Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-52-SC1, 2018

Geoscientific Model Development Discussions

Interactive comment on "Requirements for a global data infrastructure in support of CMIP6" *by* Venkatramani Balaji et al.

R. Abernathey

rpa@ldeo.columbia.edu

Received and published: 4 April 2018

Authors: Ryan Abernathey, Naomi Henderson (Lamont Doherty Earth Observatory of Columbia University), Niall H Robinson, Jacob Tomlinson (Informatics Lab, Met Office, Exeter), Kevin Paul, Joseph Hamman (National Center for Atmospheric Research), Jiawei Zhuang (School of Engineering and Applied Sciences, Harvard University), Daniel Rothenberg (ClimaCell, Boston, MA), Matthew Rocklin (Anaconda Inc)...all on behalf of the **Pangeo Project** (https://pangeo-data.github.io/)

We commend the WIP for the rigorous and thoughtful assessment of the global data infrastructure needed to support CMIP6 and beyond. This paper identifies many important challenges related to CMIP data replication, provenance, and scientific repro-

ducibility. Absent, however, is a discussion of the computational challenges associated with the analysis of CMIP datasets and the relationship between data archives and computing resources. Our overall recommendation for revising the paper is to give more attention to this important question. The authors of this comment believe that enabling efficient, accessible, scalable computation on CMIP data should inform the design of the global infrastructure. Instead of encouraging users to download the data to their local systems, we should be encouraging users to bring their computing to the data. This can be achieved by working more closely with national computing centers and by placing CMIP data in cloud storage, where it is directly accessible to distributed computing.

As recognized in the manuscript, many of the most valuable science results from the CMIP project come from global comparisons across many models, scenarios, and ensemble members. To obtain these results, scientists must run analysis on significant fractions of the multi-petabyte CMIP archives. As anyone who performs such calculations knows, they rarely work on the first try–interactive exploration and visualization of the data is a crucial part of the scientific process. However, the computing systems deployed for the analysis of CMIP data generally fall far short of producing interactive speeds; instead researchers wait for weeks to test new ideas (we know this from personal experience). Most of these computing systems are what the manuscript calls "dark repositories," mirrors of CMIP data on servers and computing clusters owned and managed by individual research groups. In addition to disrupting the chain of tracking, provenance, and curation (as discussed in the manuscript), dark repositories are potentially financially wasteful, since the data is transmitted and duplicated over and over just for the purpose of exposing it to computation. Scientists must make an up-front judgement on which fractions they wish to mirror; they may not even use everything they download. In addition, such a priori decisions create an insidious pressure to look for "things you expect to see, in places you expect to see them."

These dark repositories are ultimately funded by agencies such as the US National

Science Foundation and its international counterparts, via equipment purchases and technical support staff salaries. No one really knows how many dark repositories there are and how much they cost in aggregate. Despite the prevalence of dark repositories, users are probably frustrated with their performance on terabyte-scale, let alone petabyte-scale calculations.

A key technical consideration is that, on standard servers and workstations, most CMIP-style data analysis is heavily I/O bound rather than compute bound i.e. it is limited strongly by the rate at which data can be read from storage. Fortunately, more scalable ways for climate scientists to interact with large datasets are starting to emerge. Intelligent subsetting and lazy loading can circumvent the need for bulk downloads. Furthermore, when such datasets are placed on distributed storage attached directly to distributed computing, the time-to-result for a given analysis can be reduced by orders of magnitude, ultimately resulting in faster scientific progress. NCAR's CMIP analysis platform is a good example, with CMIP data stored on GLADE (Globally Accessible Data Environment), a high performance parallel filesystem accessible from the compute nodes of the Cheyenne supercomputer. Users with access to this platform are much less likely to want to create their own dark repositories, since they enjoy the combination of high performance computation and comprehensive data access. Although storage on GLADE is expensive compared to a single dark repository, it's probably cheaper than ten dark repositories in aggregate.

While traditional supercomputers can meet some of the data-analysis needs of CMIP users, they were not designed for this purpose and are probably overkill for it. We believe that an ideal data analytics system for these problems has the following properties:

1. Low administrative hurdles to sign up and log in, even for new, junior, or industry users
2. Easy web access for popular interactive environments like Jupyter notebooks

3. Easy web access on the open internet for automated web services and mobile apps
4. Dynamic and immediate allocation of interactive compute resources at modest sizes (hundreds rather than millions of cores) even if those sessions may have to grow or shrink during the allocation, depending on external use
5. Cheap costs, sacrificing the high performance network and rich CPU/Memory ratio of super-computing centers, and replacing them with commodity networking and locally attached storage
6. Co-location with the relevant datasets

Data analytics clusters are growing within existing computing facilities today that have some (but rarely all) of the properties above. Cloud computing, however, is ideally suited to the storage, processing, and distribution of extremely large, shared datasets today. Both, government-sponsored cloud-style data centers, and the commercial cloud (e.g. Amazon Web Services, Google Cloud Platform, Microsoft Azure, etc.) merit consideration. Data stored in cloud storage is directly accessible from cloud computing instances within the same network, providing effectively infinite data bandwidth to distributed processing systems. In this paradigm, no data needs to be downloaded at all; if the CMIP data were already in cloud storage, users would pay only for the compute time they need to do their analysis. The cost of hosting 2PB of data on any of the commercial cloud providers is roughly $500K USD per year (https://cloud.google.com/products/calculator/id=8ee0d849-a19b-44ab-b546-1b0c0dbe775d). This is no small sum, but it is likely much less than the collective operating budget of the ESGF nodes. The overall financial cost to funding agencies might even turn out to be less if individual research groups were persuaded to abandon their dark replicas and associated local data storage and computation costs in favor of cloud computing. Furthermore, commercial cloud providers might also

provide hosting for free, as they already do for many other scientific datasets (e.g. https://aws.amazon.com/public-datasets/, https://cloud.google.com/public-datasets/), if they think it will bring them customers from academia and industry.

Beyond academic research, CMIP data hold strong commercial value in sectors such as insurance and energy. If CMIP datasets can be liberated from closed institutional infrastructure, such consumers can more easily combine them with co-located domain specific datasets to gain insights and derive economic benefits. NOAA (an agency already contributing model development and simulation resources to CMIP) has recently adopted such an approach to power their Big Data Project through Cooperative Research and Development Agreements and could provide an example for future development within the climate science community.

The scientific payoff from co-locating CMIP data with distributed computing resources would be immense, both accelerating reproducibility and driving innovation in data analysis methodologies–including new machine learning and artificial intelligence techniques. But leveraging full advantage of distributed systems for analyzing climate data requires more than raw hardware; it also requires software which allows climate scientists to parallelize their calculations in a simple, efficient and transparent way, permitting them to focus on science rather than the details of the underlying computing system. A central focus of the Pangeo project is to develop these tools on distributed computing systems and deploy them on high-impact geoscience problems. The building blocks for such software exist, for example, in the scientific Python community in the form of packages such as xarray (https://xarray.pydata.org), Iris (http://scitools.org.uk/iris/), Dask (https://dask.pydata.org), and Jupyter (https://jupyter.org/), but require engagement from the broader climate science community to reach their full potential for our field. We stand ready to work with WCRP and ESGF to help our community transition to a cloud-based future.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52, 2018.