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

- 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 An automatically updated table showing which simulations have been archived by each model;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 .

When dealing with multi-model databases, scientists and other stakeholders are increasingly faced with questions about the suitability of that data for their purposes, a question that was not addressed by these initial documentation efforts. For example, what is the difference between model A and model B? Which simulations of the 20th 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? How does this model conform to the CMIP aerosols protocol? Are volcanoes included and how? The climate modelling community identified early the need for comprehensive and standard metadata for climate modelling to address such questions (as in the European Network for Earth System Modelling, ENES, http://enes.org). The whys and wherefores and issues associated with any particular simulation form the scientist's experiment notebook and sharing this key information widely is also a quality and transparency insurance. Proper and comprehensive climate modelling documentation will further re-enforce the maturity, credibility and openness of our science, under increased pressure from society (Carlson, 2011; Kleiner, 2011).

The EU-funded Metafor project (see Box 1) specifically addresses these challenges. Its central aim is the development of a Common Information Model (CIM) to describe climate data and the models that produce it in a standard way . The CIM is a formal model of the climate modelling process. It includes descriptions of the experiments being undertaken, the simulations being run in support of these experiments, the software models and tools being used to implement the simulations and the data generated by the software. The CIM is organised into two components: one normative artefact the UML (Universal Modelling Language) model called CONCIM or conceptual CIM and a derived XSD/XML generated automatically called the APPCIM, or application CIM. The CONCIM is independent of the application and its concepts are organised into several packages to separate different aspects of the climate modelling process: data, software, activity, grids, quality, shared (Lawrence et al., 2011).

Following this high-level work, Metafor has been charged by the Working Group on Coupled Modelling (WGCM) via the Coupled Model Inter-comparison Project (CMIP) panel to define and collect model and experiment metadata for CMIP5. Integrated in the ESG Federation, the CMIP5 metadata pipeline is described in Figure 1 and summarized below.

Developing and using the CMIP5 metadata questionnaire

The Metafor team has developed a web-based questionnaire to collect information and metadata from the CMIP5 climate modelling groups on the details of the climate models used, how the simulations were carried out, and how the models conformed to the CMIP5 protocol requirements.

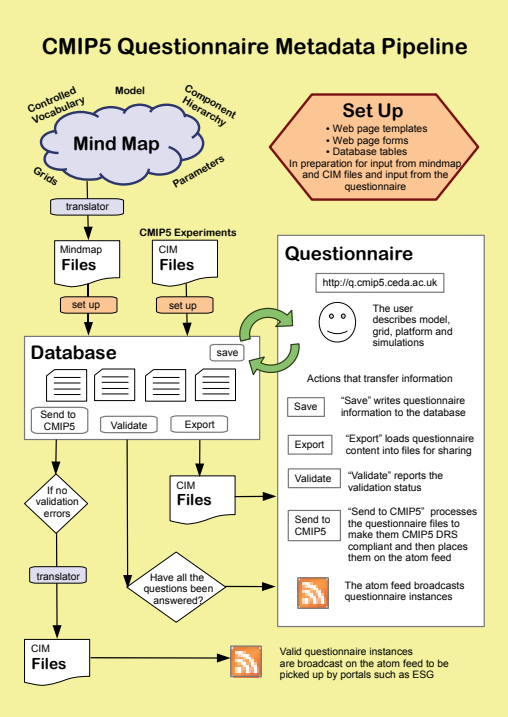

Figure 1. CMIP5 questionnaire metadata pipeline. Interviews with climate scientists helped collect basic information needed to understand models, e.g. structured and controlled vocabulary, captured in mind maps. The mind maps together with the CMIP5 protocol description are automatically transformed into a web questionnaire. Once the questionnaire is completed and validated, instances (CIM files in XML), are broadcasted and harvested by several portals (ESG Gateway, Metafor portal, vERC portal), in which the binding with the CMIP5 data files is made. See also Lawrence et al. (2011).

■ Developing standard model description with the climate modelling community

The content and structure of the model description section of the questionnaire was developed via a series of interviews with numerous climate modellers. The aim of these interviews was to find out the information that scientists need to know to be able to compare climate model simulations. Care was taken not to try to propose standards in areas where there is still active research as community agreed "standards" have yet to emerge. Besides identifying the proper questions, providing standardised responses requires specific knowledge and expertise as well as a wide community perspective. Converging on a first version proved relatively straightforward and debates among experts were easily addressed.

The interviews with domain experts were interactively summarised as mind map diagrams (Figure 2) that allowed the Metafor team to capture both the questions and the standard responses that are referred to as controlled vocabularies (CV, Moine et al., 2011). Symbols on mind map elements indicated how questions should be posed in terms of whether the users should provide one answer or many. The mind maps allowed the Metafor team not only to build up lists of controlled vocabulary, but also to build a structure for the

[1] The Coupled Model Inter-comparison Project, Phase 5

[2] www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php

[3] "A METAFOR for climate change", International Innovation, Environment, October 2010, Research Media Ltd.

way the information about climate models would be collected. Branches in the mind maps were used to illustrate model component hierarchies, and additional formatting was used to distinguish between questions about model components, questions about individual parameters or to indicate where user input should be numeric or text (Figure 2).

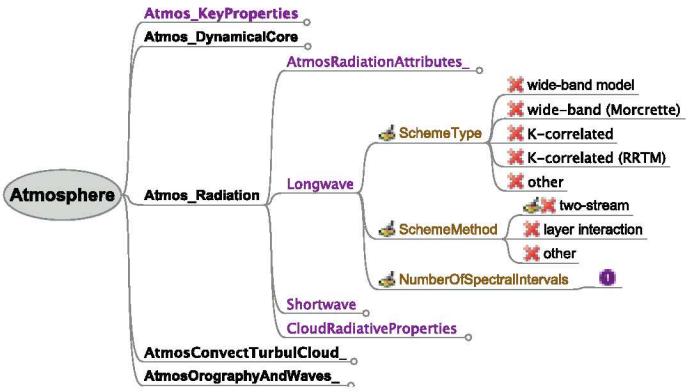

Figure 2. Example mind map for the Atmosphere longwave radiation component. Red crosses indicate that the questions about Scheme Type and Method require only one answer and the number adjacent to NumberOfSpectralIntervals indicates that numeric user input is required.

The intuitive format of the mind map diagrams enabled the scientists interviewed to give direct feedback about both the structure and content of the questionnaire and this feedback could be integrated quickly without exposing any of the questionnaire code. The current mind maps mapping the 8 realms defined by CMIP5 can be found and interactively explored under http://metaforclimate.eu/trac/browser/controlled_vocabularies/trunk/Software. This first attempt to comprehensively describe the science of an Earth System Model (it include more than 550 properties) is a unique community resource that can also be used for educational and training purposes.

Building the CMIP5 questionnaire

The mind maps, together with the precise CMIP5 experiment protocol description provided directly by the CMIP panel, were integrated into a web-based questionnaire. Automatic python parsing tools rendered the mind maps structure, questions and controlled vocabulary directly into the questionnaire (Figure 3) clearly separating climate science and IT concerns. The branching structure of the mind maps generated the hierarchy of model components, each with an associated web form. The branching structure also drove the tree navigator (to the left in Figure 3), which allows users to navigate directly to a particular model component. The controlled vocabulary captured in the mind maps generated the questions about the model components and also populated the drop-down lists of standardised responses for each web form. Attached notes in the mind maps appear as explanation tool tips in the questionnaire.

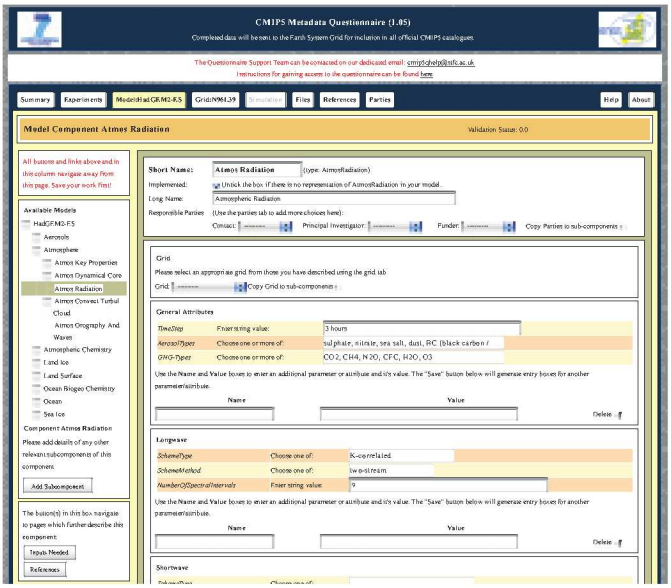

Figure 3: Screen shot of the CMIP5 questionnaire. This screen shot shows part of the entry form for describing the longwave atmosphere radiation scheme; it is generated from the mind map shown in Figure 2.

The questionnaire also allows users to enter descriptions of components that are not covered by the mind maps (see "blank" forms in Figure 3). The mind map controlled sections of the questionnaire ensure that a standardised set of metadata about each of the CMIP5 model is collected. However the questionnaire is flexible enough to allow users to describe their models in more detail if they wish. Additional terms entered by users will inform the future externally governed controlled vocabularies used by the Metafor Common Information Model (CIM).

Detailed technical information about the questionnaire and its implementation can be found in the questionnaire help documentation, in the Metafor document repository (http://metaforclimate.eu/Documents.htm) and in Moine et al. (2011). The separation of concerns described above, coupled with the generic implementation of the questionnaire as a whole, allows the questionnaire to be 'specialised' for other metadata collection projects through the supply of different controlled vocabularies, as currently developed within the Metafor and IS-ENES European projects for non-CMIP5 applications.

The CMIP5 metadata questionnaire was launched in Nov 2010 (http://q.cmip5.ceda.ac.uk), and is now in use by most of the CMIP5 modelling centres. Box 2 presents a short introduction to questionnaire use. The process to gather the required information represents a significant investment from modelling groups. First experience by several groups indicates that several weeks of interviews of many experts are likely needed, even though the process of filling up the questionnaire once that information is obtain is relatively straightforward. This information will represent the public documentation of the models and simulations provided by the modelling groups to the wider community and stakeholders. To ensure this metadata is provided in time for the analysis stage of CMIP5, Metafor offers comprehensive user support. Help systems and documentation have been developed by a dedicated team to support the users of the questionnaire. These include a dedicated email address solely for questionnaire issues (cmip5qhelp@stfc.ac.uk) and webcasts and interactive web seminars to publicise and train users of the questionnaire. A CMIP5 Questionnaire helpdesk handles all queries relating to the metadata requirements for CMIP5 and ensures replies within two working days.
Once a questionnaire instance has been completed, it is validated against a set of validation rules. The first of these is to ensure completeness of the information so that a comprehensive description is provided, while the second is to ensure consistency between related elements of metadata so that this description is meaningful. Validation may be performed at any point during the completion of a questionnaire and provides the user with an indication of the extent to which the metadata provided constitutes a valid metadata record, and a guide as to how much more information will be required before this is the case.

Once questionnaire instances have been validated into CIM XML standard instances, they are made freely available on the questionnaire atom feeds (Figure 1). The content of the questionnaire instances will be hosted and displayed in the ESG Gateway hosted by the Program for Climate Model Diagnosis and Intercomparison (http://pcmdi3.llnl.gov/esgcet/). The Curator project has worked closely with the ESG team and METAFOR to develop a metadata display for ESG, and to complete a metadata pipeline that takes questionnaire output and propagates it through the PCMDI or other (ENES's vERC, Metafor) portals (Figure 1).

Looking ahead

This first comprehensive metadata collection for climate modelling is an ambitious undertaking by the community and, used for CMIP5, will provide the most comprehensive metadata of any climate model inter-comparison project. Because it is a pilot project, many aspects will need to be revisited after this first experience, coupled with the need for a governance structure to both maintain and develop the CIM and the associated controlled vocabularies. Discussions are underway on how to best organise this important legacy of the EU Metafor and US Curator projects. Looking beyond CMIP5, the CIM and the associated standards have the ambition to become more ingrained within modelling groups (as with netCDF/CF) as a means of automatic documenting of model configurations and simulation runs (as currently planned by the Hadley Centre, NCAR, IPSL and other modelling.

Acknowledgements

METAFOR is funded by the EU 7th Framework Programme as an e-infrastructure (project # 211753). We thank Karl Taylor, Ron Stouffer, Sandrine Bony and Gerald 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. Charlotte Pascoe, Gerry Devine, Allyn Treshansky and Marie-Pierre Moine provided figures and material for this article.

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.

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.

Moine, M.P., C. Pascoe, A. Alias, V. Balaji, P. Bentley, G. Devine, R. Ford, E. Guilyardi, B. N. Lawrence, S. Valcke, 2011: Development and Exploitation of a Controlled Vocabulary in support of Climate Modelling, IEEE, submitted.

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, IEEE, submitted.

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: The CMIP5 Experiment Design. Bull. Amer. Meteorol. Soc., submitted.

Box 1: The METAFOR project

"The Common Metadata For Climate Modelling Digital Repositories" (METAFOR http://metaforclimate.eu, 2008-2011) is a Europe-US collaboration project that seeks to address the problems associated with metadata (data describing data) identification, assessment and usage. This 2.5 M€ project, which groups 12 institutions, is led by Prof. Eric Guilyardi from NCAS-Climate/University of Reading and managed by Dr. Sarah Callaghan from BADC. Metafor has developed a Common Information Model (CIM, currently at version 1.5) to standardise 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.

Box 2: Filling up the CMIP5 metadata questionnaire: a user perspective

Charlotte Pascoe and Gerry Devine, in charge of the CMIP5 questionnaire user support group.

The CMIP5 metadata questionnaire can be accessed at http://q.cmip5.ceda.ac.uk. Although the different sections of the questionnaire can, to some degree, be completed in any order, following a suggested route can significantly reduce the time needed. Initially users are advised to complete their range of auxiliary information, namely references (publications, web pages etc), files (that have been used as inputs to models for example), and details of those responsible parties, whether an institution or individual scientists, involved in the centre's CMIP5 simulations. Having this information completed prior to filling out the more complex sections of the questionnaire means that this information is on hand to attach directly to, for example, the different component sections of the model.

Having completed the auxiliary information, it is then suggested that users complete the descriptions of the different grids that they have used as well as the computing platforms on which their simulations have been deployed. The next step is to complete the description of the climate model itself and, naturally, is where the largest investment of time will occur. Within the model section of the questionnaire, the users will be able to navigate the different components using the navigation tree on the left-hand panel. Users are free to fill out the details of each component in any order they see fit and will in general, for each component, be asked to provide some high-level information, name, description, references etc, more intricate questions about the properties of each component (driven primarily by the mind maps), and details of how this component is traditionally coupled to other components. There are currently 8 top-level 'realm' components each of which has on average approximately 6 or 7 sub-components.

The final stage of the questionnaire is to complete the information about the climate simulation itself. To do so, it is required that the model and platform description have already been initiated. In the simulation section, the user will fill out the 'specifics' of the modelling workflow, e.g. the particular CMIP5 experiment that the model was run, details of how long, or over what time period, the model was run for, any configured model settings imposed for this particular model run, as well as giving details of how the simulation conformed to those requirements that the CMIP5 experiment requested.

At any stage of the process, the user can return to a 'summary' page that details all the grids, platforms, models, and simulations that are currently being documented for that particular centre. From this same page, a user can create a duplicate copy of, for example, a previously completed grid, to act as a starting point for a new, but similar in nature, grid description.

The questionnaire has a "Test centre" area where users can experiment before filling out information in their own respective centre pages, and a read-only "Example centre" which gives examples of the sorts of information that is expected. The Test and Example centres are freely accessible but only those users who have an OpenID issued by an ESG Federation OpenID provider can request access to individual modelling centre pages. The Example centre contains a read-only example of elements of the questionnaire (kindly provided by the UK Met Office Hadley Centre which already completed the description of its models and several experiments).

Satellite Observations for CMIP5 Simulations

Joao Teixeira[1], Duane Waliser[1], Robert Ferraro[1], Peter Gleckler[2], Gerald Potter[3] and colleagues

1 Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA
2 Program on Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, California, USA
3 NASA Goddard Space Flight Center, Greenbelt, Maryland, USA

Summary

The objective of this project is to provide the community of researchers that will access and evaluate the CMIP5 climate model results access to analogous sets (in terms of variables, temporal and spatial frequency, and periods) of satellite observational data. This activity is being carried out in close coordination with corresponding CMIP5 modeling activities and directly engages the observational (e.g. mission and instrument) science teams to facilitate production of the corresponding data sets and associated documentation.

Background

Observations play an essential role in the development and evaluation of climate modeling systems. In particular, observations from satellite platforms often provide a global depiction of the climate system that is uniquely suited for these purposes.

The goal of this project, funded by the National Aeronautics and Space Administration (NASA) and the Department of Energy (DOE), is to provide selected satellite observations for the diverse research that will result from the 5th phase of the World Climate Research Programme's Coupled Model Intercomparison Project (CMIP5). This standard experimental protocol facilitates the community-based study of coupled earth system model simulations, and is expected to be a centralizing resource for the upcoming 5th Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR5).

Taylor et al (2009) describe in detail the protocol for CMIP5, which defines the scope of the simulations that will be undertaken by the participating modeling groups. For several of the prescribed retrospective simulations (e.g, decadal hindcasts, AMIP and 20th Century coupled simulations), observational data sets can be used to evaluate and diagnose the simulation outputs.

However, the pertinent observational data sets to perform these particular evaluations have not been optimally identified and coordinated to readily enable their use in the context of CMIP5

Main Tasks

Given the importance of the observations to the assessment process, along with the range and complexity of the observational datasets needed for a robust assessment, a simple framework to identify, organize and disseminate them for CMIP5 is currently underway in this project.