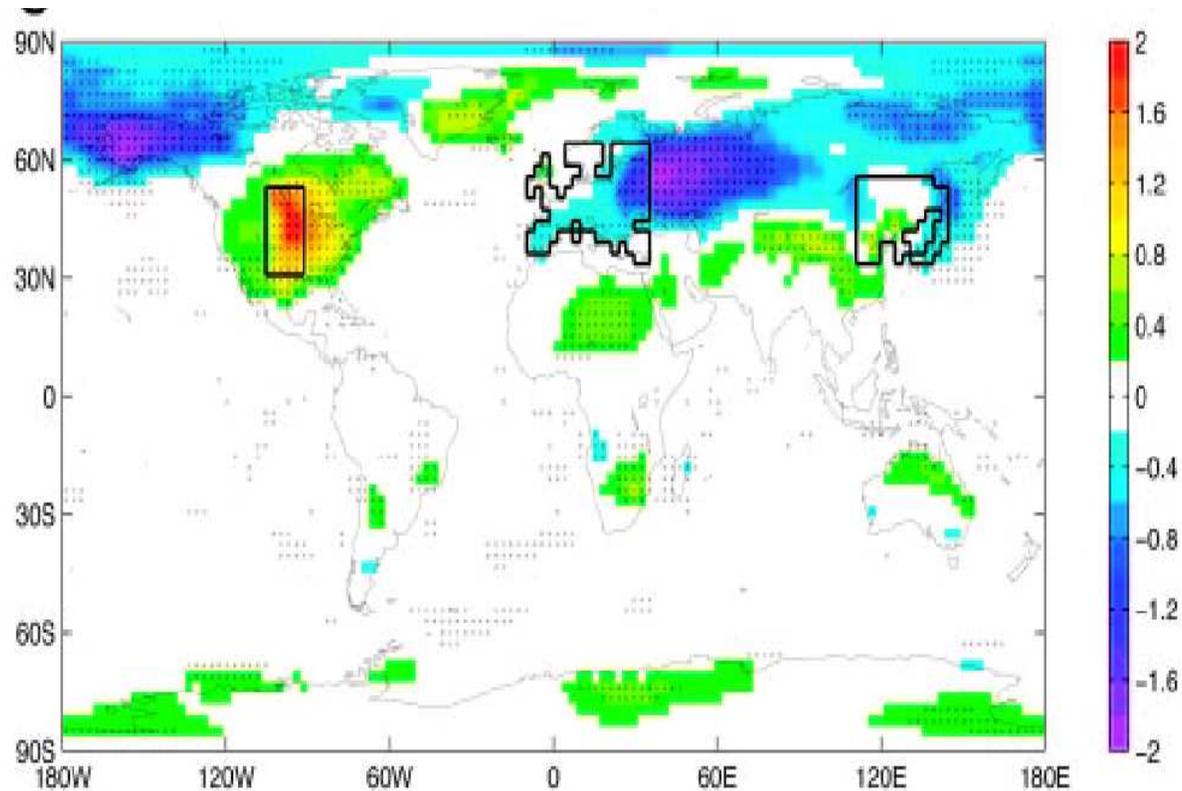
Scenario 3: disease vectors in a changing climate



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. PRISM and ESMF provide powerful substrates for facilitating this study.
- 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*. Curators are a new element in software enabling the convergence of model and data frameworks.
- Leadership in standards lies with custodians of international multi-model data archives well connected to data consumers, and will be embedded in the modeling frameworks.
- While building fully-featured systems, let's not neglect the low end... see e.g TGICA Data and Capacity Building Initiative for developing and transition economies (part of IPCC).