Proceedings of the 32nd Asia-Pacific Advanced Network Meeting

The Earth System Grid Federation: Delivering globally accessible petascale data for CMIP5

Dean N. Williams[1], Bryan N. Lawrence[2], Michael Lautenschlager[3], Don Middleton[4] , V. Balaji[5,]*

1 Program for Climate Model Diagnosis and Intercomparison , Lawrence Livermore National Laboratory, U.S.
2 NCAS/British Atmospheric Data Centre, STFC Rutherford Appleton Laboratory, U.K.
3 DKRZ: Deutshes Klimarechenzentrum, Germany
4 National Center for Atmospheric Research, U.S.
5 Princeton University, USA,

E-Mails: williams13@llnl.gov, bryan.lawrence@stfc.ac.uk, (lautenschlager@dkrz.de, don@ucar.edu, balaji@princeton.edu

* Tel +1-609-452-6516

DOI http://dx.doi.org/10.7125/APAN.32.15

Abstract: The fifth Coupled Model Intercomparison Project (CMIP5) will involve the global production and analysis of petabytes of data. The Program for Climate Model Diagnosis and Intercomparison (PCMDI), with responsibility for archival for CMIP5, has established the global "Earth System Grid Federation" (ESGF) of data producers and data archives to support CMIP5. ESGF will provide a set of globally synchronised views of globally distributed data – including some large cache replicants which will be persisted for (at least) decades. Here we describe the archive requirements and key aspects of the resulting architecture. ESGF will stress international networks, as well as the data archives themselves – but significantly less than would have been the case of a centralised archive. Developing and deploying the ESGF has exploited good will and best efforts, but future developments are likely to require more formalised architecture and management.

Keywords: keyword; keyword; keyword.

1. Introduction

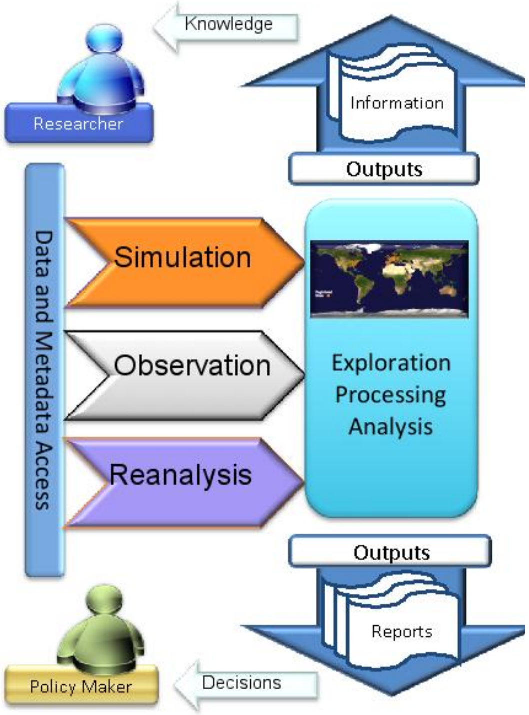

Figure 1. The climate community is making revolutionary changes in data integration and exploration. ESGF integrates heterogeneous data and metadata sources (i.e., simulation, observation and reanalysis) into a common infrastructure, improving access for both scientists and non-scientists.

The Earth system modeling community challenges itself by carrying out large, globally coordinated, model intercomparison projects. These projects were originally designed to evaluate the state of the art in earth system modeling, but the third Coupled Model Intercomparison Project CMIP3 provided an "ensemble of opportunity" [1] heavily used for the fourth assessment report (AR4) of the Intergovernmental Panel for Climate Change (IPCC). Accordingly, much of the global community now sees these Model Intercomparison Projects (or "MIPS") as primarily for doing projections, even as many in the modeling community see them primarily as tools for improving the ability to make projections. As a consequence the active fifth Coupled Model Intercomparison Project (CMIP5) incorporates many more numerical experiments than previous MIPS, covering the requirements of multiple communities. This increase in experiments, coupled with bigger and more diverse user communities, and increased volumes of output, has meant that the previous methodologies for handling MIP data distribution are simply not practical. In this paper we present a description of the Earth System Grid Federation (ESGF), a global consortium of data providers and data archives, aimed at solving this problem.

Led by the Program for Climate Diagnosis and Intercomparison (PCMDI, at the U.S. Lawrence Livermore National Laboratory), the successful delivery of the CMIP3 archive for the World Climate Research Programme was one of the reasons why model data analyses became such an integral part of the IPCC AR4. However, the scale of the CMIP3 archive will be dwarfed by that required for CMIP5. As a consequence, PCMDI initiated, via the Global Organisation for Earth System Science Portals (GO-ESSP), the establishment of a global federation to provide data archival and access for CMIP5. This culminated in late 2009 with the formation of the ESGF, which essentially consists of a club of data providers – mostly modeling groups, but not exclusively – along with some major data archive centres who are providing what is effectively a set of global caches of important data. Some of those data archives are also committed to persisting that data for the long term, i.e. they are providing curation functions. CMIP5 drives ESGF and is the primary motivator for additional collaborators to add other climate data sets (including ground based observational, reanalysis, and earth-observation data sets) to the ESGF. The only entry criteria to joining the ESGF is the ability to expose data using the ESGF software. Current ESGF members are distributed across Europe, the Americas, Asia and Australia.

In the remainder of this paper, we describe the requirements which have driven the construction of the ESGF and the resulting architecture. We describe the key challenges around delivering the architecture. It will be seen that the ESGF is an incipient activity which has aggressive timescales to deliver globally accessible, and performant, petascale data services. As such the successful delivery of ESGF requires major computational efforts which stress international network infrastructure and will eventually require significant coordination, beyond the available "best efforts" basis.

2. Requirements

The sequence of events around climate prediction are described in Figure 1: data and metadata access are integral in the exploitation of simulations, observations and reanalyses. With the right tools in place, data exploration, post processing and analysis yield new physical insights (information and knowledge) for the research community, even as they can be used to generate reports leading to decisions and policy. However, this depiction is deceptively simple: in practice both the research and policy communities are diverse, with differing requirements on the tooling for exploration, process and analysis, and even differing requirements on the underlying data.

Some of this diversity can be seen in the range of experiments being supported by CMIP5; three major categories are being undertaken:

1. Projection experiments, with a range of characteristics and inputs, aimed at providing the best available climate projections on a centennial scale,
2. Evaluation experiments, aimed at understanding processes within models, and specific capabilities (including ability to simulate paleoclimate), and
3. Decadal predictions (mostly as hindcasts) where the aim is mostly to evaluate the state of the art, and identify how to make improvements, rather than provide predictions for a wider community.

The full set of experiments [2] includes dozens of specific experiments across these categories, and so diverse users are expected, from the entire climate impacts community in the first category, to primarily the physical climate sciences community in the second, and a mixture of both in the third – as well as those in the incipient climate services industry and their clients (including insurance, food production etc) – all of whom are interested in evaluating capability as much as they are in predictions per se. Not surprisingly, this diversity leads to a diversity of data management requirements.

The CMIP5 modeling community includes in excess of twenty modeling groups, most of whom are running more than one major model configuration, with varying parameters, at a range of resolutions for various durations. While the CMIP5 data protocol (available at the CMIP5 web site, http://cmip-pcmdi.llnl.gov/cmip5) defines requested model output and an accompanying output format, the wider CMIP5 data provision includes earth observation data, and data written on complex model grids. The use of some data is restricted to specific conditions, and some is effectively completely open access, with the choice down to the original data provider. Surveys of modeling groups suggest that the total volume of requested output produced will be around 3.3 petabytes – and we expect modeling groups to write many more petatabyes of data beyond that requested (estimates range to in excess of 10 PB).

CMIP3 produced a central archive of 35TB, which resulted in over a petabyte of downloads by thousands of users over six years or so. The increased volume from CMIP5 arises for a number of reasons, including greater model complexity, higher resolution, and a variety of initializations used to produce ensembles of simulations. For some communities, CMIP3 data volumes are still challenging as are the file formats; for them, CMIP5 would be completely inaccessible without more sophisticated interfaces.

Timescales will be an issue; while it is not yet clear when all the requested data will be available, for papers to be assessed in the IPCC 5th Assessment Report (AR5), they will need to be submitted by mid 2012. Even if all the data were available by mid 2011 (unlikely), and allowing six months for data analysis, that suggests a user data acquisition period of at most six months. Even those who have sophisticated data analysis systems face problems (for example, the estimated volume of requested ocean fields for the decadal experiments is about 45TB – which for many has to be acquired, stored, and used, within that AR5 timeframe). These acquisition problems rebound on the data archives: if a centralised archive were to try and provide access to all of the 3.3 Petabytes of data for one hundred different users in six months, they would need to sustain approximately 160 gigabits per second of data delivery for that entire time! While CMIP5 usage is not easy to predict, extrapolation from CMIP3 along with a larger expected user community, suggests these could be conservative requirements in terms of downloads; clearly then, a centralised solution cannot suffice for CMIP5 (or any future MIP).

One of the reasons for such high download volumes for CMIP3 was the difficulty of getting subsets of appropriate data in time and space, so many users downloaded much more data than they needed. Many users also calculated the same higher order products (statistics etc). Often they downloaded more than they needed since the available metadata was not extensive. These experiences mean that ESGF needs to provide sophisticated interfaces to do server side calculations and visualizations, and deploy a number of methods to mitigate against high volume downloads by all users: provide subsetting tools, pre-calculating key statistics, provide better model and simulation metadata, and replicate data so as to support parallelization (on a global scale).

Experience tells us that the modeling groups will provide data, then (themselves, or others) realise that there are problems with that data – and subsequently provide replacement versions. Such problems are mostly not scientific in origin, and result from manual handling of massive quantities of data under aggressive time pressure, leading to incorrect labeling, inappropriate scaling and a host of easily identified (and fixed) problems when the data are finally looked at. Previous archives have not always been careful about version control in such situations, and it has not always been clear which data has eventually been used. In the case of CMIP3 and AR4 an additional problem has been that it was not clear which data was in the CMIP3 archive at any time, and so in some cases where users have used "all the CMIP3 models" or "all the AR4 models" it is non-trivial to go back and be sure what data has been used. For CMIP5 there has been a concerted effort to make sure that version control has been designed into the solution so that these problems can be avoided.

On the archive side, we have three further requirements: all services need to respect the license requirements of the underlying data, we need to protect against malicious attack aimed at data damage, and we need to protect against intended or unintended denial-of-service (the latter would occur when attempted downloads exceed the archive capacity).

As it is clear that a distributed solution is required, that too brings requirements of its own: data needs to be discoverable, wherever it is. Data needs to be replicated, and the individual replicants need to be discoverable and distinguishable in their own right. Users need to be sure that all replicants are identical. Finally, because data providers need evidence of use; logging, notification, and citation are all necessary, so that wherever data are obtained, originators and service providers can gain credit. Data citations (including interfaces and the data itself) need to be robust beyond the expected life times of much of the software infrastructure.

3. The ESGF Architecture

ESGF was born out of a number of initiatives to handle diverse, distributed data access for the climate community: In the U.S., the "Earth System Grid" (ESG, [3]), in the UK, "The NERC DataGrid" [4] , and in Germany, the Collaborative Climate Community Data Processing Grid (C3-Grid [5]) . However, the

Dominant contribution has been that of the ESG. As a consequence, the ESGF architecture is curently a more mature version of the original ESG, extended and modified by both the code and experiences of the other partners.

There are five key information classes which underpin the ESGF: the data itself; the "data metadata" which exists within the data files (both described on the CMIP5 website); the "model and experiment metadata" created externally and ingested into the ESGF system [6]; the "quality metadata" (which describes intrinsic checks on data fidelity rather than the extrinsic scientific quality, [7]); and “federation metadata” (to support user management and system deployment, [8]).

SGF exploits this information using four major components: data nodes, gateways, federation metadata services (to support authentication and authorization), and data services to be deployed adjacent (or on) the data nodes. Key relationships between data services, nodes and gateways are shown in Figure 2. Data is exposed to the federation by data nodes – there are likely to be many data nodes. Each data node provides a THREDDS (Thematic Realtime Environmental Distributed Data Services, [9]) catalog interface (itself not protected by any security middleware) and a number of data interfaces (each protected by security middleware). The security middleware [8], heavily informed by the NERC DataGrid experience, is not discussed in detail further here, except to note that it requires user management (for registration, and to assign authorization credentials) and the deployment of associated policy enforcement points.

Three key data interfaces include (1) a Browse interface to provide a hierarchical view of the data, (2) an OpeNDAP [10] interface to provide subsetting facilities and direct programmatic access to the data archive, and (3) a GridFTP [11] service to provide high bandwidth data download. Additional interfaces shown are a Product interface, allowing more access to more sophisticated data services – primarily the Live Access Service [12] and a Management interface to allow the extraction of logging information etc. It is likely that further data services will be deployed with the data nodes as the ESGF matures.

Data and data metadata are ingested into the data node via the Publisher component. This component parses the data files to populate the node database, which itself populates the THREDDS catalogue. Data nodes may support deep archives but are expected to hold most data on disk.

It is expected that most modeling groups will deploy a data node; those that do not will send data to another site to expose via their data node(s). Those that deploy their own nodes may well expose considerably more than just the requested data – additions might include extra ensemble variables, more variables, or higher temporal resolution output. Additional data nodes are also being deployed into ESGF by other communities, one example being the provision of validation and evaluation data from earth observation.

Each data node publishes information about the data contents of the node to a gateway via the publishing API. There are expected to be at least eight gateways, each of which shares information about the data holdings of its associated data nodes by the Open Archives Initiative Protocol for Metadata Harvesting (OAI/PMH). Each gateway can also ingest model metadata via the polling of an Atom (RFC4287) feed from the Metafor questionnaire (see [6]). Some gateways support user registration, but this can also be handled out-of-band, since the federation uses a whitelist of OpenID providers for authentication (see [8]). The gateways provide a semantic search facility as well as hierarchical browse. The former is based on the parsing of the THREDDS catalogue information into an OWL ontology and subsequent ingestion into a triplestore (it is these triples which are exchanged by OAI/PMH). The gateways will also provide the ability to generate editable WGET scripts for customising data downloads.

Three of the gateways hold special status, as they are associated with a set of three data nodes which will attempt to provide long term persistence of a cache of replicated data for the IPCC data distribution centre; the three are PCMDI, the British Atmospheric Data Centre (BADC), and the German Climate Computing Centre, DKRZ. This replicated data will not be the entire 3.3 PB of requested data, and has been chosen to be what the community expect to be the most heavily used. Estimates of the appropriate cache volume have fluctuated as modelling groups have declared their intentions, but it is expected to be near 1.5 Petabytes.

Some other gateways are expected to be associated with cache replicates, and may well persist them long term, but they have no obligation to do so. The replicants, along with the originating data nodes with the remainder of the data, are key to delivering both the parallelisation of access to the ESGF data resources, and the global distribution of high bandwidth access. Currently there are expected to be cache replicants in Australia, Japan, Europe, and the U.S. Data which has passed the appropriate quality checks (see below) and held in the appropriate part of the archives will be identified and listed in a “replication manifest”, and these manifests will be used to drive replication aimed at global synchronization.

4. Current Status and Deployment Challenges

There are three major challenges facing ESGF:

1. How to initialise the data holdings of the petascale replicants from the data holdings at the originating modeling centres, and keep them synchronized (this too challenges networks; at 1 Gb/s it will take approximately 100 days to move 1PB);
2. How to quality control the data as it enters ESGF to avoid both the expensive global replication of “incorrect” data, and the entry of such data into the scientific ecosystem;
3. How to ensure that the software underlying the ESGF can be deployed easily, and evolve alongside other activities where it is deployed.

The status of ESGF as of April 2011 is that there are several gateways and data nodes operational, but very little CMIP5 model data is in the system. However, it is expected that the bulk of the data will be provided during the remainder of 2011, which means that the replicants are expected to be populated to petascale during the remainder of 2011.

There are several possible mechanisms for populating and continuing the synchronization of the replicates: we could rely on couriers, and cycle TB-scale disks around the world; we could rely on the existing academic networks, or we could set up our own dedicated network links. The current plan is a mixture of both the first two options: where possible we will exploit the existing networks, tests suggest that multi-Gb/s bandwidth can be delivered between the australian node (at the Australian National University) and PCMDI, and near 1 Gb/s between BADC and PCMDI. (It has not been trivial for the ESGF community to get such high bandwidths in place, subtle issues with router and system configuration have meant considerable work at most sites.) In principle we can move up to 10 TB per day at 1 Gb/s, which is likely to suffice, since we don't expect data to arrive all at once. If it does, or if the networks into or out of some of the other replicant sites cannot cope, then we will resort to physical disks – but this will be the last resort, since the manual handling involved will be onerous, lead to errors, and likely to introduce complications into the replication scheduling.

Along with user downloads, these synchronization data flows will stress both networks and I/O at the archive centres: PCMDI in the U.S. and BADC in Europe are both expecting to have to handle multiple synchronization datastreams as well as significant user downloads. Peak loads are expected to be filling the 10 Gbit/s wide area network capacity currently available at the BADC, and PCMDI (which expects an even higher load) has configured two 10 Gb/s links, and is moving towards 100 Gb/s in 2012 and projecting 1Tbp/s in 2015.

Data and metadata is being quality controlled at a number of levels: the first level of quality control carried out during the initial publication process is effectively syntactic; are all the correct attributes present and using appropriate vocabularies? The second level of data quality control will test that data falls within expected extremes and produce plots that can be eyeball sampled to pick up unphysical discontinuities (as might happen if data from the wrong variable was inadvertently written into the wrong output stream). Second level metadata quality control will be carried out by the Metafor team [6]. A third level of quality control will result in more stringent manual investigation of output, and some feedback between the archive teams and the modelling teams, to result in the formal publication of datasets via the World Data Centre for Climate at DKRZ [7].

One of the most difficult problems facing ESGF is the deployability and evolvability of the software infrastructure. When an archive is in one institution, choices and compromises can be made within one management domain. Clearly petascale resources cannot be deployed identically at each institution: there will be existing site software policies, expertise, and infrastructure with which the new archives will have to “play nicely”. Each institution will have it's own threshold for, and mechanisms for, providing redundancy and/or high availability. All institutions are involved in multiple projects, each running with their own timescales.

The ESGF solution will be to develop independent components which provide defined interfaces, and to plan on evolving around those interfaces – with deployment using standard components suitable for deployment in a range of environments. However, there is a tension between providing “easy-to-install” scripts and the necessity for flexibility of implementation (for example, BADC deploys a set of database machines which are configured for high availability – but the default data node installer expects the database on the same node as the THREDDS server). These issues are encountered day-to-day in an ad-hoc manner, since there is no central resourcing for architectural design, nor a formal governance procedure for agreeing on changes. Decisions on how to proceed are generally made at weekly telecons via consensus, but all partners recognize the need for more formal project management and governance, particularly as the architecture is now moving away from being based on ESG alone. While GO-ESSP has provided an umbrella for the activity, none of the institutions involved has funding available to take on architecture and project management for a global federation – and no other mechanism has yet been established to generate the funding.

5. Future Work

In the next twelve months, the ESGF will be concentrating on CMIP5 support, improving and hardening the performance of the components deployed and ensuring operational user support is efficient and effective. Improvements in metrics and user notification associated with data changes will be incorporated in the systems. The underlying software of ESGF is then expected to evolve further, with more modularization, and clearer well documented interfaces. Improved and extended services are expected on the data nodes, targeted particularly at the less experienced users of climate data.

From a data perspective, new observational data will be expected, and new model simulations from a major international initiative to evaluate regional climate modeling will likely be acquired. Archive scales are expected to reach exascale within the decade.

Acknowledgements

The ESGF is supported by the hard work of many individuals, in many countries – too many to name here. It is also supported by a variety of funders in many countries. However, since the current ESGF is heavily based on the Earth System Grid as well as the NERC Data Grid and C3-Grid, we pay particular acknowledgment to Department of Energy in the U.S., the Natural Environment Research Council in the U.K., and the German Federal Ministry of Education and Research.

References

1. Tebaldi, C.; R. Knutti. The use of the multi-model ensemble in probalistic climate projections. Phil. Trans. Roy. Soc., 2007, 365, 2053-2075.
2. Taylor, K. E; R. J. Stoufer; G.A Meehl. A summary of the CMIP5 Experimental Design. 2011. Available at http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf
3. Williams, D.; Coauthors. The Earth System Grid: Enabling Access to Multimodel Climate Simulation Data. Bull. Am. Met. Soc., 2007, 90, 195-205.

4. Lawrence, B.N.; R. Lowry; P. Miller, H. Snaith; A. Woolf. Information in Environmental DataGrids. Phil. Trans. Roy. Soc., 2009, 367, 1003-1014.
5. Kinderman, S.; M. Stockhause; K. Ronneberger. Intelligent Data Networking for the Earth System Science Community in German e-Science Conference, 2007. Available at http://edoc.mpg.de/get.epl?fid=36067&did=316512&ver=0
6. Lawrence, B. N.; Coauthors. Describing Earth System Simulations. 2011.
7. Stockhause,M.; H. Höck, M. Lautenschlager; F. Toussaint. The Quality Assessment Concept and its Application to CMIP5 at the World Data Centre for Climate. 2011.
8. Kershaw, Philip; Rachana Ananthakrishnan; Luca Cinquini; Dennis Heimbigner; Bryan Lawrence. A Modular Access Control Architecture for the Earth System Grid Federation Submitted to GCA2011.
9. Domenico, Ben; John Caron; Ethan Davis; Robb Kambic; Stefano Nativi. Thematic Real-time Environmental Distributed Data Services (THREDDS): Incorporating Interactive Analysis Tools into NSDL. J. Dig. Inf., 2, 2002. Available at http://journals.tdl.org/jodi/article/view/51.
10. Cornillon, P.; J. Gallagher; T. Sgouros. OPeNDAP: Accessing data in a distributed, heterogeneous environment. Data Science J. 2003, 2, 164-174.
11. Allcock, W.; J. Bresnahan; R. Kettimuthu; M. Link; C. Dumitrescu; I. Raicu; I. Foster. The Globus Striped GridFTP Framework and Server In SC ’05: Proceedings of the 2005 ACM/IEEE conference on Supercomputing, 2005.
12. Hankin, S.; Callahan, J.; Sirott, J.. The Live Access Server and DODS: Web visualization and data fusion for distributed holdings. 17th Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, American Meteorological Society, Albuquerque, NM, 2001.

© 2011 by the authors; licensee Asia Pacific Advanced Network. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).