The Earth System Grid Federation: software framework supporting CMIP5 data analysis and dissemination

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2011: The CMIP5 Experiment Design. Bull. Amer. Meteorol. Soc., submitted.

Uppala S, D. Dee, S. Kobayashi, P. Berrisford and A. Simmons, 2008: Towards a climate adapt assimilation system: Status update of ERA-Interim, ECMWF Newsletter, 115, 12-18.

Van der Linden. P and J.F.B. Mitchell (eds). 2009: ENSEMBLES: Climate Change and its Impacts: Summary of Research and Results from the ENSEMBLES Project. Met Office Hadley Centre: Exeter.

Wang, Y., et al., 2004: Regional climate modeling: progress challenges and prospects. Journal of the Meteorological Society of Japan, 82, 1599-1628.

Wilby, R. L., and T. M. L. Wigley, 2000: Precipitation predictors for downscaling: Observed and General Circulation Model relationships. International Journal of Climatol., 20, 641–661,

The Earth System Grid Federation: Software Framework Supporting CMIP5 Data Analysis and Dissemination

Dean N. Williams[1], Karl E. Taylor[1], Luca Cinquini[2], Ben Evans[3], Michio Kawamiya[4], Michael Lautenschlager[5], Bryan N. Lawrence[6], Don E. Middleton[7], and the ESGF contributors.

1 Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory (LLNL).
2 Jet Propulsion Laboratory (JPL), National Aeronautics and Space Administration (NASA).
3 Australia National University (ANU) National Computational Infrastructure (NCI).
4 Japan Agency for Marine-Earth Science and Technology (JAMSTEC)
5 German Climate Computing Centre (DKRZ).
6 National Centre for Atmospheric Science / British Atmospheric Data Centre (BADC).
7 National Center for Atmospheric Research (NCAR).

The Earth System Grid Federation (ESGF) is a coordinated international collaboration of individuals and institutions that is developing, deploying and maintaining software infrastructure for the management of model output and observational data. The goal of this effort is to facilitate advancements in Earth System Science. Through the ESGF alliance, governed under the Global Organization for Earth System Science Portals (GO-ESSP), the team has developed an operational system for serving climate data from multiple locations and sources. Model simulations, satellite observations, and reanalysis products will all be served from a distributed data archive. Researchers worldwide can now access ESGF data holdings through any of the gateways hosted by ESGF partners, including laboratories in the U.S. funded by the Department of Energy (DOE), the National Science Foundation (NSF), the National Aeronautics and Space Administration (NASA), and the National Oceanic and Atmospheric Administration (NOAA), and at laboratories elsewhere, for example at the Australian National University (ANU) National Computational Infrastructure (NCI), the British Atmospheric Data Center (BADC), the Max Planck Institute for Meteorology (MPI-M) German Climate Computing Centre (DKRZ). A good place to start if one wants access to CMIP5 output is the CMIP5 website and the "getting started" document .

In planning for CMIP5, the ESGF has built on the success of the earlier Earth System Grid (ESG) project, which served CMIP3 model output and with US DOE support was led by the Program for Climate Model Diagnosis and Intercomparison (PCMDI). CMIP5 has driven all ESGF development work and has attracted the interest of others seeking to make their data widely available and easy to use (e.g., CORDEX and satellite measurements for CMIP5, discussed elsewhere in this publication (see Jones et al. and Teixeira et al., this issue). The ESGF aims to:

- Support the current CMIP5 activity, and prepare for future assessments;
- Develop data and metadata facilities for inclusion of observations and reanalysis products in CMIP5;
- Enhance and improve current climate research infrastructure capabilities through involvement of the software development community and through adherence to sound software principles;
- Foster collaboration across agency and political boundaries;
- Integrate and interoperate with other software designed to meet the objectives of ESG: e.g., software developed by NASA, NOAA, ESIP, and the European ES-INES;
- Create software infrastructure and tools that facilitate scientific advancements.

The software deployed in ESGF has been developed using an open-source approach, and all participants are encouraged to contribute to the ongoing development of the infrastructure.

A detailed view of the components and capabilities provided by the ESGF is shown and described in Figure 1.

Not all aspects of the end-to-end preparation and archiving of CMIP5 model output are formally organized as part of the ESGF, but the entire process will be briefly summarized here . It starts with the individual modeling groups running models, following the CMIP5 experiment specifications described in Taylor et al. (2009, 2011). The CMIP5 output structure and metadata requirements ensure that analysts can read and interpret output from all models in a uniform way and is perhaps the most valuable aspect of CMIP5. A software library, called CMOR (Climate Model Output Rewriter), helps insure conformance with the requirements, while somewhat reducing the burden imposed on the modeling groups in preparing output. The simulation data, stored in files using the netCDF library and consistent with the Climate and Forecast (CF) metadata conventions , are then placed on an ESGF data node located either at the modeling center or one of the ESGF data centers. The data is then "published", which is a procedure that records information in the ESGF catalog and makes the data visible to users through ESGF gateways.

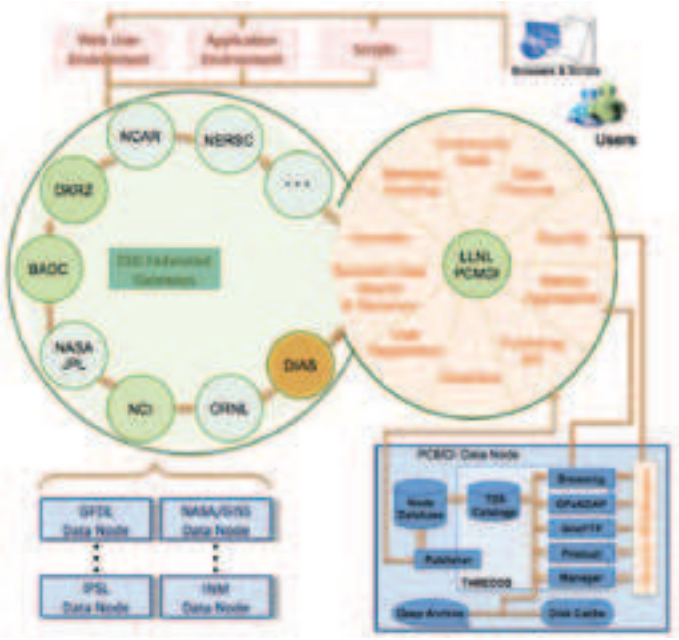

Figure 1: The figure shows how users can access ESGF data using web browsers, scripts, and soon by client applications. Conceptually, ESGF is composed of two interacting parts: ESGF gateways (indicated in green) and ESGF data nodes (indicated in blue). The expanded view shown for the gateway and data node hosted by PCMDI provides a more detailed picture of the various components and capabilities that are duplicated at the other the sites. Gateways handle user registration and management and allow users to search, discover, and request data. Data nodes are located where the data resides, allowing data to be "published" (or exposed) on disk or through tertiary mass store (i.e., tape archive) to any gateway. In addition, some advanced data nodes can handle data reduction, analysis, and visualization. ESGF currently comprises eight gateways, and four of these (indicated by darker shade of green) are special because they host replicas of a substantial number of the CMIP data sets. Users have access to all data from the federation regardless of which gateway is used.

As the modeling groups prepare the simulation output, they are expected to complete the METAFOR questionnaire (described by Guilyardi et al., this issue) whereby information is gathered to document the models and the simulations. The documentation is subsequently made accessible to users through the ESGF gateways using software developed collaboratively by the Curator, METAFOR, and ESG projects. An important new addition to CMIP5 is that a three-step quality control (QC) procedure is being applied to all model output. If the ESGF software can successfully read and obtain catalog information from the data files during "publication", then Level 1 of the QC procedure is passed. Level 2 QC involves extensive examination of metadata and data for self-consistency and conformance to standards. These quality checks are performed using the "QC Tool" developed at DKRZ. For example, a variable must have a recognized CF "standard name" attribute and its data values are checked to determine whether they fall within a range of values expected for the variable. Level 3 QC provides a few additional self-consistency checks and is passed only when the modeling group providing the data has agreed that the data should be permanently entered into the ESGF data holding (although it can subsequently be flagged as being flawed). DKRZ maintains a database of the results for every QC check performed by their tool anywhere within the ESGF archive.

For CMIP5, the ESGF may eventually archive 3 petabytes (PB, 1 PB = 10^{15} bytes) or more of officially requested data . Some simulation output is expected to be of interest to only a handful of specialists, whereas a few fields are likely to attract widespread interest (e.g., surface temperature and precipitation). To assure the high interest data are preserved and that this subset of output is readily accessible worldwide, it will be replicated by one or more of the ESGF partners. Replication does not proceed until QC Level 2 is passed.

The final step (after QC Level 3 has been passed) is to assign a Digital Object Identifier (DOI) to each dataset within the CMIP5 collection. To give appropriate credit to the data providers, these DOI's should be cited when results are published based on the CMIP5 model output.

Although some CMIP5 model simulations were completed in early 2010, it was not until March of 2011 that the first model output was published and made available to users. By May of 2011 some simulation output was available from modeling groups in the UK, Russia, France, and the U.S.A. A current list of models that have contributed to the CMIP5 archive can be found at the CMIP5 website . There are also observational datasets that will become available soon through ESGF which have been written in the same structure with similar metadata conventions as the CMIP5 model output. Teixeira et al. (this issue) describe one such effort.
At present the available CMIP5 model output can be viewed only via web browsers hosted by the following ESGF gateways, with data centers planning to replicate a significant subset of the model output highlighted by the asterisks:

- PCMDI*: http://pcmdi3.llnl.gov/esgcet/home.htm
- BADC*: http://cmip-gw.badc.rl.ac.uk/home.htm
- DKRZ*: http://ipcc-ar5.dkrz.de/home.htm
- JPL: http://esg-gateway.jpl.nasa.gov/home.htm
- NCAR: http://www.earthsystemgrid.org

- NCI*: http://esg.nci.org.au/esgcet/home.htm
- NERSC: http://esg.nersc.gov/esgcet/home.htm
- ORNL: http://esg.ccs.ornl.gov/esgcet/home.htm

Note that regardless of where data may be located, all holdings are visible at any ESGF gateway that is configured to display it. Thus a user can browse the federation's holdings from any gateway and obtain the data of interest. A help desk staffed by ESGF collaborators provides support to CMIP5 users across the federated system.

With CMIP5 data now being served, the ESGF federation is working to improve various aspects of the system by adding new capabilities that should better meet the needs of users. Among the improvements expected over the next several months are:

1. A simpler scripting method for downloading files;
2. An enhanced search capability;
3. An automatically updated table showing which simulations have been archived by each model;
4. A notification service to advise users when errors are found in datasets;
5. A straight-forward method to report errors discovered in the data and to provide feedback to the modeling groups about their simulations;
6. A list of publications based on CMIP5 model output, as recorded by users through a web form;
7. General system enhancements related to scaling to millions of datasets and petabytes of data volume;
8. An online visualization capability that will allow users quick inspection and comparison of datasets from multiple locations;
9. An enhanced capability to perform server-side data reduction and calculations, which will reduce the volume of data transferred to the users via the Internet.

References

Taylor, K.E., R.J. Stouffer, and G.A. Meehl, 2009: A summary of the CMIP5 Experimental Design. http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf.

Taylor, K.E., R.J. Stouffer, and G.A. Meehl, 2011: An Overview of CMIP5 and the Experiment Design. Bull. Amer. Meteor. Soc., submitted.

The CMIP5 model and simulation documentation: a new standard for climate modelling metadata

Eric Guilyardi[1], V. Balaji[2], Sarah Callaghan[3], Cecelia DeLuca[4], Gerry Devine[5], Sébastien Denvil[6], Rupert Ford[7], Charlotte Pascoe[3], Michael Lautenschlager[8], Bryan Lawrence[3], Lois Steenman-Clark[5], Sophie Valcke[9]

1 NCAS, University of Reading, UK and IPSL, Paris, France
2 GFDL, Princeton, USA
3 NCAS-BADC, STFC, UK
4 NCAR, Boulder, USA
5 NCAS, University of Reading, UK
6 IPSL, Paris, France
7 University of Manchester, UK
8 DKRZ, Hamburg, Germany
9 CERFACS, Toulouse, France

Together with the data transformation towards a standard format and the archiving of output files in the distributed ESG Federation, the standard model and simulation documentation process is an essential part of the CMIP5 process. The development of the associated metadata and web questionnaire is described in this article.

Climate modelling metadata: sharing the climate scientist's notebook

The outputs of climate models are increasingly used, not only by the climate scientists that produce them, but also the growing number of stakeholders which study climate change as well as policy-makers and the enlightened public. Climate modelling data is stored in huge and complex digital repositories (Overpeck et al., 2011). Hence, archiving, locating, assessing and making sense of this unique resource requires accurate and complete metadata (data describing data). Climate model simulations, such as those prepared for CMIP5 , involve several component models (atmosphere, ocean, sea-ice, land surface, land ice, ocean biogeochemistry, atmosphere chemistry) coupled together that follow a common experimental protocol (Taylor et al., 2009; 2011). Each of these component models can be configured in many different ways, including not only different parameter values but also changes to the source code itself. Component models, or even compositions of component models, can have multiple versions, and individual component models can be coupled together and run in a myriad of different ways. The range of possibility is immense. Until now, this key information can only be found in the climate scientist's experimental notebooks, hence largely under-documented in the output data itself. Community multi-model database provided the first incentive for a common description, as for instance initially proposed for CMIP3 .