Documenting climate models and their simulations

Eric Guilyardi (NCAS, University of Reading, UK and IPSL/LOCEAN, France), V. Balaji (GFDL, US), Bryan Lawrence (NCAS, University of Reading and STFC, UK), Sarah Callaghan (STFC/BADC, UK), Cecelia Deluca (NOAA, US), Sébastien Denvil (IPSL, France), Michael Lautenschlager (DKRZ, Germany), Mark Morgan (IPSL, France), Sylvia Murphy (NOAA, US), Karl E. Taylor (PCMDI, US) and the METAFOR team

Abstract

The results of climate models are of increasing and widespread importance. No longer is climate model output of sole interest to climate scientists and researchers in the climate change impacts and adaptation fields. Now non-specialists such as government officials, policy-makers, and the general public, all have an increasing need to access climate model output and understand its implications. For this host of users, accurate and complete metadata (i.e., information about how and why the data were produced) is required to document the climate modeling results. Here we describe a pilot community initiative to collect and make available documentation of climate models and their simulations. In an initial application, a metadata repository is being established to provide information of this kind for a major internationally coordinated modeling activity known as CMIP5 (Coupled Model Intercomparison Project, Phase 5). It is expected that for a wide range of stakeholders this and similar community-managed metadata repositories will spur development of analysis tools that facilitate discovery and exploitation of earth system simulations.

Main text

The range of possible results inherent in climate simulations can be examined using ensembles of simulations to explore the influence of climate model formulation as well as those of initial and boundary condition changes. Multi-model ensembles are now being used both in seasonal to inter-annual forecasts and on longer timescales and under specified external forcing, also for projections of climate change. The range of results produced by different models can be partly attributed to the chaotic nature of the climate system itself, but the range can also be dominated by differences in model formulation. Today, the differences and similarity between models is not well-documented nor widely understood either by the research community or by potential users of climate model output. Here we describe a project that seeks to collect better documentation of these climate models and their simulations. Such documentation is a necessary step towards understanding how the diversity of existing models influences the differences and uncertainty in their predictions.

The need to understand structural uncertainty among climate models

The climate system features a complex interplay of processes and feedbacks, ranging from ecosystems to radiative transfer. Numerical climate simulations therefore involve several “component models” which are coupled together into an “earth system model” to simulate the atmosphere, ocean, sea-ice, land surface, and land ice and which can account for processes governing the physical, chemical and biological behavior of the system. The complexity of these models, gauged in terms of the number of the processes that are represented, continues to grow. The complexity of each individual process is scientifically challenging enough that there is sometimes disagreement as to the best way to represent them in models. Without an obvious correct choice, the corresponding "structure" of a model must be considered uncertain. This is referred to as "structural uncertainty" (see also Palmer 2012). In different climate models, the interactions of clouds with the large-scale atmospheric circulation and with radiation, for example, have been encoded in a number of different ways, which partially accounts for the structural uncertainty evident in the climate projections produced by these models (Hawkins and Sutton 2009).

In this context, multi-model ensembles are now being used to explore the range of results arising from structural uncertainty. Different models perform the same numerical experiments under agreed-upon common protocols which specify, for instance in the case of 20th century simulations, the imposed external forcing (solar, volcanic, and anthropogenic). Such comparative studies of models, internationally organized under formal model intercomparison projects, have now become standard practice. They have spawned new lines of research into the evolution and genealogy of models (e.g. Masson and Knutti 2011) and into the skill of the model ensemble mean relative to the individual models (Reichler and Kim 2008, Gleckler et al. 2008, Santer et al 2009). Studies of this nature require precise descriptions of models so that their features can be contrasted and compared.

The expanding user community of climate model output

The results of multi-model climate simulations are also of increasing importance to broad segments of society. Not only scientists and researchers in the climate change impacts and adaptation fields, but also non-specialists such as policy-makers, local government officials and the general public now have a need to locate and understand the implications of climate simulations. Climate simulation data is stored in huge and complex digital repositories (Overpeck et al, 2011, Williams et al. 2011). For the fifth phase of the Coupled Model Inter-comparison Project (CMIP5, Taylor et al. 2012), organized by the Working Group on Coupled Modelling (WGCM) on behalf of WMO's World Climate Research Program (WCRP), over a million individual datasets and several petabytes of data will be generated. Locating and making sense of this community resource requires accurate and complete metadata. This is especially important for climate model simulation output, since many different variants of each of the component models exist, and it is essential to document which model version produced each dataset. A model version may differ from others, for example, not only in the values specified in the parametric representations of various processes, but also in the algorithms incorporated in the source code itself.

The need for climate modeling groups to make metadata available

Until now, much of the critical information needed to describe the model configurations and the experiments could only be found in the notebooks of individual climate scientists and the "comment statements" found in their computer codes; hence it was largely inaccessible by the broader community. Multi-model databases provided the first strong incentive for developing a common approach to recording descriptions. When dealing with multi-model databases, scientists and other stakeholders are increasingly seeking information about the suitability of available data for their purposes. Prior to CMIP5, this information was difficult to obtain. Thanks to recent community efforts, it should now be possible to obtain answers to questions like: What differences are there between the GFDL CM2.0 and GFDL CM2.1 models? (The only difference is the atmospheric dynamical core). Which simulations of the

20^{th} century have daily output data and use Turbulent Kinetic Energy (TKE) vertical mixing in the ocean? What is the grid resolution near the equator or over Europe in the IPSL-CM5A simulator? Are volcanoes included in the MIROC5 simulator and how? The current set of initial questions has for now (during this proof of concept phase) a scope targeted to climate scientists but, as described below, extension to other fields and stakeholders will be natural.

Many climate model configuration choices are determined by experimental requirements, but these still usually leave room for some differences in how a particular simulation is performed by each model. Hence, in addition to detailed documentation of the models themselves, the experiment conditions must also be fully documented. This information is not only important for scientific interpretation but, under increasing scrutiny from society, it is also demanded of a science that purports to be mature, credible, and open (Carlson, 2011; Kleiner, 2011). As a consequence, in early planning stages of CMIP5 the climate modelling community committed to collecting a comprehensive and standardized set of metadata for the climate model simulations. An important additional benefit of archiving such information is that this ensures that the conditions under which model simulations were performed will be understood well into the future (decades and beyond) when data may still be relevant not only for scientific reasons, but also possibly of historical interest. This "data curation" activity depends on such comprehensive information collected when the data are produced.

The METAFOR project and CMIP5 questionnaire

Acknowledging these challenges, a part of the climate research community committed itself to achieving the ambitious goal of defining, collecting, and making accessible model and experiment metadata for CMIP5. The aim was to make generally available an unprecedented level of detail in describing the models and simulations (Guilyardi et al. 2011). Funding for this international effort is being provided by the European Union (http://metaforclimate.eu, http://enes.org) and by the United States (http://earthsystemcurator.org), and guidance and encouragement is being provided by the WCRP. In the initial phase of the project, climate and information technology experts worked together to identify the information that would need to be collected to describe models and their simulations. The various types of metadata of interest were then organized into a new conceptual model, called the CIM (Common Information Model, Lawrence et al. 2011). This conceptual model was applied to the specific needs of CMIP5, and a metadata entry tool was developed to collect the information (the CMIP5 questionnaire, http://q.cmip5.ceda.ac.uk/).

[Box 1 here]

Climate scientists and modelers were initially expected to have the most interest in metadata that would help them understand the differences among the various simulations performed by different models. Hence, the content and structure of the model description section of the CMIP5 questionnaire largely reflect the needs of these groups. The input from climate modelers was obtained through direct interviews. Besides identifying the appropriate questions to include, a list of standardized responses was also developed. Converging on a first version of the questionnaire proved relatively straightforward and disagreements among experts were usually easy to address. Care was taken not to impose uniformity in areas where consensus within the climate research community has not yet been achieved and where agreed "standards" have yet to emerge. In addition, because of the sheer complexity of climate modeling and the finite resources available, a decision was made to limit the scope of the questions, leaving some areas of interest, such as the description of specific model tuning approaches or the choices of metrics used in model evaluation, for future development.

The modelling groups involved in CMIP5 are now entering information into a metadata catalogue, and the documentation for about 20 models and hundreds of simulations has already been recorded by the model developers through the web interface questionnaire. Data portals can harvest the information contained in the resulting machine-readable files and render it in a form more usable to humans. Given that different users will want to explore this database in different ways, it is essential to engage directly with various individual communities so that tools can be developed to address their specific needs. Such tools will depend on standard technical interfaces, which have already been developed, and the first tools aimed at displaying and searching the metadata are also now becoming available. Further development planned both in Europe and in the US will eventually enable more complex analysis of the metadata (e.g., determining the differences between two model versions).

[Box 2 here]

Conclusions and future outlook

As described above, this undertaking by the climate modelling community to collect and make accessible metadata in support of CMIP5 will provide the most comprehensive set of multi-model climate simulation metadata to date. Beyond CMIP5, the intention is that the CIM and the associated standards should become increasingly adopted by climate modelling frameworks in much the same way that promotion of the CF conventions (see http://cf-pcmdi.llnl.gov/) led to standardization of climate model output. To ensure continuity while allowing evolution, a governance structure to maintain and further develop the CIM and the associated "controlled vocabularies" is being proposed which will build on the structure already in place for governing the CF conventions. The term "controlled vocabularies" refers to the pre-defined and limited set of words, phrases, and names that comprise the metadata description.

Several funded projects (listed from http://es-doc.org) have already begun to build on this initial effort and to extend its scope and use in several directions (e.g. developing new tools, addressing the needs of additional user communities.). A possible extension, for example, is to link such model and simulation descriptions to the suite of model performance metrics now being devised by WCRP (http://www-metrics-panel.llnl.gov/wiki). Such metrics measure how the models perform compared to observations; linking these with detailed model description might help in determining the origin of some model errors.

Another project has succeeded in linking the CIM metadata for the EU ENSEMBLES simulations to the University of Cantabria downscaling portal (https://www.meteo.unican.es/downscaling/ensembles), thereby helping to meet the needs of the impacts community. It is expected that new tools will be developed to provide a synthesis of information in various CIM documents and to produce easily configurable, scientifically meaningful summaries of, for example, the differences between two models or two simulations. In time, it is expected that a diversity of users would engage their own experts to devise alternative ways of melding climate model output and documentation to best meet their needs.

Continuing developments and investments to record and archive climate model metadata are only part of a longer-term effort that should provide ongoing benefits to the community of users accessing climate model output. For instance, this model metadata archive will provide a much more comprehensive and up-to-date description of climate models than is typically available in journal articles or reports. Beyond the raw documentation, these community-managed metadata repositories will spur development of analysis tools for a wide range of stakeholders, providing a form of "Google advanced search" suitable for finding and exploiting simulations of the earth system.

Box 1: The METAFOR project

"The Common Metadata For Climate Modelling Digital Repositories" (METAFOR http://metaforclimate.eu, 2008-2011) project is a Europe-US collaboration that addressed the problems associated with metadata (data describing data) identification, assessment and usage. This EU funded 2.5 M€ project, which involved 12 institutions, has been led by UK's National Centre for Atmospheric Science (NCAS) at the University of Reading. METAFOR has developed a Common Information Model (CIM, currently at version 1.9) to standardize descriptions of climate data and the models that produce it. METAFOR has secured a mandate from the World Climate Research Programme's Working Group on Coupled Modelling (WGCM) to define and collect model and experimental metadata for the Coupled Model Intercomparison Project Phase 5 (CMIP5) project. METAFOR is taking the first step in doing for climate data what search engines have done for the Internet: putting users of climate data in touch with the information they need. Following the completion of the project in late 2011, the European Commission review characterized METAFOR as "*a very successful project, which should become a blueprint-element for other projects in the data infrastructure domain*". The funding for a continued development of METAFOR activities is now secured under the EU IS-ENES2 project (http://enes.org), starting early 2013.

Box 2: Using the CMIP5 model and simulation description portals

As part of its system to distribute the CMIP5 climate modelling data, the Earth System Grid (ESG) has been the first user interface to include a capability to display the model and simulation documentation (http://www.earthsystemgrid.org/home.htm). This metadata can be found by either selecting "Simulation Metadata" from the Search category or via a pull down menu. Selecting Realm="Earth system" in the faceted search will retrieve the upper-most component (e.g. CESM, Hadgem2) of each of the metadata documents. The user will note that the metadata is divided into tabs. In the current version of the interface, the "Properties" tab contains contact information, simulation information, and the technical properties of the machine the simulation was run on. Information about the CMIP5 protocol can be found under the "Experiment Tab" and similarly information about the model's grid can be found under the "Grids" Tab. Initial conditions and boundary conditions are located under "Inputs" and links to associated datasets are under "Outputs". To delve into the scientific formulation of the model, it is necessary to select a particular model component from the tree on the left and then click on the "Scientific" sub-tab under "Properties".

The CMIP5 metadata information can also be accessed through another portal (http://es-doc.org), which follows more closely the structure of the CIM documents. Here, models, grids or simulations can be explored independently, although this view may confuse the non-experts. As a third illustration, the CMIP5 metadata catalogue is also being used to generate model and experiment description tables in support of the IPCC 5th Assessment Report (scheduled for publication in 2013).

These early uses of the model and simulation description database developed for CMIP5 do not yet enable easy comparison of specific models or simulations, but more tailored applications will appear as the community takes ownership of this new resource and the associated tools. The single entry page http://es-doc.org/cmip5 will point to the different CMIP5 metadata resources as they become available and evolve.

Acknowledgments:

METAFOR has been funded by the EU 7th Framework Programme as an e-infrastructure (project # 211753). We thank Ron Stouffer, Sandrine Bony and Jerry Meehl for the strong support provided via the CMIP and WGCM panels. Mark Elkington from the Met Office/ Hadley Centre provided extensive comments on the beta testing of the CMIP5 metadata questionnaire. We acknowledge the enthusiasm, dedication, inspiration and hard work of the Metafor and Earth System Curator teams and in particular Charlotte Pascoe, Gerry Devine, Hans Ramthun, Allyn Treshansky, Marie-Pierre Moine, Frank Toussain, Rupert Ford and Sophie Valcke.

References:

Carlson, D., 2011: A lesson in sharing. *Nature*, **469**, 293, doi:10.1038/469293a.

Kleiner, K., 2011: Data on demand. *Nature Climate Change* **1**, 10–12. doi:10.1038/nclimate1057.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, doi:10.1029/2007JD008972.

Guilyardi E., V. Balaji, S. Callaghan, C. DeLuca, S. Denvil, R. Ford, M. Lautenschlager, B. Lawrence, L. Steenman-Clark, S. Valcke and the METAFOR group, 2011: The CMIP5 model and simulation documentation: a new standard for climate modelling metadata. *CLIVAR Exchanges*, **56**, 42-46

Hawkins E, Sutton R (2009) The potential to narrow uncertainty in regional climate predictions. Bull Am Meteorol Soc 90:1095–1107. doi: http://dx.doi.org/10.1175/2009BAMS2607.1

Lawrence, B.N., V. Balaji, P. Bentley, S. Callaghan, C. DeLuca, S. Denvil, G. Devine, M. Elkington, R. Ford, E. Guilyardi, M. Lautenschlager, M. Morgan, M.-P. Moine, S. Murphy, C. Pascoe, H. Ramthun, P. Slavin, L. Stenman-Clark, F. Toussaint, A. Treshansky and S. Valcke, 2011: Describing earth system simulations, *Geophys. Mod. Dev.*, *submitted.*

Masson, D. and R. Knutti, 2011: Climate model genealogy. *Geophys. Res. Letts.*, **38 (8),** L08,703.

Overpeck, J.T., G. A. Meehl, S. Bony, and D. R. Easterling, 2011: Climate data challenges in the 21st century. *Science*, **331**, 700-702, doi: 10.1126/science.1197869.

Palmer T., 2012: Towards the Probabilistic Earth-System Simulator: a vision for the future of climate an weather prediction. *To appear in Quart.J.Roy.Meteor. Soc.*

Reichler, T. and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Met. Soc.*, **89 (3)**, 303—312.

Santer, B.D., K.E. Taylor, P.J. Gleckler, C. Bonfils, T.P. Barnett, D.W. Pierce, T.M.L. Wigley, C. Mears, F.J. Wentz, W. Brüggemann, N.P. Gillett, S.A. Klein, S. Solomon, P.A. Stott, and M.F. Wehner, 2009: Incorporating model quality information in climate change detection and attribution studies. *Proc. Nat. Acad. Sci.*, **106**, 14778-14783, doi:10.1073/pnas.0901736106.

Taylor, K.E., R.J. Stouffer, G.A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, doi:10.1175/BAMS-D-11-00094.1, in press.

Williams, Dean N., Bryan N. Lawrence, Michael Lautenschlager, Don Middleton and V. Balaji. 2011: The Earth System Grid Federation: Delivering globally accessible petascale data for CMIP5 in *Proceedings of the 32nd Asia-Pacific Advanced Network Meeting*, New Delhi, August 2011 doi:10.7125/APAN.32.15.