Convergence of model frameworks and data frameworks

V. Balaji
Princeton University and NOAA/GFDL

PRISM Community Meeting
CERFACS
Toulouse FRANCE
16 November 2005

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?

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...

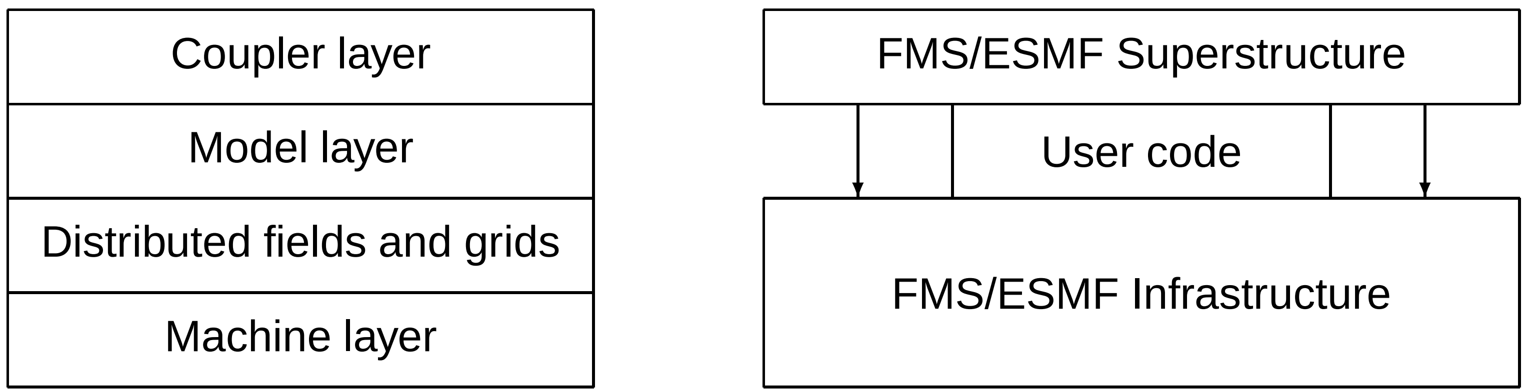
The development of modeling frameworks came in part from the realization that the engineering involved in climate research was “too loud”.

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”.

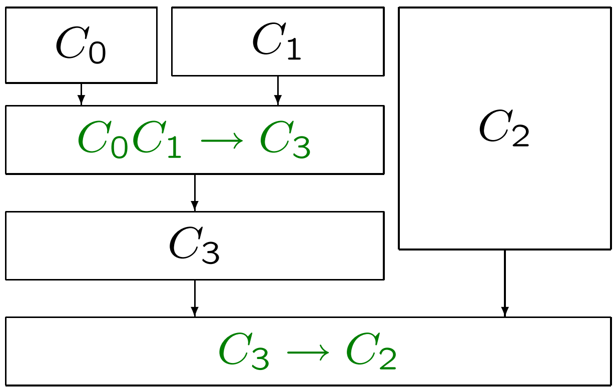
FMS/ESMF Overview

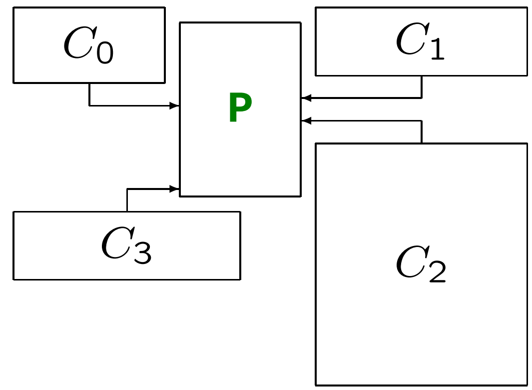


- Infrastructure provides simple interface to parallel communication and I/O, captured in a datatypes called **`fields`** and **`grids`**.

- 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.

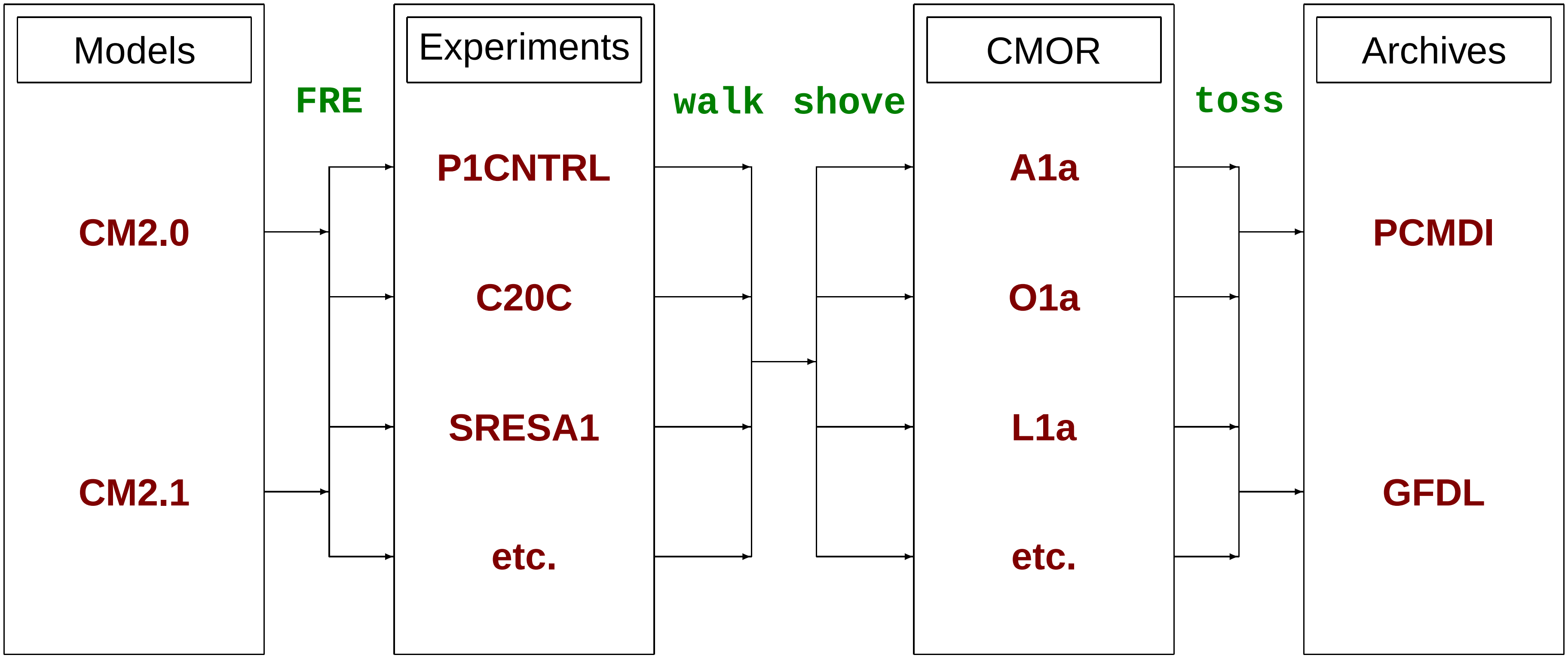
The IPCC data archive

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 $\sim$200 scientific papers aimed at providing consensus and uncertainty estimates of climate change, from $\sim$20 state-of-the-art climate models worldwide.

While the data archive is an unprecedented boon for researchers analysing the output (***consumers***), the process of generating data for the archive was quite onerous for the modeling centres (***producers***).

Needed: a food web for the data ecosystem!

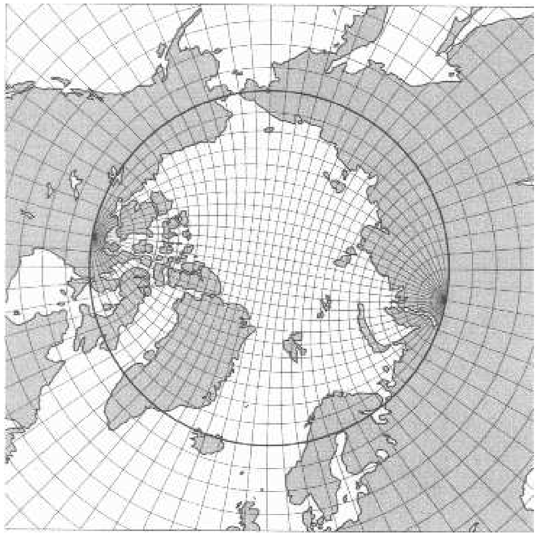
The IPCC data pipeline at GFDL

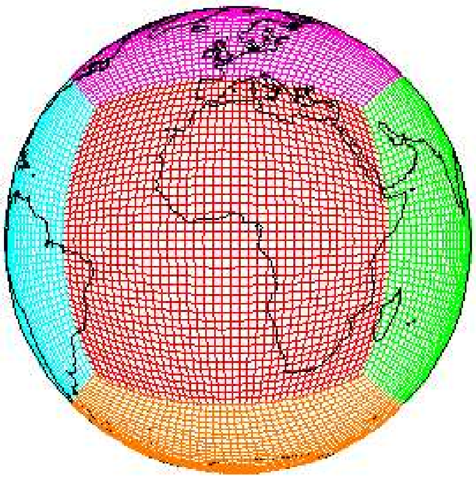


The process was time- and data-intensive, with multiple access episodes for the same datasets. Clearly it would be ideal if FRE already produced compliant data.

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. In the middle 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.





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***.

Organization of metadata

The hierarchy of model data structures provides useful insights on how to organize metadata. In the hierachies of modeling frameworks, the following terms are often used:

- object-oriented programming;
- component-based design;
- services (and on the data portal side, web services).

How do we distinguish between objects, components and services?

Roger Sessions, ***ACM Queue***, 2004: ***Fuzzy Boundaries: Objects, Components, and Web Services***.

Objects

- Objects are constructs of which you can create one or many instances within a ***process***. You interact with an object through its specified interfaces or methods.

- In parallel computing environments, we can generalize from ***process*** to ***communicator***: the object is now a ***distributed object***.

- Examples of objects are ***fields*** and ***grids***. Model data structures tell us that these are separate, linked objects. But current practice in creating datasets bundles grids along with fields. The problem is that two fields which share a grid (say, the u and v components of a vector field) may be in separate datasets, and there is no way of ensuring that they share a grid.

Components

- Components work similarly to objects, but in a more general computing ***environment***. The ***state*** of a component may be extracted and passed to other components by a ***coupler***.

- Components within an environment may be asynchronous.

- Examples of components may ESMF components, FMS component models, PRISM models, FLUME composable entities. Model metadata should identify components clearly: current CMOR metadata only requires model metadata for the entire application.

- Fuzzy boundaries: ensemble models may treat components as objects, spawning many instances.

- Are components ostensibly labelled “atmosphere”, say, sufficiently similar that a single physical interface may be defined? Or, to put it another way, to what extent to two such components share a **`state`**?

- Do different models see component granularity the same way? What incompatibilities are introduced if one model treats atmospheric chemistry say, as an indivisible entity within an atmosphere component, whereas another treats them as independent **`component`**s?

Web services

In a web service, there is no guarantee that the invoker of a method shares an environment with where the service provider resides. The data transferred cannot be simple data, but is wrapped in an agreed-upon common language and vocabulary: usually some dialect of XML.

In general, the resolution of the “fuzzy boundary” problem is based on performance, for example:

Can components be web services (i.e GRID computing)? In principle, why not? In practice, probably no: consider a typical coupled model integrating at the rate of 10 model years/day, exchanging data betwen components every 15 model minutes: this works out to an exchange every 250 msec. This cannot tolerate current Internet latencies ($\sim$100 msec).

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.

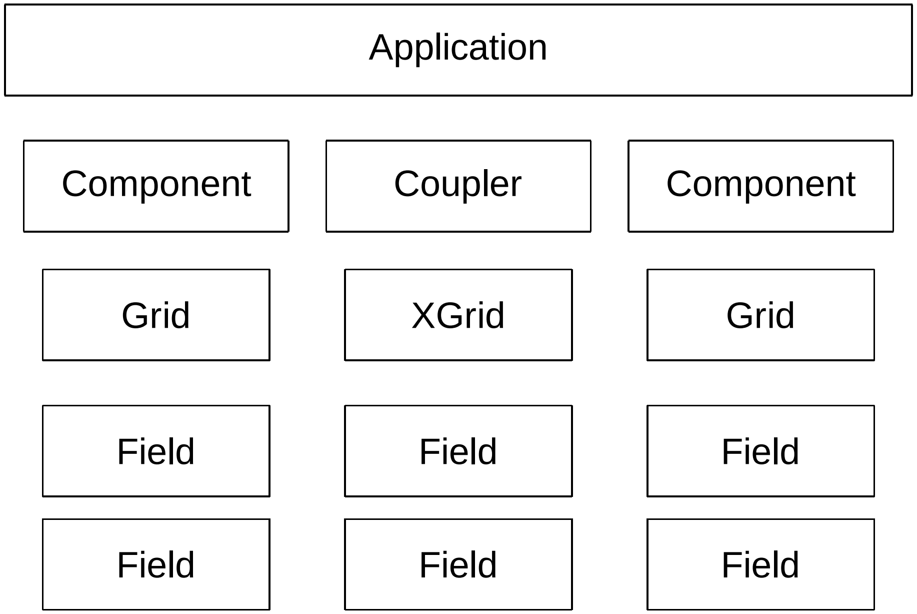
Geospatial information: location information: latitude, longitude, elevation. This set of standards unites a much larger community (mobile phones, GIS), in which our community has begun to play a role. We can provide some useful extensions toward 3D and 4D data.

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(?)

- 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***.

Scenario 1: dynamically generated data catalogues

http://www-pcmdi.llnl.gov/ipcc/data_status_tables_files/sheet002.htm...

Time-Independent Land Surface Data Availability **(as of 17 Au**

available

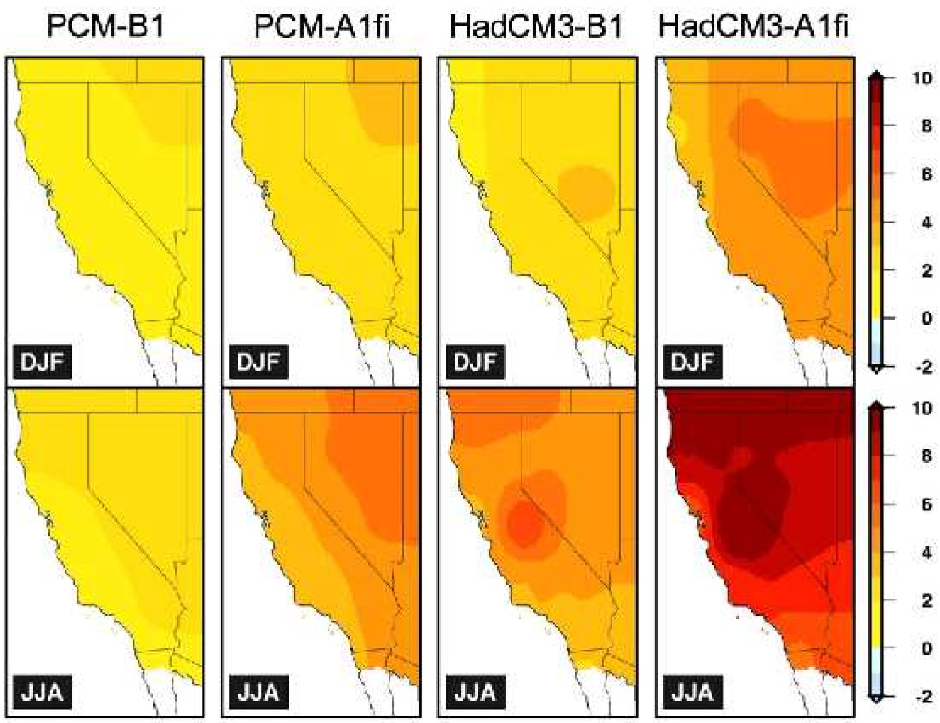
	PIcntrl	PDcntrl	20C3M	Commit	SRESA2	SRESA1B	SRESB1	1%
BCC-CM1, China	0	1	1	0	1	0	1	
BCCR-BCM2.0, Norway	0	0	0	0	0	0	0	
CCSM3, USA	1	1	1	1	1	1	1	
CGCM3.1(T47), Canada	1	0	1	1	1	1	1	
CGCM3.1(T63), Canada	1	0	1	0	0	1	1	
CNRM-CM3, France	1	0	1	1	1	1	1	
CSIRO-Mk3.0, Australia	1	0	1	1	1	1	1	
ECHAM5/MPI-OM, Germany	1	0	1	1	1	1	1	
ECHO-G, Germany/Korea	1	1	1	1	1	1	1	
FGOALS-g1.0, China	3	0	3	3	0	3	3	
GFDL-CM2.0, USA	1	0	1	1	1	1	1	
GFDL-CM2.1, USA	1	0	1	1	1	1	1	
GISS-AOM, USA	1	0	1	0	0	1	1	
GISS-EH, USA	1	0	1	0	0	1	0	
GISS-ER, USA	1	0	1	1	1	1	1	
INM-CM3.0, Russia	1	0	1	1	1	1	1	
IPSL-CM4, France	1	1	1	1	1	1	1	
MIROC3.2(hires), Japan	1	0	1	0	0	1	1	
MIROC3.2(medres), Japan	1	0	1	1	1	1	1	
MRI-CGCM2.3.2, Japan	1	1	1	1	1	1	1	
PCM, USA	1	1	1	1	1	1	1	
UKMO-HadCM3, UK	1	0	1	1	1	1	1	
UKMO-HadGEM1, UK	1	0	1	0	1	1	0	

a shaded box indicates that at least some, but not necessarily all, fields o

1 of 1 11/13/2005 02:37 PM

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

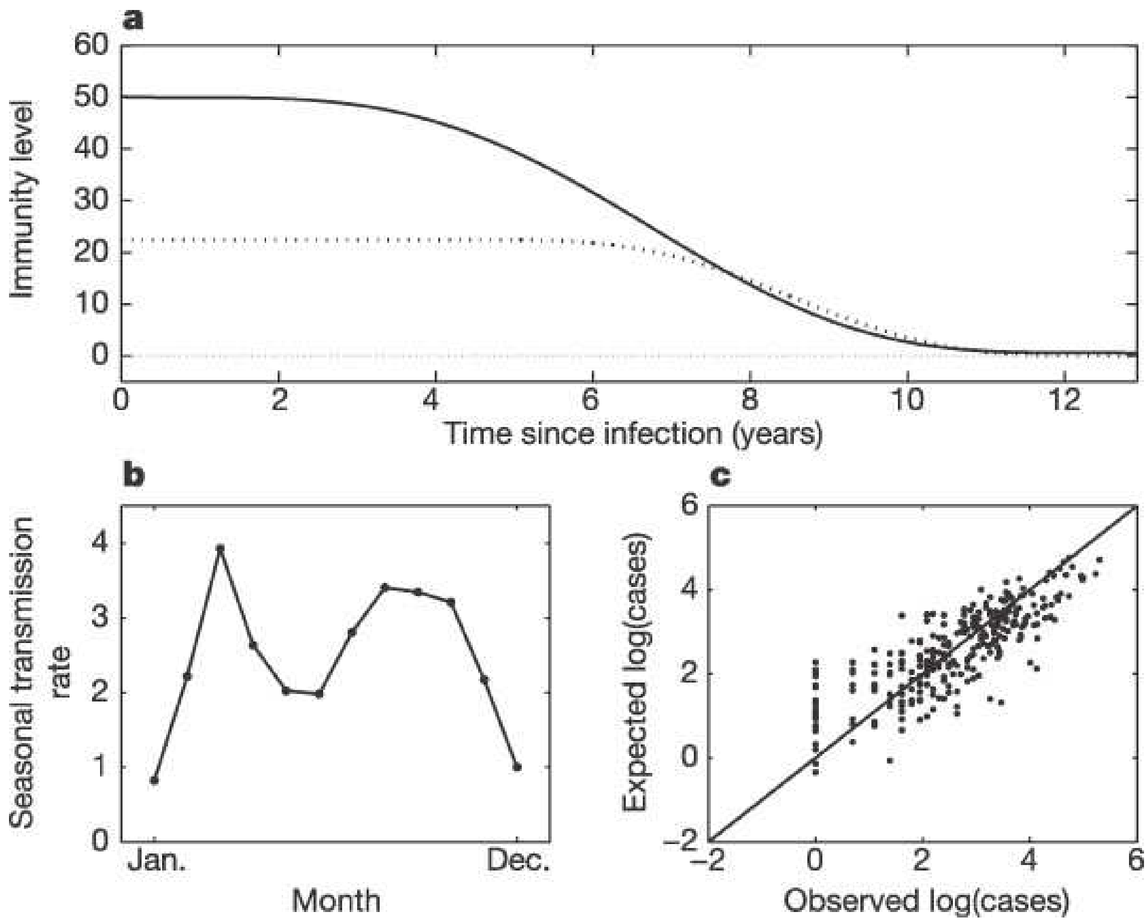
Scenario 2: statistical downscaling of climate change projections



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).

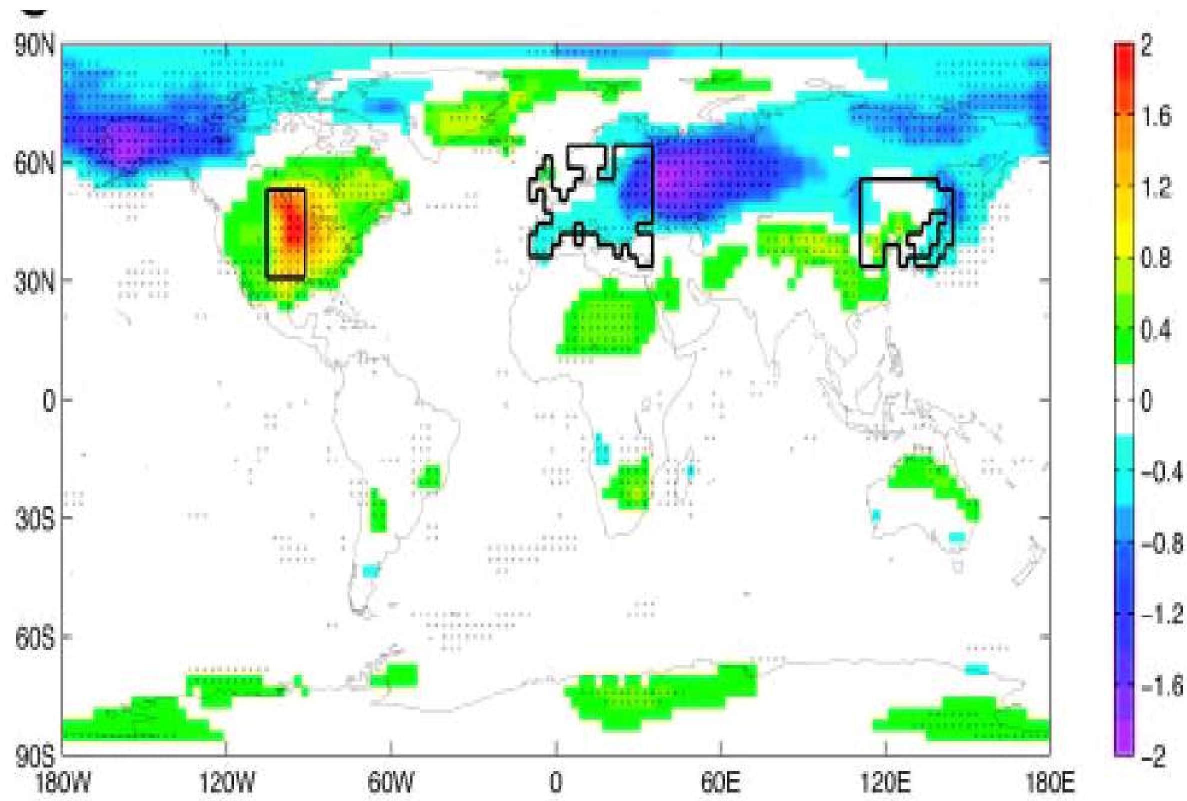
Scenario 3: disease vectors in a changing climate



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. 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.

- 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).