Contribution ID: **0**

Type: **Oral Presentation**

Ensuring Data Readiness within Climateprediction.net

Thursday, 4 April 2019 16:00 (20 minutes)

Climate change is both one of the grand scientific challenges of the moment and also one of the most politically charged. As such it is essential that research in this area operates in as transparent a manner as possible. This takes on extra relevance when considering studies that either lead towards political or social impact such as those feeding into the IPCC special report on 1.5degrees or impact studies of extreme event attribution, where some parties are even now trying to approach 'blaming' other parties for climate change.

Within the Climateprediction.net program we have long considered it essential, due to the nature of how we obtain our simulation results, to make our data open access. Until now this has mostly been through personal contact with the co-ordinating team located at the University of Oxford by any researcher that wishes to use the data. Whilst this is still an effective method, it is clear that we must disseminate more than just the final raw data, but also for that data to be truly effective we must include the experiment design so that users are able to fully understood how the data was created. As such the full data chain must be understood and accessible, from the persistent capture of experimental design, simulation management, steps in data readiness for results and finally archiving and publishing.

We have developed a number of these processes independently but are now in a position where we can describe the interconnection of these such that we can show end to end provenance on all studies now undertaken using CPDN;

1. Management of experiment definition Experiments are defined within CPDN using a number of configuration files, both model specific and XML based which manage all aspects of the experiment. They are managed via a repository system as well as their ingestion into the main database of the CPDN experimental setup as well as using the Trello project management tool. This allows remote submitters of work to interact with the core team within Oxford as well as keep all communication about the batch between the researchers stored for future reference. With these combination of steps and systems we have established the first steps in the provenance chain.
2. Simulation Management Once the workunits are submitted they are subsequently picked up by BOINC clients registered with the project and executed. As these are long running applications then to ensure that computational resource is not wasted the capability of check pointing and uploading intermediate results from the workunit back to the project servers is used extensively. The presence of these results and all the information about the systems which processed each workunit are kept within the main CPDN project database. This allows us to investigate any possibility of machine dependant results etc as well as understanding if there are any platform specific performance issues. The data that is returned is sent using the community standard, netcdf. This has the advantage of being a self-describing data format, where the content, diagnostics etc. that are contained within are all described within the headers of the files themselves.
3. Readiness of Results Data Once we have reached the end of the workunit then the results are all uploaded to the specified upload server which will curate and manage the results data for that workunit. Alongside this the client communicates with the core system scheduler to notify it that a task has completed and that results have (normally) completed successfully. The results data is presented to the scientist as a large number of individual files, one per workunit. As such utilising these large number of files could be severely challenging it has been necessary to develop a set of CPDN specific tools to allow simple interaction by the data using scientist with their data. These tools are managed within the Github software management platform. The processed data is again created using the netcdf data standard and at this point that the data is ready for analysis and subsequent publication of the science results normally occurs.
4. Archive and publication With many publishers as well as funding agencies it is necessary to curate your open data for a period beyond its last use of publication date. As such we have also determined that to increase the possible utilisation of the dataset that it would be welcome to make the data available using community repositories and descriptors which the research who may reuse the data greater insight into how it has been created. As such working with repositories such as the UK NERC Centre for Environmental Analysis repository and publishing venues such as Elsevier Data in Brief or Nature Scientific Data. This allows greater depth in the description of the available datasets as well as linkage to scientific results that have been already generated using them.

Summary

Climateprediction.net is the largest climate simulation facility available to the general research community currently. As such it has partners in eleven different countries and has performed over 200 million years of simulation of both the whole earth system using HadCM3 global coupled model and using Atmosphere only regional models within the Wether@Home sub experiments. Generating in excess of 0.5PB of open access climate data we have needed to create a robust pipeline to ensure that provenance of the data is maintained. This has become especially necessary within the extreme event attribution experiments of Weather@Home projects where we can use up to 120k member ensembles to generate robust attribution statements. This pipeline must operate from experiment definition through to final archiving of the data when the main experiment has completed it utilisation and at all times we must be able to return to the previous step to understand where a particular piece of data has come from.

Primary author: Prof. WALLOM, David (University of Oxford)

Co-authors: Dr RASHID, Mamun (Kings College London); Mr UHE, Peter (Univeristy of Bristol); Dr SPARROW, Sarah (University of Oxford); Dr LI, Sihan (University of Oxford)

Presenter: Prof. WALLOM, David (University of Oxford)

Session Classification: Earth & Environmental Sciences & Biodiversity Application

Track Classification: Earth & Environmental Sciences & Biodiversity Applications