Earth System Model Curator

V. Balaji
SGI/GFDL Princeton University

ESMF Team Meeting, Princeton
13 May 2003

Emerging metadata standard

- There are emerging standards and conventions to allow some aspects of datasets to be standardized for easy sharing across institutions (CF convention for variables: the `standardName` attribute).

- There are attempts to extend this to other aspects of datasets such as descriptions of grids, interpolators and masks (GFDL proposal for a standard `gridSpec` file).

- Standards such as these allow the development of portals into distributed data servers, and a true Data GRID: where datasets and data may be manipulated through tools either client-side (download and process) or server-side (process and download). (NOMADS, ESG, CAPRI...)

Layered metadata

Standardized descriptions of variables and grids in datasets is necessary but not sufficient step: the data, especially model output, is only useful when the researcher has 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 includes:

- Description of model components: e.g FMS BGRID atmosphere, land and sea ice coupled to MITgcm ocean.
- Description of grid configurations and resolutions.
- Choice of physics packages and input parameters.

ESMF and PRISM are emerging standards that allow the development of the model metadata layer, based on the `importState`, `exportState` and `configAttr` data structures.

Convergence of models and datasets

Given the existence of a model metadata layer, *the same descriptor can be used as model input and model output.* This means:

- the files that are used to configure, build and launch a model (written in, say, XML) contain the same physical information that must be written to the output dataset for a comprehensive description of how the data was generated.

- This information can also be stored in a relational database of model configurations and datasets: the Earth System Model Curator. Such a DB would allow experiment comparisons, high-level queries, experiment redesign, next-generation publication of scientific results.

Potential use scenarios

Climate scientists setup (assemble components, configure input parameters); comparisons (run configurations, results, with data); branch runs, ...

Impacts studies query models by pattern, couple biogeochemistry model either offline with dataset or online with model.

IPCC, MIPs descriptions of intercomparisonns, setup new MIPs, archive MIP results.

Policymakers High-level access to swathes of model data.

Publication link datasets to publications; introduce interactive aspect to publication; annotation of data, certification and quality control.

Portability automatic best-practice configuration appropriate for platform.

Proposal

The proposal would be to unite the data (NOMADS, ESG) and model (ESMF) communities with climate scientists (IPCC, CMIP) to develop the model metadata layer, and the relational database of models and data that would be based on it.

This effort would be closely allied with the PRISM/CAPRI efforts in the same domain.

Elements of proposal

Physical interfaces development of comprehensive physical interfaces for model components.

Hierarchical metadata development of a semantic web of model and data descriptors.

Relational database of model experiments and observational and model datasets.

Data annotation certification by assigned authority, or *à la* Google. Links with scientific results and peer-reviewed literature.

Web portal interfaces to query operations, comparisons, client- and server-side data analysis.