DATA MANAGEMENT AND ANALYSIS OF THE HIGH-RESOLUTION MULTI-MODEL CLIMATE DATASET FROM THE PRIMAVERA PROJECT

Jon Seddon, Met Office, Exeter, UK, jon.seddon@metoffice.gov.uk

Ag Stephens, Centre for Environmental Data Analysis, STFC Rutherford Appleton Laboratory, UK

This project has received funding from the European Union's Horizon 2020 Research & Innovation Programme under grant agreement no. 641727.

PRIMAVERA OVERVIEW

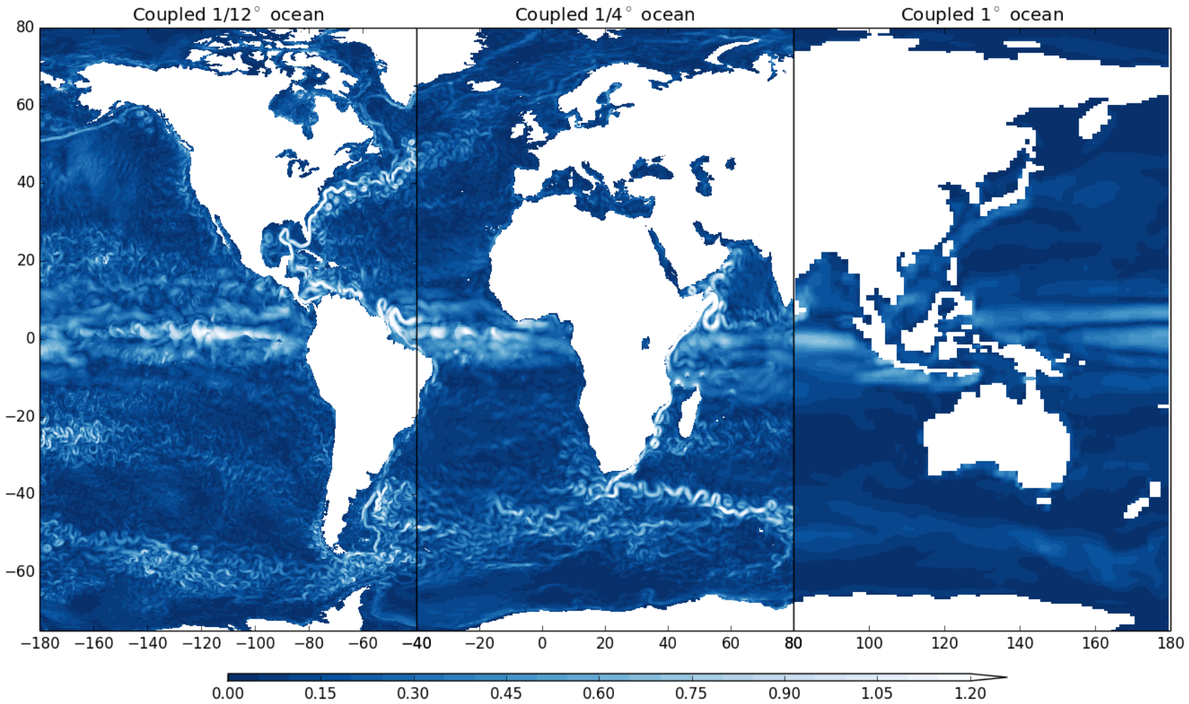
PRIMAVERA is a European Union Horizon 2020 funded project. There are 20 project partners based across Europe and over 100 researchers working on the project.

PRIMAVERA aims to generate:

- a new generation of advanced and well-evaluated high-resolution global climate models,
- simulations and predictions of regional climate with unprecedented fidelity,
- for the benefit of governments, business and society in general.

PRocess-based sIMulation:

AdVances in high resolution modelling and European climate Risk Assessment



Ocean surface currents from HadGEM3-based global coupled (atmosphere-ocean/sea-ice) models at three different resolutions - (left) 25km-1/12 degree, (middle) 60km-1/4 degree, (right) 130km-1 degree (courtesy of Malcolm Roberts)

https://www.primavera-h2020.eu/

PRIMAVERA'S DATA CHALLENGE

An important component of PRIMAVERA are the Stream 1 and Stream 2 simulations. Seven different global climate models were run at standard CMIP6 resolution and at a higher resolution (typically around 25 km in the atmosphere). These simulations follow the HighResMIP[1] protocol, which is part of Coupled Model Intercomparison Project Phase 6 (CMIP6). Each of the seven models has run the following experiments:

- highresSST-present – atmosphere-only 1950 to 2014
- highreSST-future – atmosphere-only 2015 to 2050
- spinup-1950 – 30 years of coupled spinup
- hist-1950 – coupled 1950 to 2014
- highres-future – coupled future 2015 to 2050
- control-1950 – coupled 1950 to 2050 with a fixed 1950s forcing

For Stream 1, each model ran one ensemble member at standard and high resolution. For Stream 2, some models ran additional ensemble members and other models performed new runs to demonstrate some of the new physics code that has been developed.

At the start of the project, it was estimated that these simulations would generate over 2 petabytes of data. PRIMAVERA has 400 terabytes (TB) of disk storage allocated to it at JASMIN. Therefore most data would have to be held on tape and only the data currently being analysed could be held on disk. The PRIMAVERA Data Management Plan was developed to allow the model data to be transferred to JASMIN, stored there and catalogued so that each file can be tracked and searched as it was moved between tape and disk.

The PRIMAVERA Data Management Tool (DMT) is the software developed to implement the Data Management Plan at JASMIN. Users use the DMT's web interface to query the available data and request that data is restored from tape to disk.

[1] Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., Chang, P., Corti, S., Fučkar, N. S., Guemas, V., von Hardenberg, J., Hazeleger, W., Kodama, C., Koenigk, T., Leung, L. R., Lu, J., Luo, J.-J., Mao, J., Mizielinski, M. S., … von Storch, J.-S. (2016). High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geoscientific Model Development*, *9*(11), 4185–4208. https://doi.org/10.5194/gmd-9-4185-2016

JASMIN SUPER DATA CLUSTER

http://www.jasmin.ac.uk/

JASMIN is funded by the Natural Environment Research Council and the UK Space Agency and operated by the Centre for Environmental Data Analysis (CEDA), part of the Science and Technology Facilities Council (STFC). JASMIN is located at STFC's Rutherford Appleton Laboratory (RAL), Oxfordshire, UK. JASMIN consists of:

- over 39 petabytes of disk storage
- interactive data analysis servers
- 4000 cores of compute in the LOTUS batch cluster system
- a low-latency high-bandwidth network connecting the storage and compute
- a fast connection to JANET, the UK academic network, which has a fast connection to the European GÉANT research network
- access to the RAL tape library for offline storage of data

STFC are PRIMAVERA project partners. The disk storage is split into units called group workspaces, each group workspace is around 100 TB in size.

DATA MANAGEMENT PLAN

1. Generate – seven different climate models are run on HPCs across Europe.

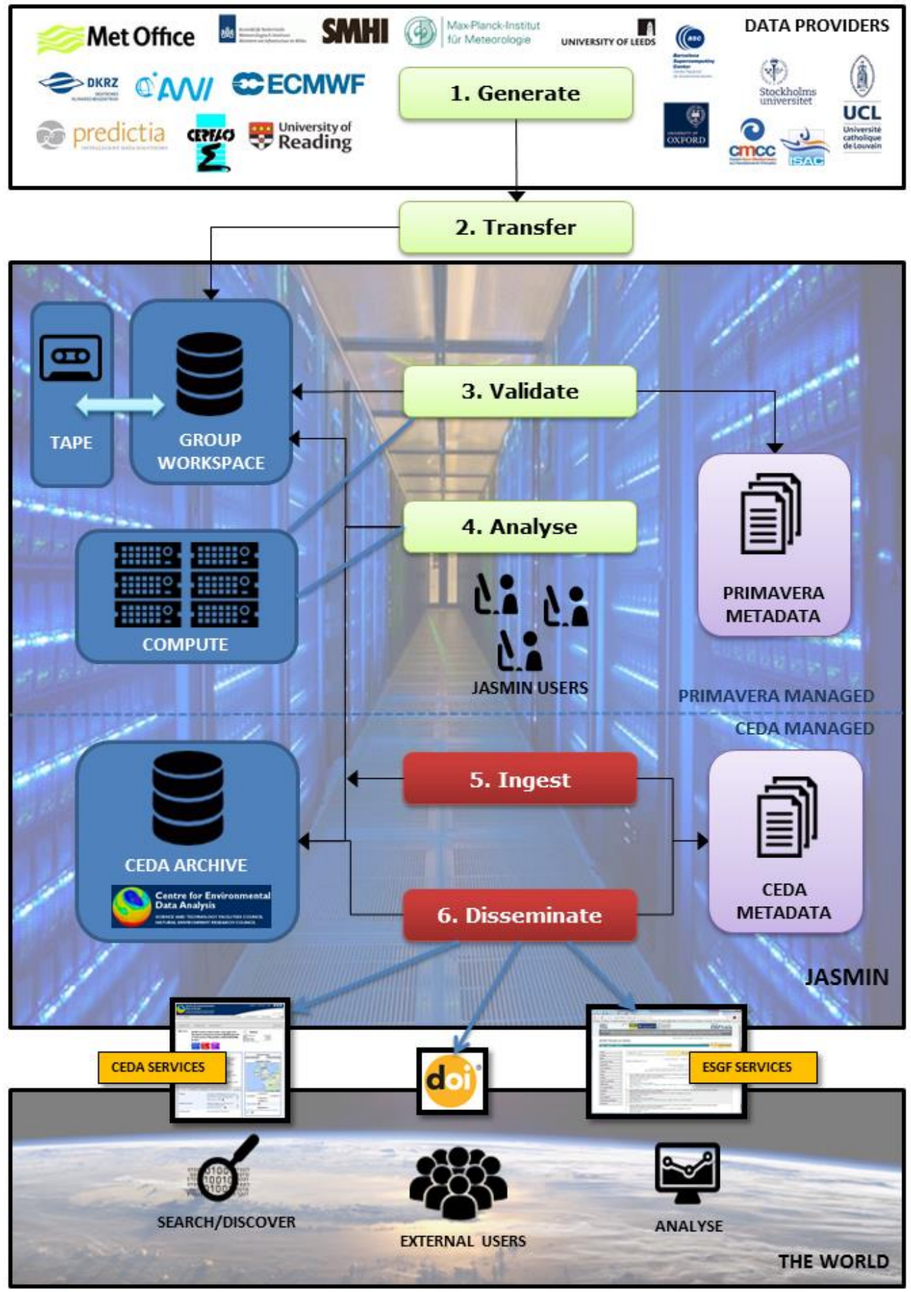
2. Transfer – data is transferred to the PRIMAVERA group workspaces at JASMIN.

3. Validate – uploaded files are checked for their compliance with the CMIP6 and PRIMAVERA standards. Their metadata is added to the PRIMAVERA database and then they're moved to tape to create space for more files to be uploaded.

4. Analyse – users check the DMT's web interface for available data. If it's only available on tape then they use the DMT to request that it's restored to disk. They can develop their analysis on JASMIN's interactive data servers and then run it on the complete dataset using the LOTUS batch cluster. When they're happy with their results then they can mark the data as complete. If the data isn't required by any other users then it's deleted from disk to free-up space for other users' data.

5. Ingest – data is ingested into the CEDA archive and published on CEDA's Earth System Grid Federation (ESGF) node.

6. Disseminate – data is made available to the global community through CEDA's ESGF node. Through the ESGF, data is discoverable and citable by the allocated DOIs.

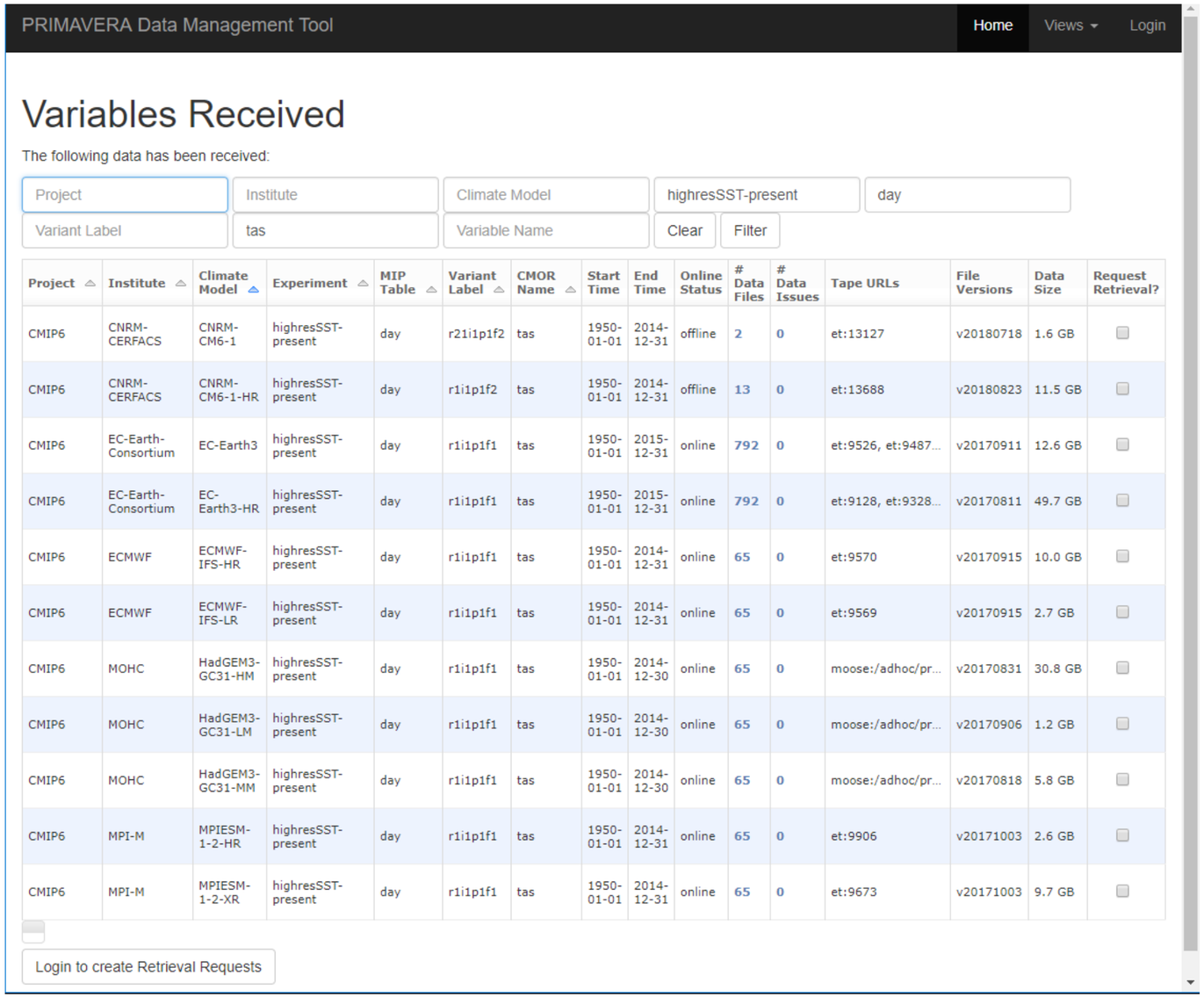


DATA MANAGEMENT TOOL

The DMT was written in the Python programming language and uses the Django web framework. It is installed on a dedicated server at JASMIN that can be accessed from the Internet. The server also hosts the PRIMAVERA PostgreSQL database.

Users can use the DMT's web interface to query the available data. The screenshot here shows a query for the tas variable (surface air temperature) from the day table of the highresSST-present (atmosphere only) experiment. Data with an "Online Status" of "online" means that all files for that variable are available on disk, "partial" means that some files are on disk and some are only available on tape, and "offline" means that the files are only available on tape.

Users can click in the "Request Retrieval" column to indicate that they are currently working on a variable. If that variable isn't currently on disk then it will be queued to be restored from tape to disk. The user will be emailed when the data has been restored. When a user has finished with a variable then they mark it as complete on another page in the DMT. If the variable isn't being used by any other users then it will be deleted from disk to create space for other data.



CONCLUSIONS

The combination of JASMIN and the Data Management Tool allowed the 100 PRIMAVERA scientists spread across Europe to analyse the almost 2 PB of data produced during the project.

Rather than having to copy data from the HPCs to the many institutes where it was being analysed, JASMIN allowed all users to work where the data was stored. Users had to learn how to use JASMIN, but it is similar to the existing Linux facilities at many institutes. The existing JASMIN documentation, and the new training materials developed by PRIMAVERA, helped scientists adapt their existing workflows and tools to run at JASMIN.

We would recommend that multi-model or multi-institute projects similar to PRIMAVERA consider the use of JASMIN or similar facilities. JASMIN only has a limited capacity and so cannot support all of the projects that would benefit from having access to it. The funding and development of further facilities similar to JASMIN should be considered.

PRIMAVERA data management legacy:

- The Stream 1 and Stream 2 HighResMIP simulations are now being made available to the global community through CEDA's ESGF node - https://esgf-index1.ceda.ac.uk/.
- The PRIMAVERA Data Management Tool is available under an open source license for other projects to use at https://github.com/PRIMAVERA-H2020/primavera-dmt. The current DMT implementation makes assumptions about the PRIMAVERA data and JASMIN and so will require refactoring before it can be used by other projects.
- PRIMAVERA Deliverable 9.6 will review the Data Management Plan and will document the lessons learnt that can be applied to other projects (due July 2020).