The Earth System Curator: integration technology for models and data

ESMF Community Meeting, Boulder

V. Balaji (**balaji@princeton.edu**)[1]

[1]Princeton University

NOAA/GFDL

30 May 2007

Outline

Outline

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* forecasts delivered to the public;

Decision support models routinely run for climate policy, energy strategy, risk pricing.

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 will require a radical shift in the way we do modeling: an infrastructure for moving the building, running and analysis of models and model output data from the "heroic" mode to the routine mode.

Curators: a noughties technology

- The *comparative study of climate simulations* (e.g IPCC) 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.
- The key element in the integration will be a curator. A curator begins begins with a 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*.
- ESC – the Earth System Curator – is a pilot project building prototype elements of such a system. The current project is funded by NSF and brings together NCAR, Princeton, MIT and GA Tech.

Curators: a noughties technology

- The *comparative study of climate simulations* (e.g IPCC) 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.
- The key element in the integration will be a curator. A curator begins begins with a 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*.
- ESC – the Earth System Curator – is a pilot project building prototype elements of such a system. The current project is funded by NSF and brings together NCAR, Princeton, MIT and GA Tech.

Curators: a noughties technology

- The *comparative study of climate simulations* (e.g IPCC) 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.
- The key element in the integration will be a curator. A curator begins begins with a 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*.
- ESC – the Earth System Curator – is a pilot project building prototype elements of such a system. The current project is funded by NSF and brings together NCAR, Princeton, MIT and GA Tech.

Linking model and data frameworks

Community data frameworks (e.g ESG, an ESC partner) are under development, at various institutions, informally linked by the GO-ESSP. For model output data to be scientifically useful, the researcher must have some knowledge of how the data was produced. Model data requires a *model's eye view* description of the data, another layer of metadata, which might include:

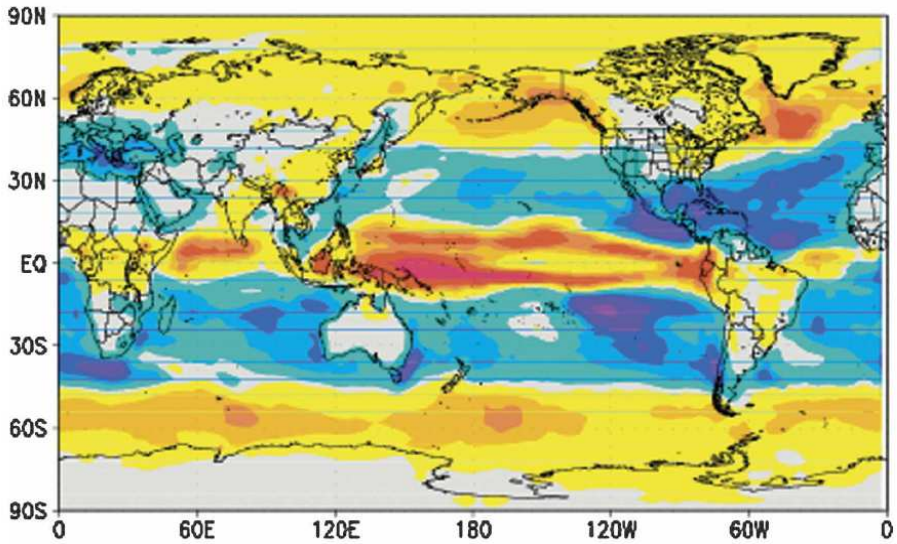
- Description of model components: e.g GEOS-5 atmosphere, land and sea ice coupled to MIT ocean.
- Description of grid configurations and resolutions.
- Choice of physics packages and input parameters.
- Model state and its fields.

ESMF and PRISM are emerging standards that allow the development of the model metadata layer, based on the state data structures and its base classes. (Think **`State`**, **`Grid`**, **`LocStream`**, `...`)

Semantic vs. syntactic, discovery vs. use

Descriptive metadata can be succinct, and can be used to discover certain aspects of the data. But almost any serious use requires deeper knowledge. The boundary between *discovery* and *use*, *semantic* and *syntactic*, is blurred by the use of controlled vocabularies and ontologies.

Graphics such as this from Held and Soden (2006) are so routinely produced from the IPCC AR4 database that we've ceased to marvel at it. This is a composite of output from 20 models worldwide, run with minimal coordination.



Model and grid metadata

Physical fields: standard vocabulary for describing the relevant physical quantities (viz. CF **`standard_name`**).

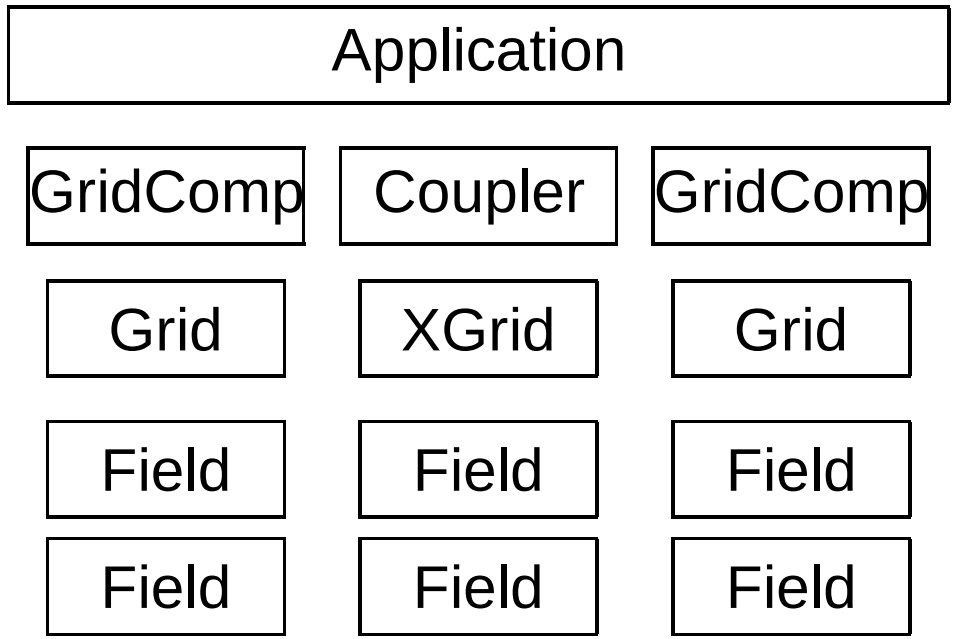
Geospatial information: location information: latitude, longitude, elevation. This set of standards unites a much larger community (mobile phones, GIS), in which our community has begun to play a role. We can provide some useful extensions toward 3D and 4D data.

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.

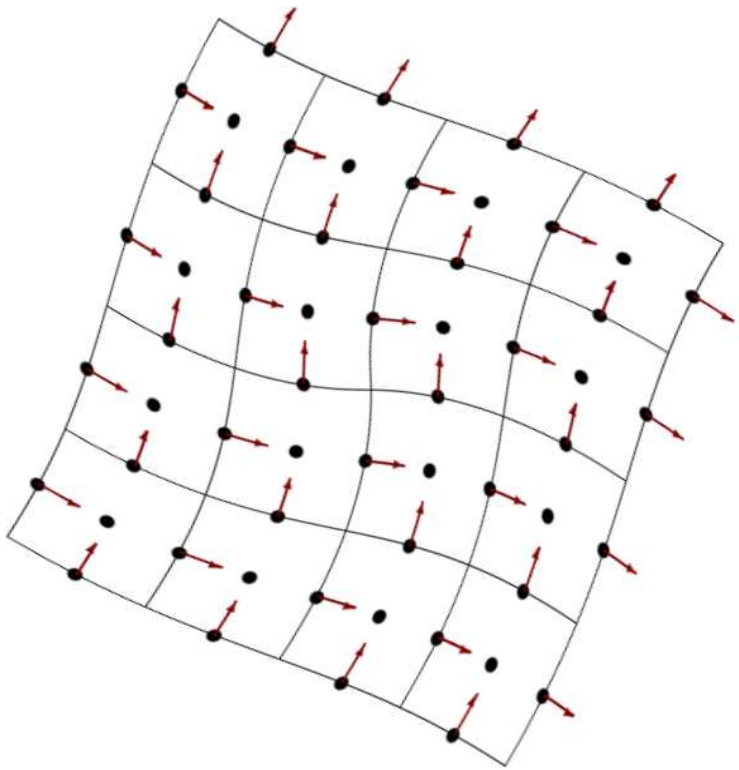
Model metadata

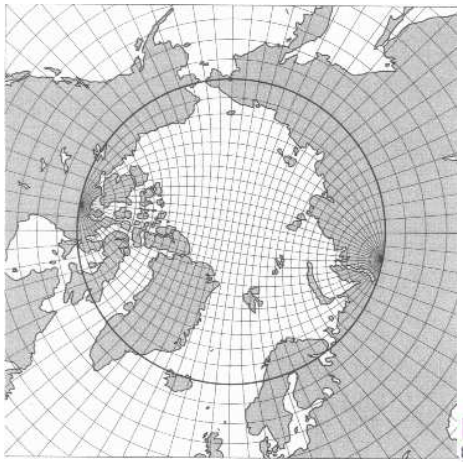
- Application metadata: experiment, scenario, institution, contact: currently covered by CF/CMOR.
- Component metadata: physical description of component: currently covered by CMOR, extended by NMM.
- Coupler metadata: inventory of export and import fields, interpolation methods. Currently covered by OASIS4 XML, not exported to model output. Associated with an XGrid.

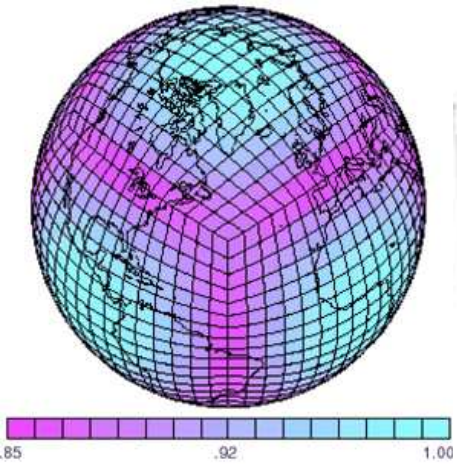


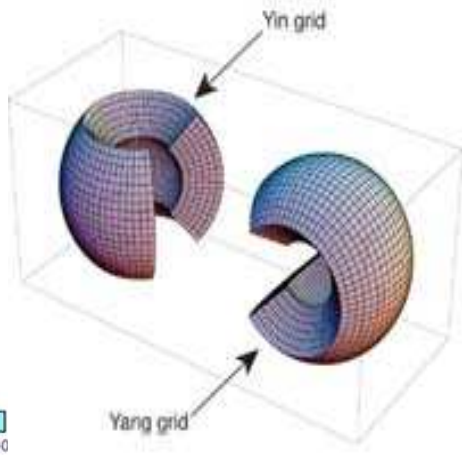
Grid metadata

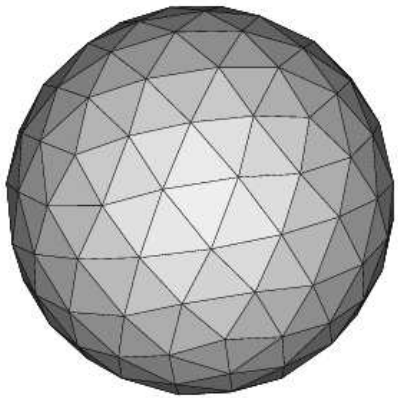
The Mosaic Gridspec is now under consideration as a draft CF standard. Contains information for differentiation, integration and regridding on generalized grids. Currently implemented for coupling models at GFDL (cubed-sphere atmosphere, tripolar ocean, lat-lon land and river grids) and as an XML schema for data web services in the European Genie project.











Outline

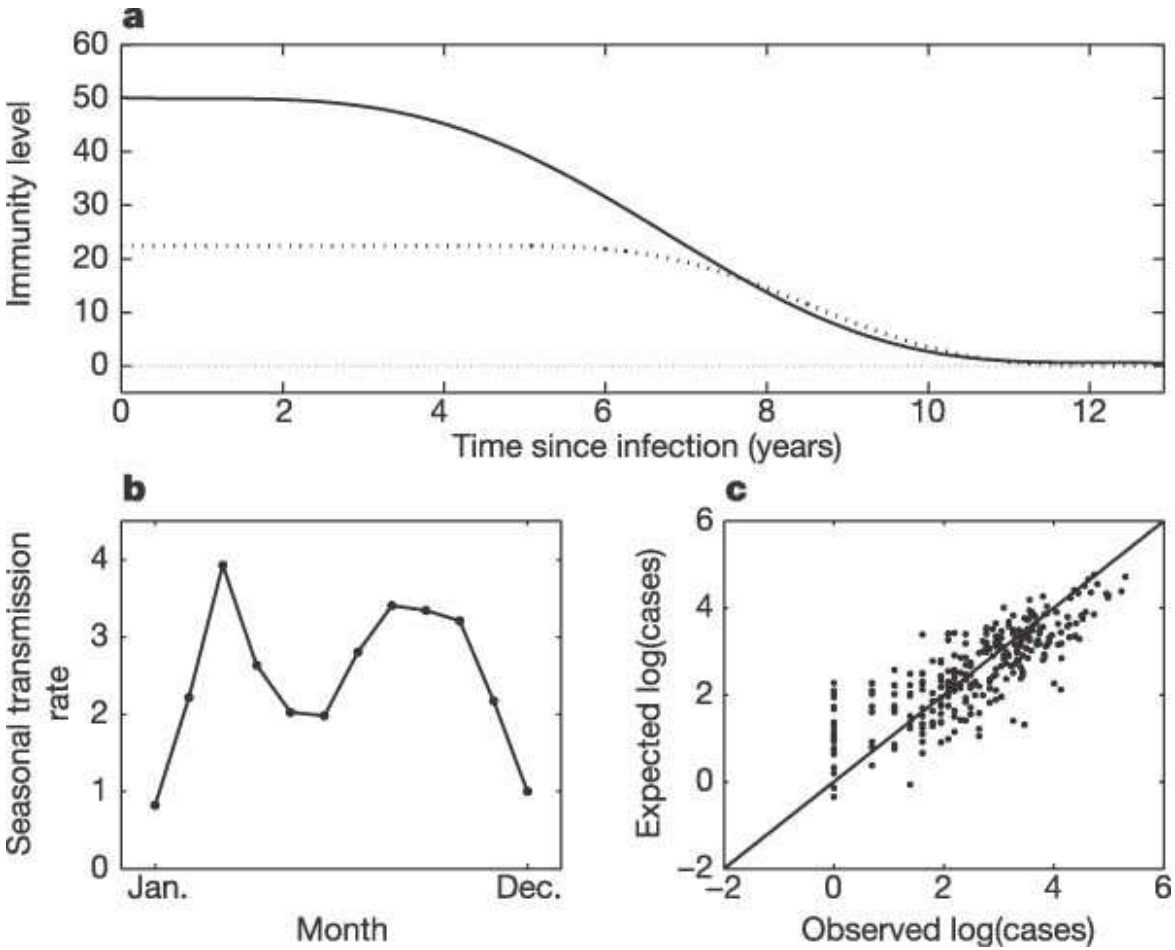
Dynamically derived data catalogues

Table 1 The models used in the present study, including, configurations (near the equator) and number of years of simulations

Model	Institution	Atmosphere resolution	Ocean resolution	Length picntrl	Length 1pctto2x	Length 1pctto4x
CCSM3	NCAR (USA)	T85L26	1.125°×0.27°L40	230	150	n/a
CGCM3.1(T47)	CCCMA (Canada)	T47L31	1.85°×1.85°L29	500	150	150
CNRM-CM3	Meteo-France/CNRM (France)	T63L45	2°×0.5°L31	390	100	110
CSIRO-Mk3.0	CSIRO (Australia)	T63L18	1.875°×0.84°L31	380	10	n/a
ECHAM5/MPI-OM	MPI-M (Germany)	T63L31	1.5°×0.5°L40	332	100	81
FGOALS-g1.0	LASG/IAP (China)	T42L26	1°×1°L33	150	80	n/a
GFDL-CM2.0	GFDL (USA)	2.5°×2°L24	1°×0.33°L50	500	100	160
GFDL-CM2.1	GFDL (USA)	2.5°×2°L24	1°×0.33°L50	500	150	160
GISS-AOM	NASA/GISS (USA)	4°×3°L12	4°×3°L16	251	n/a	n/a
GISS-EH	NASA/GISS (USA)	5°×4°L20	2°×2°L16	500	80	150
GISS-ER	NASA/GISS (USA)	5°×4°L20	5°×4°L13	400	100	n/a
INM-CM3	INM (Russia)	5°×4°L21	2.5°×2°L33	330	n/a	n/a
IPSL-CM4	IPSL (France)	2.5°×3.75°L19	2°×0.5°L31	230	80	n/a
MIROC3.2(hires)	CCSR/NIES/FRCGC (Japan)	T106L56	0.28°×0.1875°L47	100	10	n/a
MIROC3.2(medres)	CCSR/NIES/FRCGC (Japan)	T42L20	1.4°×0.5°L43	500	100	150
MRI-CGM2.3.2	MRI (Japan)	T42L30	2.5°×0.5°L23	350	150	150
PCM	NCAR (USA)	T42L18	0.66°×0.5°L32	350	96	90
UKMO-HadCM3	HadleyCentre (UK)	3.75°×2.5°L19	1.25°×1.25°L20	341	10	n/a
UKMO-HadGEM1	HadleyCentre (UK)	1.875°×1.25°L38	1°×0.33°L40	80	10	n/a
SINTEX T30	IPSL/INGV (France,Italy)	T30L19	2°×0.5°L31	200	n/a	n/a
SINTEX T106	INGV/IPSL (Italy,France)	T106L19	2°×0.5°L31	100	n/a	n/a
SINTEX T106mod	IPSL/INGV (France,Italy)	T106L19	2°×0.5°L31	100	n/a	n/a
HadOPA	CGAM/IPSL (UK,France)	3.75°×2.5°L19	2°×0.5°L31	100	n/a	n/a

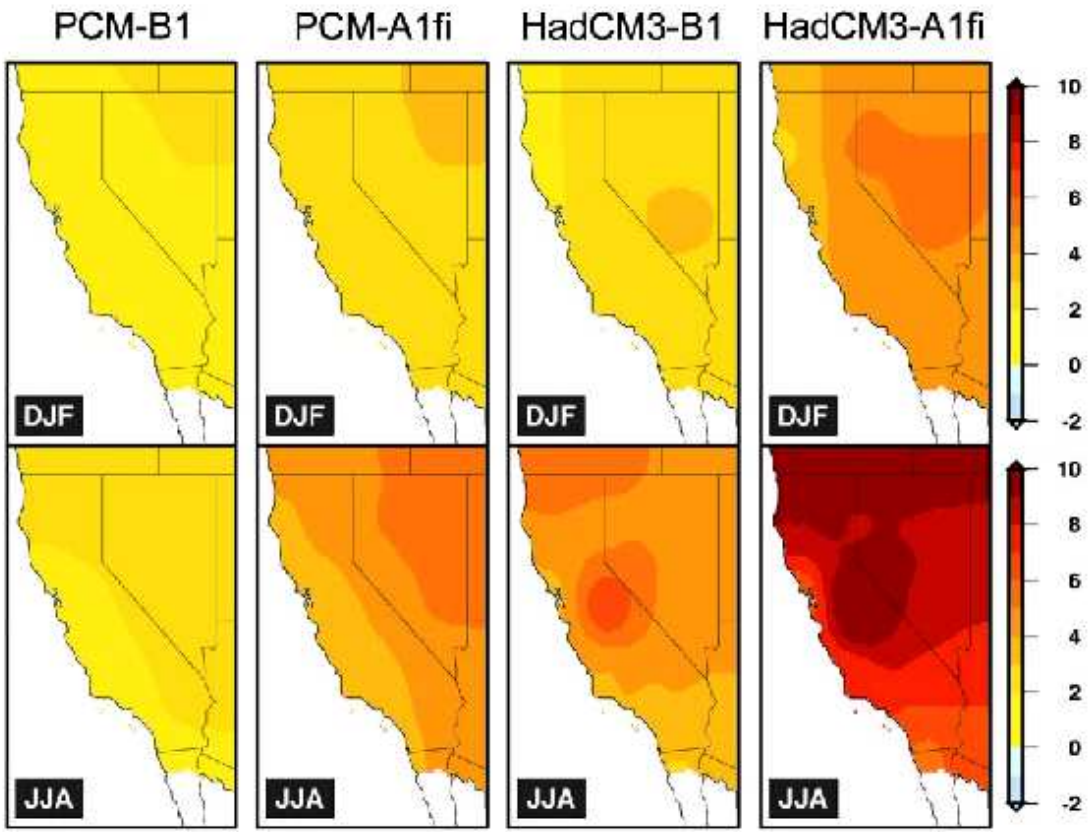
From Guilyardi (2006): for the most part, tables such as this are laboriously filled by hand. Currently deployed data frameworks (PCMDI, GFDL, DDC) are becoming capable of dynamically generating these tables.

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. This usage is typical of IPCC WG2 users.

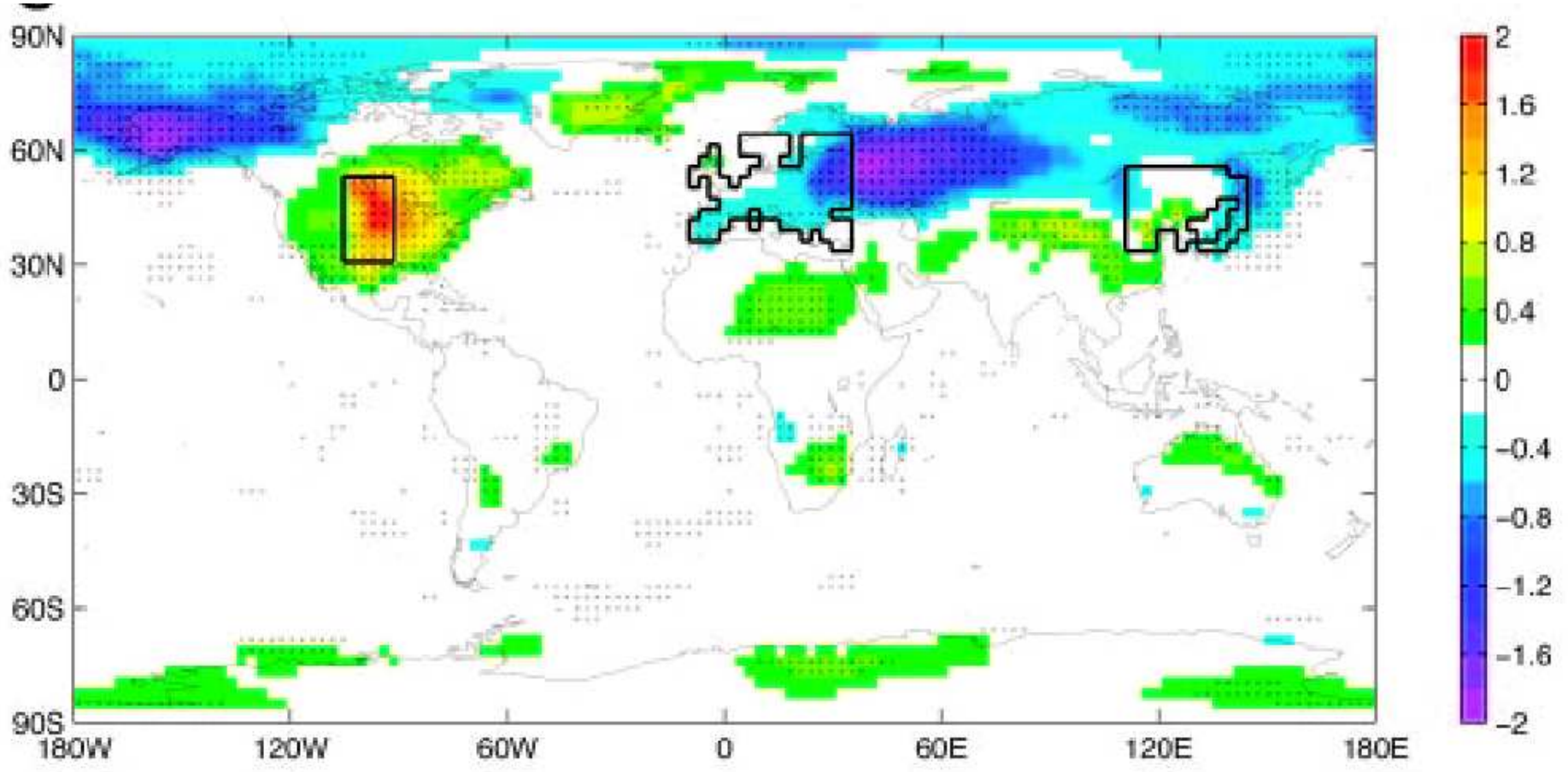
Statistical downscaling of climate change projections



Hayhoe et al, *PNAS*, 2004: *Emissions pathways, climate change, and impacts on California.* Uses daily data for “heat degree days” and other derived quantities.

What if it requires data beyond that provided by IPCC AR4 SOPs (1960-2000)?

Alternate energy sources

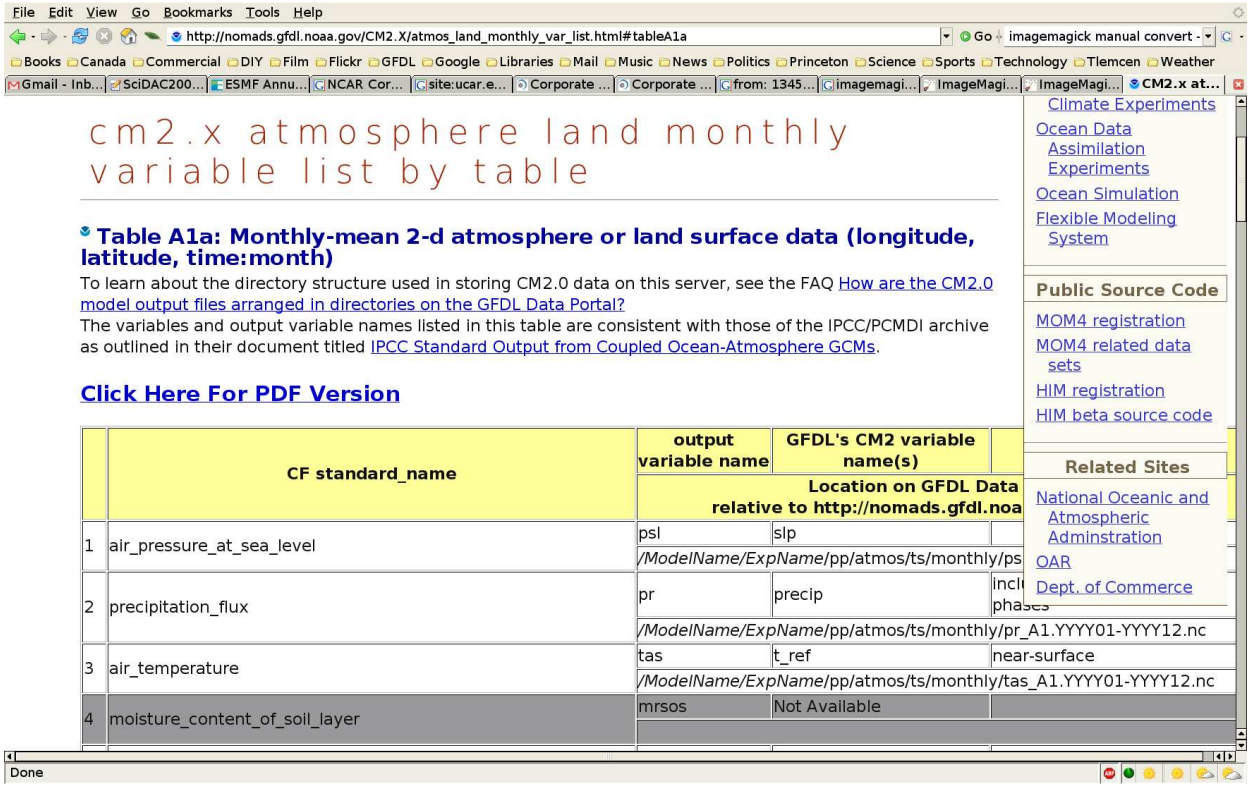


Keith et al, *PNAS*, 2005: *The influence of large-scale wind power on global climate.*

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

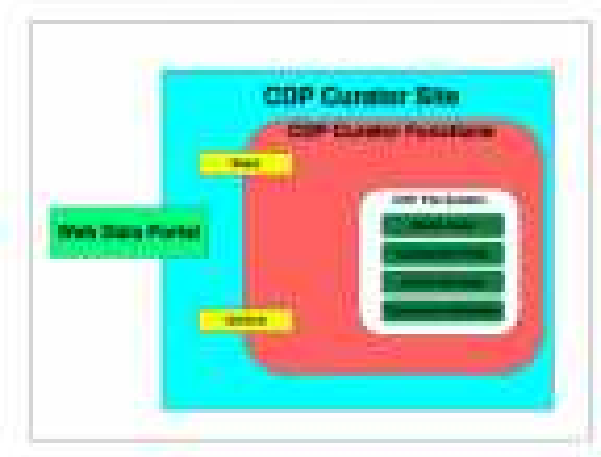
Outline

GFDL Curator and CDP Curator



GFDL Curator currently creates dynamic data catalogues from metadata (currently not from the Curator schema itself, but from metadata embedded in the datasets: integration with schema is underway).

The CDP Curator provides an interface both query and archival of ESMF components.



Operational use of model frameworks

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

- Source code maintenance across many repositories;
- Model configuration, launching and regression testing encapsulated in XML descriptors;
- Relational database for archived model results;
- Standard and custom diagnostic suites;
- Branching: "descent with modification".

FRE: a prototype for CRE

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, ...`

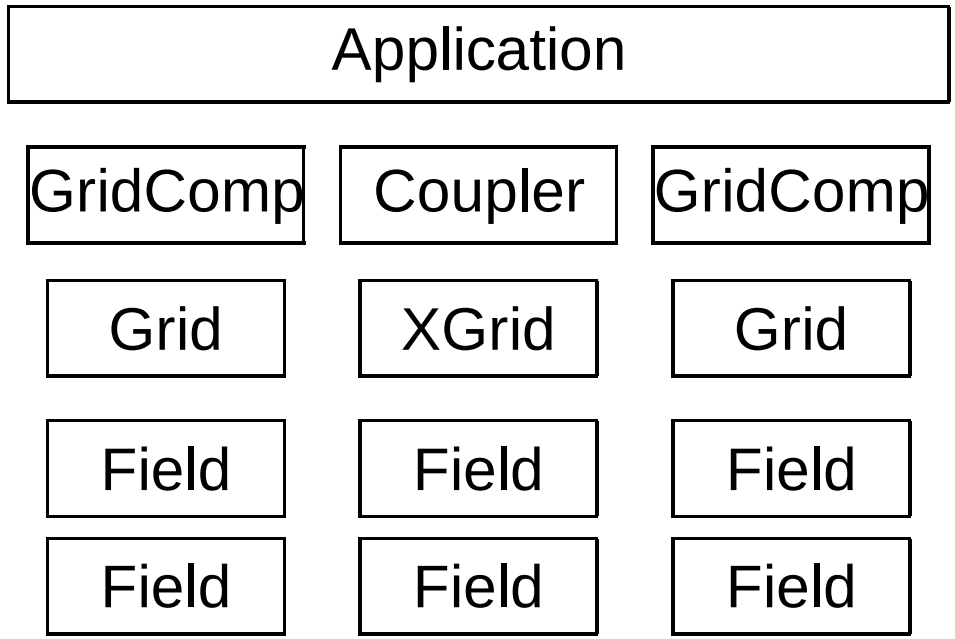
The Regression Test Suite (RTS) is a set of tests that are run continuously on a set of FMS models to maintain and verify code integrity.

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. We are currently merging FRE and ESC schema to make FRE serve as a prototype Curator Runtime Environment.

`http://www.gfdl.noaa.gov/~fms/fre`

ESC links to ESMF and MAPL

- ESMF data structures (**`Grid`**, **`Field`**, **`LocStream`**, **`ConfigAttr`**, **`State`**, ...) are ideal containers for holding the metadata. Tools are under development for extracting the Curator metadata from ESMF components registered under the Community Data Portal.
- MAPL specifications of coupling will be encoded in Curator coupler metadata, which itself is being developed on the basis of the OASIS4 (PRISM) schema.



ESC links with other projects

- Earth System Grid is developing key technology for serving data from multiple modeling centers (IPCC, NARCCAP), and are partners in the metadata definition effort focused on AR5.
- Numerical Model Metadata (NMM) based in the University of Reading defines discovery metadata for Earth System Models, and have converged with ESC schema where there is overlap. The Metafor proposal seeks to formalize the relationship further.
- The FLUME project at the UK Met Office is developing similar metadata and is interested in looking at auto-generation of FLUME "glue" using the common information model. Also part of Metafor.

Outline

Summary

- Curators are a natural outgrowth of frameworks. By blurring the edges between models and data, between objects and services, they are a typical "fuzzy boundary" technology. (Roger Sessions, *ACM Queue*, 2004: *Fuzzy Boundaries: Objects, Components, and Web Services*.)
- The ESC project is drawing up a metadata architecture which can aggregate within a single information model, metadata layers (e.g component, coupler, campaign) to be assigned to different domains of expertise.
- A functional schema-driven runtime environment allowing composition, configuration, running, archival, and analysis of Earth system model data is already a (limited) reality, and we are currently probing its limits and imposing generalizations.

`http://www.earthsystemcurator.org`