Call: H2020-SC5-2014-two-stage

Topic: SC5-01-2014

PRIMAVERA

Grant Agreement 641727

PRocess-based climate sIMulation: AdVances in high resolution modelling and European climate Risk Assessment

Deliverable D9.6

Review of DMP and lessons learnt for future projects

Deliverable Title	Review of DMP and lessons learnt for future projects	
Brief Description	Review of outcomes of the DMP and document lessons for future EU and international big data projects.	
WP number	9	
Lead Beneficiary	Met Office	
Contributors	Jan Hegewald, Tido Semmler, AWI Pierre-Antoine Bretonnière , Louis-Philippe Caron, BSC Alessandro D'Anca, Sandro Fiore, CMCC Marie-Pierre Moine, CERFACS Chris Roberts, Retish Senan, ECMWF Matthew Mizielinski, Malcolm Roberts, Jon Seddon, Met Office Uwe Fladrich, SMHI Ag Stephens, STFC	
Creation Date Version Number Version Date	29th April 2020 1.2 24th July 2019	
Deliverable Due Date Actual Delivery Date	31st July 2020 27th July 2020	
Nature of the Deliverable	R	*R - Report*
		P - Prototype
		D - Demonstrator
		O - Other
Dissemination Level/ Audience	PU	*PU - Public*
		PP - Restricted to other programme participants, including the Commission services
		RE - Restricted to a group specified by the consortium, including the Commission services
		CO - Confidential, only for members of the consortium, including the Commission services

Version	Date	Modified by	Comments
1.0	10th July 2020	Jon Seddon	First version
1.1	21st July 2020	Jon Seddon	Combined reviews from all authors
1.2	24th July 2020	Jon Seddon	Added Project Officer suggestions

Table of Contents

List of Tables

List of Figures

1 Executive Summary

WP9 has successfully delivered 1.6 PiB of data to the JASMIN central analysis facility, which has allowed over 100 scientists to analyse the data and complete the science in the rest of the project. Many lessons have been learnt that are applicable to future EU and international big data projects. This report summarises WP9's activities and highlights the important lessons learnt.

Data Management Plans (DMPs) were developed in D9.1 and MS25 for the Stream 1 and Stream 2 simulations respectively. These DMPs were followed throughput the project and remain valid at the end of the project with no changes to them required. The project has generated a large volume of well described data for the project's users and for the global climate science community. Deliverables D9.4 and D9.5 are evidence of DMPs' success and describe the data that has been generated.

The data uploaded to JASMIN has been archived and published to the Earth System Grid Federation to make it available to the global community. The persistent identifiers allocated to this data have been included in D9.5 and on the project's external website (https://www.primavera-h2020.eu/). Documentation describing the data has been published in the community-standard ES-DOC format, which is linked from the data catalogues that the persistent identifiers resolve to, and in peer reviewed journal articles. D9.7 describes how the external community can gain access to the data and interact with project members to obtain support and collaborate, which the PRIMAVERA external website supports and will remain available for the foreseeable future.

Recommendations have been made throughout the report and are repeated here:

- Projects should use central facilities such as JASMIN to store and analyse their data. The use of such a facility should be coordinated by a tool such as the DMT.
- Wherever possible, the running of simulations, post-processing and all other processes should be automated, and tools should be developed to configure the automation directly from the machine-readable data request.
- Early effort defining data requirements, via the data request, can save a lot of work later because of the stable data request.
- Estimating the volume of data to be produced by a project is a difficult task. Generating a worst case estimate of the data volume will allow the storage and transfer resources required for a project to be sized.
- The data request should be carefully analysed during its development to prevent variables being added that are not later analysed. Unused variables require work from the modelling centres to produce, slow the models down and require additional storage. The variable access statistics from this report may be useful when designing future data requests.
- Careful consideration is needed for productive use of tape archives. Loading tapes into drives can be slow and should be minimised for data that will be read from tape multiple

times. Exploration of the organisation of data on tape is needed as it may be quicker to store an entire variable in each tape write to minimise the number of tape loads when restoring data. This comes at a cost of using disk space to cache data after the simulation and before writing to tape.

- CEDA, the operators of JASMIN, should consider the introduction of shared accounts for projects like PRIMAVERA so that processing and operations are not dependent on a single user.

- Future projects using a system such as the DMT should generate regular summaries of what data has recently been made available. Ideally, these updates should be generated automatically.

- Modelling groups should carefully evaluate the precision that values are saved with, because storing data with greater precision than is necessary wastes storage. There is some evidence that quantising variables to less than four bytes allows netCDF deflation to achieve better compression ratios and could potentially reduce the total volume of data by up to 50%. This requires further investigation using a wider range of models and variables.

- If future projects require a sustained transfer rate greater than 2.5 TiB/day then this should be achievable if there is some prior human (and potentially financial) resource dedicated to improving the current transfer systems.

- Use a versioning scheme for technical documents such as the data request so that everyone knows that it is under development and when it becomes stable.

2 Project Objectives

With this deliverable, the project has contributed to the achievement of the objectives shown in *Table 1* (DOA, Part B Section 1.1), WP numbers are in brackets:

No.	Objective	Yes	No
A	To develop a new generation of global high-resolution climate models. *(3, 4, 6)*	Yes	
B	To develop new strategies and tools for evaluating global high-resolution climate models at a process level, and for quantifying the uncertainties in the predictions of regional climate. *(1, 2, 5, 9, 10)*	Yes	
C	To provide new high-resolution protocols and flagship simulations for the World Climate Research Programme (WCRP)'s Coupled Model Intercomparison Project (CMIP6) project, to inform the Intergovernmental Panel on Climate Change (IPCC) assessments and in support of emerging Climate Services. *(4, 6, 9)*	Yes	
D	To explore the scientific and technological frontiers of capability in global climate modelling to provide guidance for the development of future generations of prediction systems, global climate and Earth System models (informing post-CMIP6 and beyond). *(3, 4)*		No
E	To advance understanding of past and future, natural and anthropogenic, drivers of variability and changes in European climate, including high impact events, by exploiting new capabilities in high-resolution global climate modelling. *(1, 2, 5)*		No
F	To produce new, more robust and trustworthy projections of European climate for the next few decades based on improved global models and advances in process understanding. *(2, 3, 5, 6, 10)*		No
G	To engage with targeted end-user groups in key European economic sectors to strengthen their competitiveness, growth, resilience and ability by exploiting new scientific progress. *(10, 11)*		No
H	To establish cooperation between science and policy actions at European and international level, to support the development of effective climate change policies, optimize public decision making and increase capability to manage climate risks. *(5, 8, 10)*		No

Table 1 – The objectives that this deliverable has contributed to.

3 Detailed Report

3.1 Developing the Data Request

The original PRIMAVERA plan (Appendix 1 of the Grant Agreement (PRIMAVERA, 2015)) envisaged that the Stream 1 simulations would be run at the start of the project and would form the basis of the analysis and development work for the duration of the project. The Stream 2 simulations would be run towards the end of the project and would incorporate the model improvements developed during the project. The Stream 2 simulations would include additional variables that were deemed necessary by user facing work packages 10 and 11. The Stream 1 simulations would be archived but were envisaged to be primarily for use by the project. The Stream 2 simulations would be submitted to CMIP6.

The HighResMIP protocol (Haarsma et al., 2016) specifies the use of the MACv2-SP EasyAerosol scheme (Stevens et al., 2017). The EasyAerosol forcings were delayed and first became available in December 2016 and so the Stream 1 simulations could not begin until then. Because of the delay there was limited time for model development work to take place between the Stream 1 and 2 simulations, and hence not enough time for the Stream 2 simulations to be as comprehensive as the Stream 1 simulations. With permission from the European Commission a revised set of deliverables was developed. The Stream 1 simulations would be PRIMAVERA's contribution to HighResMIP and CMIP6 and the Stream 2 simulations would primarily enhance the Stream 1 simulations, but also provide some new additional experiments associated with specific scientific questions.

One of the HighResMIP authors met the CMIP6 Data Request's (Juckes et al., 2020) author around 2014 or 2015 to start the development of the HighResMIP data request, which was initially based on the default CMIP5 data request. Any variables specific to Earth system model components were removed as the HighResMIP models were not intended to contain Earth system components. The global high-resolution modelling community were asked what variables they considered were necessary for analysis and these were added to the HighResMIP request. It was known that some variables such as daily variables on model levels would be too expensive and so these were removed from the request. The HighResMIP authors had to address concerns about the volume of the data in the higher temporal frequency variables but were able to show that these were necessary for the analysis of the processes occurring in the higher spatial resolution models that would be contributing to HighResMIP. Some of the HighResMIP authors were simultaneously developing the PRIMAVERA proposal and were aware that additional variables would be required for some of the planned PRIMAVERA analyses and they knew that they could not justify including these in the global HighResMIP data request.

Different user communities have different requirements from a climate model's output data. Users studying climate impacts typically require data at a high temporal frequency whereas the modelling groups are less keen on providing data at these frequencies because of the effect on model run-time and storage requirements. While the data request was being developed the user facing work packages were beginning their work and had less experience to feed into the development process. During the planning for the Stream 2 simulations in November 2018 there was a productive discussion involving all of the

relevant work packages, who by then had more experience, and an analysis of the data retrieved so far which resulted in the significantly smaller data request documented in milestone 25.

The first data request sent to the members of PRIMAVERA WP9 was based upon 00.beta27 around May 2016. As the CMIP6 data request evolved this had to be updated, which was a time-consuming process. Typical changes included variables having their name changed, variables being moved from one table to another or the names of tables changing. An automated tool was developed to help update the PRIMAVERA data request (Seddon, 2020a), but it still took a considerable amount of effort to keep the PRIMAVERA data request in sync with the changes being made to the CMIP6 data request. The PRIMAVERA data request was used by each of the modelling groups to control the output from their own models; at each new version of the data request the modelling groups had to reconfigure the output from their models.

Various levels of automation were employed by each of the modelling centres to configure their simulations and post-processing. The CERFACS model outputs CMOR compliant data directly from the XIOS I/O server, which was configured directly from the data request. The Met Office developed two tools, one to configure their model's I/O server from the data request and a second tool to configure the CMORization post-processing step from the data request. Considerable software development time is required to develop such tools, but such tools insulate groups from changes to the data request and once thoroughly tested, prevent any mistakes from occurring when configuring the models.

The CMIP6 project provides the CMOR library (Nadeau et al., 2019) to generate standards compliant netCDF files. The inputs required to configure CMOR are known as the "MIP tables" and these were forked and code was written to add the PRIMAVERA specific variables allowing all groups to use CMOR to generate compliant data if required (Nadeau et al., 2018).

The modelling groups typically began running their simulations when the CMIP6 data request was at version 01.00.07. When the Earth System Grid Federation (ESGF) became operational the minimum data request version that it would accept was version 01.00.21. A tool was therefore developed to perform fixes to the metadata in files as they were submitted to the ESGF (Seddon, 2020b).

Because PRIMAVERA was at the forefront of model development it encountered frequent problems with it wanting to start work before the rest of the community was ready and before standards and tools had become mature. Particularly in work package 9, this generated additional work, such as the problems with the data request, forcings and MIP tables not being ready. This problem is exacerbated because high-resolution simulations need to start earlier due to their significantly larger computational requirements and slower progress.

Recommendations:

- Wherever possible, the running of simulations, post-processing and all other processes should be automated and tools should be developed to configure the automation directly from the machine-readable data request.

3.2 Data Request and Transfer Modelling

At the start of the project WP9 performed a significant amount of modelling work to check that the data volumes expected could be run on the available HPC resource and then transferred to, and stored at, JASMIN successfully (Mizielinski et al., 2016). This work was useful to confirm that the planned data volumes were feasible. The WP9 leader left the project in August 2017 and there was then 0.5 full-time equivalent less resource available for WP9 work. Less time was spent on management activities and modelling the dataflow to compensate for this loss of resource. Resource was instead focussed on developing the DMT, monitoring the flow of data through the system and supporting the data providers to upload and validate their data. Fortunately, few problems were encountered and the earlier modelling work allowed the data to be safely uploaded and stored.

The modelling of the likely volume of data to be uploaded was incredibly complex. It was initially estimated that the volume of the Stream 1 simulations would be 2.4 PiB[1] but 692 TiB was uploaded in the end. Table 2 shows the actual volume of data uploaded by each model for the highresSST-present atmosphere-only experiments and the volumes that were expected to be uploaded according to milestone 22 (MS22). The actual volume uploaded was 42% of the planned volume. The number of individual variables uploaded was 52% of the initially planned variables (Seddon, 2020d). There are many reasons for these differences, including that the data volumes used in the planning were only estimates as new versions of the data request were regularly being released. It was also not known during the planning which of the priorities of variables would be produced. With the version of the data request available during the development of MS22 the HadGEM3-GC31-HM highresSST-present experiment was estimated to produce 424 GiB per year for the priority 1 variables and then an extra 254 GiB and 15,981 GiB for priority 2 and 3 respectively. In the end, most groups produced only the priority one variables.

Model	Volume Uploaded, TiB	Planned, TiB
AWI-CM-1-1-LR	0.3	6.4
AWI-CM-1-1-HR	0.3	25.6
CMCC-CM2-HR4	0.8	6.4
CMCC-CM2-VHR4	11.6	69.6
CNRM-CM6-1	4.3	3.2
CNRM-CM6-1-HR	13.1	44.8
EC-Earth3P	8.4	3.2
EC-Earth3P-HR	32.9	44.8
ECMWF-IFS-LR	0.9	1.3
ECMWF-IFS-HR	3.3	10.2
MPI-ESM1-2-HR	1.5	6.6

[1] Data volumes and rates in this paper use binary prefixes and so 1 PiB is 1024^5 bytes.

MPI-ESM1-2-XR	2.8	13.1
HadGEM3-GC31-LM	1.8	4.5
HadGEM3-GC31-HM	46.4	64.0
Total	**128.3**	**303.7**

Table 2 – The actual and planned volumes of highresSST-present data.

These assumptions in the data volume calculations estimated that it would require 2.5 bytes per grid box to store a file. Table 3 shows the actual number of bytes to store each grid box for typical files from the high-resolution model from each of the modelling centres. All modelling centres were encouraged to save their files using netCDF version 4 classic with chunking enabled and a deflation level of at least 3. The number of bytes per point will vary depending on the number of time points in each file. Storing one time point per file is inefficient because a copy of the grid needs to be included for each time point, whereas with one month or one year of data in a file then there will only need to be a copy of the grid for many time points. For regular latitude and longitude grids then the grid overhead is much smaller than for irregular grids. However, with higher spatial-resolution data then there is a limit on the number of time points that can be placed in one file while keeping the file to a manageable size.

The results from Table 3 show that there is considerable variation in the number of bytes used to store each grid box. Precipitation (variable name pr) has more variation between adjacent grid cells than temperature (variable ta) and so does not compress as well. The assumed value of 2.5 bytes is probably realistic when averaged over all variables.

Filename	**Density, bytes / point**
ta_Eday_AWI-CM-1-1-HR_hist-1950_r1i1p1f2_gn_19510101-19511231.nc	3.0
ta_6hrPlevPt_CMCC-CM2-VHR4_hist-1950_r1i1p1f1_gn_195001010000-195001311800.nc	2.1
ta_6hrPlevPt_CNRM-CM6-1-HR_hist-1950_r1i1p1f2_gr_195001010600-195004010000.nc	1.9
ta_6hrPlevPt_EC-Earth3P-HR_hist-1950_r1i1p2f1_gr_195001010000-195012311800.nc	2.2
ta_6hrPlevPt_ECMWF-IFS-HR_hist-1950_r1i1p1f1_gr_195001010000-195001311800.nc	1.5
ta_6hrPlevPt_MPI-ESM1-2-XR_hist-1950_r1i1p1f1_gn_195001010558-195012312358.nc	1.5
ta_E3hrPt_HadGEM3-GC31-HM_hist-1950_r1i1p1f1_gn_195001010300-195001302100.nc	1.7
pr_3hr_AWI-CM-1-1-HR_hist-1950_r1i1p1f2_gn_195101010130-195112312230.nc	3.8
pr_Prim6hr_CMCC-CM2-VHR4_hist-1950_r1i1p1f1_gn_195001010000-195001311800.nc	3.2
pr_3hr_CNRM-CM6-1-HR_hist-1950_r1i1p1f2_gr_195001010130-195012312230.nc	3.1
pr_3hr_EC-Earth3P-HR_hist-1950_r1i1p2f1_gr_195001010000-195012312100.nc	2.2
pr_Prim6hr_ECMWF-IFS-HR_hist-1950_r1i1p1f1_gr_195001010300-195012312100.nc	1.6
pr_Prim6hr_MPI-ESM1-2-XR_hist-1950_r1i1p1f1_gn_195001010558-195012312358.nc	2.1
pr_3hr_HadGEM3-GC31-HM_hist-1950_r1i1p1f1_gn_195001010130-195006302230.nc	3.1

Table 3 – The number of bytes to store each point of data in typical files for each model.

Table 3 shows that ECMWF atmosphere data consistently takes fewer bytes to store than the data from other models. The reason for this has not been proven but it is believed that the raw GRIB data was restored from the ECMWF tape archive using the default bitsPerValue setting of 24 bits (3 bytes per datapoint). The data was then converted to netCDF files and stored as type NC_FLOAT requiring 4 bytes. The quantisation applied during the extraction from tape

has allowed the netCDF deflation to work more efficiently, reducing the number of bytes required per point. Similarly, in the Met Office HadGEM3-GC31 files temperature data was packed with WGDOS packing code -12, quantising the temperature data, but the precipitation data was not packed. This resulted in the larger difference in density between these two variables for HadGEM3-GC31 than for other models. This is an important lesson for all models in that the precision required to store each variable should be carefully assessed to ensure that the data is not being stored with unnecessary precision. These examples suggest that reducing the precision could reduce the volume required to store data by up to 50% compared to storing with full precision.

Further reductions in data volume can be achieved by appropriate choice of netCDF deflation levels and chunk sizes. Chunk sizes should be carefully chosen based on a knowledge of expected data access patterns as the wrong size can make reading data from the file very slow.

Estimating data volumes is an art rather than a science, and a difficult one at that. In the case of PRIMAVERA it was made more difficult because the data request was changing frequently. The volumes were overestimated but allowed the resources available to the project at JASMIN to be sized.

It is recommended that:

- Estimating the volume of data to be produced by a project is a difficult task. Generating a worst case estimate of the data volume will allow the storage and transfer resources required for a project to be sized.
- Modelling groups should carefully evaluate the precision that values are saved with, because storing data with greater precision than is necessary wastes storage. There is some evidence that quantising variables to less than four bytes allows netCDF deflation to achieve better compression ratios and could potentially reduce the total volume of data by up to 50%. This requires further investigation using a wider range of models and variables.

3.3 Data Request Summary

As of 26th June 2020, 30,955 data requests had been uploaded to JASMIN comprising 1.5 pebibytes (PiB) of data. 5,941 unique data requests had been restored at least once, which is 19% of the data requests by number. These restored unique data requests contained 418 tebibytes (TiB) of data, which is 27% of the volume uploaded. The fact that more data by volume rather than by number of data requests was restored suggests that PRIMAVERA users were interested in analysing the higher volume data requests, which tend to be the higher temporal-frequency variables. However, less than a third of the data by either measure has been analysed yet. Anecdotally, some retrievals have recently been created for variables that had not been requested previously. This could suggest (but has not been measured) that users initially work with some common variables, but in the longer-term, science will be conducted with a larger proportion of the data request. However, the low volume of data that has been accessed so far suggests that some items could be removed from the data request.

A list of the variables that has been retrieved from tape to disk using the PRIMAVERA DMT is shown in Appendix B in table name order and in Appendix C in order of popularity. The code and original data to generate these has been published (Seddon, 2020d). The most frequently accessed variables are all monthly variables, including precipitation flux, surface temperature, sea level pressure, air temperature, geopotential height, eastward and northward winds, seawater potential temperature and sea water velocity (northward and then eastward). Daily variables and then sub-daily have been requested slightly less frequently but the same variables have been requested at each frequency.

PRIMAVERA was provided with a copy of the logs from the THREDDS Tomcat server from CEDA's ESGF node (CEDA, 2020), where all of the PRIMAVERA data has been published. ESGF data can be downloaded over the HTTP protocol from THREDDS or alternatively via Globus (or accessed directly at JASMIN). Therefore, these logs do not show all of the data that has been downloaded, although the Globus downloads include replication of the data to other ESGF nodes and so may not be representative of actual usage of the data. As the data has been replicated to other ESGF nodes, CEDA's logs only show downloads and do not show the total global downloads. The logs had been anonymised by replacing the IP address in them with a hash of the IP address. Most datasets have a single variable spread over multiple files and so a download has been counted as each download of that variable from each unique IP address hash, which will miss multiple downloads from any sites that use a web proxy with a single outgoing IP address and will also miss access from users at sites with a local replica of the data. Using the code in (Seddon, 2020d) the logs have been parsed and the results sorted by frequency and table name are shown in Appendix D and by popularity order in Appendix E.

Similar monthly variables have been downloaded from the ESGF as have been retrieved from JASMIN by PRIMAVERA users. From the ESGF node, after the monthly data, similar daily variables were downloaded, whereas in PRIMAVERA little daily data was retrieved and instead users appeared to use the sub-daily data. In the sub-daily data, similar variables have been accessed from ESGF and at JASMIN.

A thorough analysis of the variables retrieved from PRIMAVERA and downloaded from CEDA's ESGF node has not been completed here. However the complete list of the variables accessed from the two methods has been published (Seddon, 2020d) and may be useful when developing the data request for future high-resolution modelling projects when the constraints on the data request and available resources are better known.

Recommendations:

- The data request should be carefully analysed during its development to prevent variables being added that are not later analysed. Unused variables require work from the modelling centres to produce, slow the model down and require additional storage. The variable access statistics from this report may be useful when designing future data requests.

3.4 Data Management Tool

The Data Management Tool (DMT) was developed to the specification developed in the original Data Management Plan (deliverable D9.1). It allows data to be tracked and its flow to be controlled around JASMIN. It consists of a PostgreSQL database and custom software written in the Python programming language using the Django web framework (Django, 2019). A dedicated server (prima-dm.ceda.ac.uk) in the JASMIN managed cloud was allocated and HTTPS access to it from the Internet was allowed. The database, a web server and the DMT application were installed on the server. Access to the database was made available from all other hosts at JASMIN, allowing data intensive operations such as validating data to be run on the 4000 node LOTUS batch cluster.

The DMT software (Seddon and Stephens, 2020) is freely available under a BSD 3 Clause open source license. Development of the DMT began in April 2016 and the first data was made available via the DMT in May 2017, requiring the full-time work of one developer. Development work has continued as required since then to improve the flow of data through the system and to facilitate the publication of data to the ESGF.

The DMT has largely worked well and positive feedback was received from users (see Section 3.9). The ability for users to retrieve data when needed and to indicate when they had finished with the data meant that data was only on disk when being actively accessed. The DMT therefore enables the available JASMIN storage to be used very efficiently.

The DMT and all data analysis used the workflow shown in Figure 1. Users check the DMT's website to see if the data that they require is currently available. If it is only currently available on tape, then they can use the DMT to request that a subset of it is restored from tape to disk. They are sent an email by the DMT when this subset is available on disk and can then work on JASMIN's interactive servers to develop and test their analysis code on this subset of data. When they are ready to run their analysis they use the DMT to request that the full dataset is restored from tape to disk and then run the analysis on the LOTUS batch cluster. Once users are happy with the results they can mark the dataset as complete in the DMT. If no other users have indicated that they are working with this dataset then the DMT deletes the data from disk to make space available for other users' data.

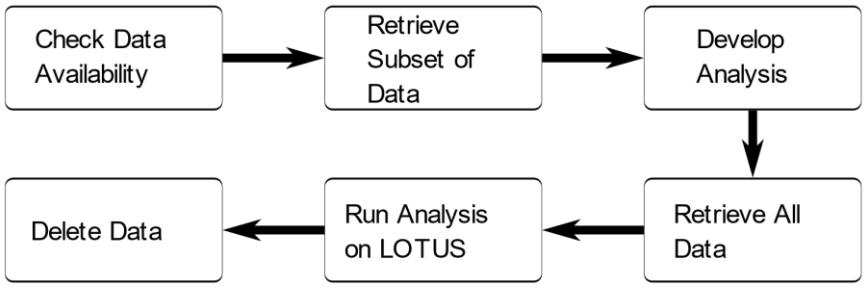


Figure 1 - The workflow required from users working with data in PRIMAVERA.

Scripts were written to automate all tasks in the DMT's workflow, such as validating data, moving validated data from tape to disk and restoring data from tape to disk. Initially the DMT's developer would monitor the data submissions and retrieval requests that users

created in the DMT and would submit the scripts to LOTUS or run them on the tape hosts as necessary. As the developer gained more confidence that the processes were working as intended then these scripts were triggered automatically by daemons (e.g. cron jobs or other scripts) on the tape hosts. After a considerable amount of effort, all processing was triggered automatically requiring little human intervention other than monitoring of the group workspace usage and ensuring efficient use of group workspaces.

Data providers were asked to upload data in chunks of no more than 5 TiB and were recommended that chunks of 0.5 TiB were ideal, to allow the validation software to process each chunk in a time that would fit into the standard LOTUS queue. The validation software was modified so that the number of parallel processes could be specified. If users uploaded data in larger chunk sizes, then the automatic validation of data was stopped and the validation was submitted to LOTUS using a larger number of cores, allowing it to run in the available time for the LOTUS queue. Ideally, if users uploaded data in chunk sizes of around the recommended 0.5 TiB then the automatic processes meant that only several TiB of group workspace was required to cache data being uploaded. However, if the validation software found errors in files' metadata then data would remain on group workspace for several weeks until the data providers fixed the metadata. This rarely happened, limiting the volume of group workspace required for data being uploaded.

There was a known issue with the DMT that made it occasionally unstable, which was commented upon by several users in the survey. The symptoms for the users were that the DMT's web page would not load and would take up to 15 minutes before it would load again. The cause of this issue is that the received data queries were generated by foreign key relationships linking the data requests and data files in the DMT's database. During testing and in the first year of the operation the number of records in the DMT was small enough that this worked well. Eventually the number of records became too large and that the web server timed out before these relationships had been resolved. As an immediate workaround the timeout on the webserver was increased, the server running the database and web server was given more processors and memory, and a quick query web page was created that forced users to include a search term in their query, which reduced the query to a size that ran quickly enough. Rather than calculating the status of each data request by checking its constituent data files whenever a web page is loaded, a full fix would cache each data request's status in the database allowing for a rapid generation of the web page without needing to follow any foreign key relationships. The cached status would be updated whenever a constituent file is modified, and possibly also periodically (e.g. overnight) to maintain consistency. This fix is a significant change to the DMT that would likely require the Python code for every operation that modifies a data file object to be updated and tested and is likely to take at least two weeks of full-time work and has not yet been implemented.

JASMIN has no facility for shared accounts and so the DMT and retrievals were run from the DMT developer's user id. This made it difficult for other users to monitor the DMT's processes when the developer was not available. There was also the risk if the developer was unable to work for a long period of time due to illness or injury then the flow of data

could have been stopped, giving the project a "bus factor"[2] of one. The use of shared or project accounts at JASMIN would give projects hosted there some additional resilience.

Recommendations:

- Include an API, perhaps a RESTful one that is already supported by Django, so that users can automate their workflows. One user managed to do this using an HTML GET request (e.g. https://prima-dm.ceda.ac.uk/received_data/?climate_model=HadGEM3-GC31-HH) and then screen scraping the response. A RESTful API would have made their work easier.
- Refactor the DataRequest objects in the DMT so that their status is cached in the database rather than requiring foreign-key relationships being followed to DataFile objects when generating web page views.
- CEDA, the operators of JASMIN, should consider the introduction of shared accounts for projects like PRIMAVERA so that processing and operations are not dependent on a single user.

3.5 Data Availability Summary for Users

During the peak of the uploads there was so much data arriving that it was difficult for users to know when new datasets had arrived. Periodically (monthly during the peak upload period), a table was manually compiled and emailed to the project showing the latest data availability (e.g. monthly and daily, or all except 1 hourly from experiment X with model Y) from each modelling centre. During Stream 2 a script was developed to generate a table of all available ensemble members. Users found these summaries a useful tool of when new datasets became available rather than having to regularly query the DMT. It is therefore recommended that:

- Future projects using a system such as the DMT should generate regular summaries of what data has recently been made available. Ideally, these updates should be generated automatically.

Developing software to generate such textual updates could be difficult. A simpler form of software generated update would show, for each model and experiment, for each table, the variables and time frames that had been recently uploaded. When data is being uploaded at a high rate in the middle of the project as happened in PRIMAVERA then this amount of information could become overwhelming for users and a human written summary may also be useful.

3.6 Storage on Group Workspaces and Tape

PRIMAVERA was allocated 440 TiB of storage at JASMIN split across five group workspaces (disks). It was estimated at the start of the project that the data volume was likely to be 2.5 PiB and so the data management plan described how the Data Management Tool (DMT) would allow the data to be split across tape and disk. The project was very privileged to have

[2] https://en.wikipedia.org/wiki/Bus_factor

access to this volume of disk storage as many other projects do not have access to such resources.

The split of the storage across five group workspace volumes added some complexity to the DMT, but also had the benefit that if one GWS became full unexpectedly (because more data was uploaded or retrieved than expected for example) then uploads or retrievals on the other four group workspaces were not affected. There is no mechanism at JASMIN for implementing quotas on individual users on a group workspace and so the WP9 leader had to closely monitor the group workspace usage and monitoring plots were produced daily to enable this.

One of the group workspaces contained an area known as the “cache” where users could store large intermediate datasets that they had been generated. This cache area was often around 20 TiB and would eventually grow large enough that its group workspace would be almost 100% full. When this happened, users who had the largest directories were emailed reminding them of the size of their cache directory. These users were typically able to free-up enough space to allow all users to continue working normally.

The five group workspaces were typically 90% full with the uploads and data being analysed moving through the system relatively quickly. There were only a few occasions during the project when data providers had to be asked to stop their uploads, or data users had to be emailed to remind them to mark data that they had finished with as complete. This did require close and active monitoring and active management of the disk usage by the WP9 leader to ensure that the group workspaces did not become 100% full.

Data is produced by climate models in chronological order and so it is convenient to upload it to JASMIN in time limited chunks, for example all variables one year at a time. Once data has been uploaded to JASMIN it is validated and then moved to tape. However, users typically analyse a small number of variables at a time but for all time periods. When restoring data from tape, the loading of a tape into a drive takes a significant part of the restore time. Using a variable from a 65 year highresSST-present experiment as an example, if a variable has been uploaded one year at a time then there may need to be 65 tapes loaded to restore one variable every time it is needed for analysis. Once this had been realised then data providers were encouraged to upload a variable (or variables for smaller lower-frequency variables) at a time rather than a year at a time. There may be a cost to doing this as all years may need to be extracted from tape and stored on disk during post-processing or prior to upload to JASMIN. This typically only happens once, whereas some variables are restored from tape to disk many times for analysis by different users. It is recommended that:

- Careful consideration is needed for productive use of tape archives. Loading tapes into drives can be slow and should be minimised for data that will be read from tape multiple times. Exploration of the organisation of data on tape is needed as it may be quicker to store an entire variable in each tape write to minimise the number of tape loads when restoring data. This comes at a cost of using disk space to cache data after the simulation and before writing to tape.

3.7 CMORization

There is a discussion in Appendix F of the approach taken by each of the modelling groups to producing CMOR compliant data. The approaches vary depending upon the output format of the model and the computing facilities available. The common factor affecting all of the centres was the ability to process the volume of data involved in a timely manner using the volumes of disk space available. Many groups had to run the post-processing on HPCs rather than Linux clusters for performance and storage reasons. One centre was able to output CMORized data directly from their model, which eliminates the need for a post-processing step.

3.8 Transfer rates achieved

Table 4 shows the data rates that each group reported that they achieved when transferring data to JASMIN. Groups were encouraged to use parallel transfer protocols such as GridFTP or BBCP as it was assumed that these would give the best transfer rates. As long as they were achieving adequate transfer rates then most groups chose to use the protocols that they were most familiar with. The best rate achieved was over 200 $MiBs^{-1}$, equivalent to 16.5 TiB per day, which was challenging to validate and move the data to tape. The typical rate achieved was 30 $MiBs^{-1}$ (2.5 TiB per day). The project assumed at the start that it would be possible to achieve 5TiB per day. As the transfer rates achieved were acceptable then no effort was put into improving these. It is therefore recommended that:

- If future projects require a transfer rate greater than 2.5 TiB/day then this should be achievable if there is some prior human (and potentially financial) resource dedicated to improving the current transfer systems.

Transfer from	Rate achieved	Protocol used
Toulouse, France	13 MiB s^{-1}	4 BBCP jobs in parallel, with 4 streams each
Hamburg, Germany	20 to 55 MiB s^{-1}	globus-url-copy
Lecce, Italy	34 MiB s^{-1} average, 69 MiB s^{-1} peak	gridftp, with 4 concurrent FTP connections 8 process in parallel
Bologna (Cineca), Italy	200 to 300 MiB s^{-1}	5 parallel `rsync -av -e "ssh -c arcfour"`
Barcelona, Spain	13 MiB s^{-1}	rsync
Exeter, UK	30 MiB s^{-1}	5 `moo get` in parallel
Reading, UK	85 MiB s^{-1}	4 parallel `rsync -rvz -rsh="ssh -c arcfour"`

Table 4 – Data rates achieved (Seddon et al., 2020)

3.9 User Survey

A survey was sent to members of PRIMAVERA to gather feedback on how successful the data management had been and what lessons could be learnt. The full set of responses can be seen in Appendix A – Data Management Plan User Survey. There were 24 unique responses to the survey.

		Agreeing the data request (and coping with its many versions)	Data management planning and the data request	Process of including requested diagnostics in your model	Data conversion process at your institute	Data upload to JASMIN
Score (1 is very difficult and 7 is very	**1**	1	1	0	1	1
	2	0	1	1	1	1
	3	2	0	2	1	0
	4	1	2	1	2	1
	5	0	0	2	0	0
	6	1	1	1	1	4
	7	0	0	0	0	0
# Responses		5	5	7	6	7
Mean		**3.4**	**3.4**	**4.0**	**3.3**	**4.4**

Table 5 - Number of responses for each score from the data providers.

The survey was split into two sections and the first was intended to be for members of WP9 who had supplied data to the project. Responses were received from seven members, who graded the tasks that they had performed, where 1 was very easy and 7 was very difficult and the results are shown in *Table 5*. The process of agreeing the data request and planning the data request at the start of the project were rated as neither difficult nor easy, but with a large spread of answers. Future projects could improve this score by:

- Use a versioning scheme for the data request so that everyone knows that it is under development and when it becomes stable.
- Put maximum effort into the development of the data request so that it rapidly becomes stable at the start of the project.

If, as in PRIMAVERA, the data request is dependent on an external project then there will be less control over the data request.

There was a balanced range of answers regarding how easy it was to include the requested diagnostics in groups' models. There were many new high-temporal frequency variables in the PRIMAVERA and HighResMIP data requests and so these may have contributed to some of this difficulty.

There was a tendency for the groups to find the data conversion process at their institute difficult. Many groups used the external CMOR software to run this process. The volume of data to convert will have added to these difficulties. It is not known how this process can be made easier. There was significant software development performed at the Met Office to assist in the conversion process and so this will hopefully become easier in future projects.

Most groups found the upload of data to JASMIN relatively easy, but a minority found it difficult or very difficult. No specific feedback was received in the comments suggesting the cause of this. Informal feedback received during the project suggested that difficulties could have been due to the network and facilities at data providers, or due to times when

the data transfer servers at JASMIN were unstable and connections dropped during transfers.

The comments received from the data providers are included in Appendix A. The comment about the occasional slowness or responsiveness of the DMT is due to the known issue.

		Querying the data request to see which variables were scheduled to be produced.	Use of the DMT to discover and query the PRIMAVERA data available at JASMIN.	Retrieving data from tape to group workspace/disk.	Accessing the data on the JASMIN group workspaces.	Analysing the data at JASMIN?
Score (1 is very difficult and 7 is very easy)	**1**	0	1	1	1	1
	2	1	1	0	1	0
	3	0	0	0	2	1
	4	0	0	1	0	2
	5	3	1	1	0	0
	6	7	11	10	7	8
	7	4	7	6	10	4
# Responses		15	21	19	21	16
Mean		**5.8**	**5.9**	**5.9**	**5.8**	**5.5**

Table 6 - Number of responses for each score from the users of PRIMAVERA facilities at JASMIN.

Table 6 shows the results from the questions asked to the users of PRIMAVERA's facilities at JASMIN. The answers are predominantly towards the higher end of the easy to use scale. The users believe that the DMT and accessing and analysing data at JASMIN was relatively easy. A recommendation for future projects is:

- Projects use central facilities such as JASMIN to store and analyse their data. The use of such a facility should be coordinated by a tool such as the DMT.

There were two users who found the DMT very difficult to use. The only feedback received during the project was about the occasional slowness of the DMT due to its known issue. There were five users who reported having varying levels of difficulty accessing or analysing the data at JASMIN but there is not enough information to understand these difficulties. There was help documentation available at https://help.jasmin.ac.uk/, videos produced by PRIMAVERA and email support available from JASMIN and WP9. JASMIN is a complex system and so these difficulties are not unexpected. This is a reminder that there can never be too much help and documentation available.

The comments received from the users in the survey suggest several useful improvements that could be made to the DMT.

		The documentation for accessing and using JASMIN	The email support from the CEDA helpdesk for accessing and using JASMIN
Score (1 is not at all useful and 5 is very useful)	1	1	0
	2	0	0
	3	2	2
	4	4	6
	5	13	11
# Responses		20	19
Mean		4.4	4.5

Table 7 - Number of responses for each score from the users about the JASMIN support.

Table 7 shows the responses to two questions about the support received from JASMIN. These two questions are scored differently to the previous questions with 1 being not at all useful and 5 being very useful. In general, the JASMIN documentation and support was scored neutrally or very useful.

4 Lessons Learnt

Please see the recommendations made in the bullet points throughout Section 3. The recommendations are also summarised in Section 1.

5 Links Built

WP9 has been working closely with WPs 4, 5 and 6 throughout the project to generate and upload the data. WP9 has been working with and supporting the users in all work packages throughout the project. The DMT was demonstrated at several conferences and the audience were invited to download and try the DMT software themselves.

References

CEDA: CMIP6 CEDA ESGF Node, [online] Available from: https://esgf-index1.ceda.ac.uk/search/cmip6-ceda/ (Accessed 2 June 2020), 2020.

Django: Django, [online] Available from: https://djangoproject.com/, 2019.

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., Mizuta, R., Nobre, P., Satoh, M., Scoccimarro, E., Semmler, T., Small, J. and von Storch, J.-S.: High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6, Geosci. Model Dev., 9(11), 4185–4208, doi:10.5194/gmd-9-4185-2016, 2016.

Juckes, M., Taylor, K. E., Durack, P. J., Lawrence, B., Mizielinski, M. S., Pamment, A., Peterschmitt, J.-Y., Rixen, M. and Sénési, S.: The CMIP6 Data Request (DREQ, version 01.00.31), Geosci. Model Dev., 13(1), 201–224, doi:10.5194/gmd-13-201-2020, 2020.

Mizielinski, M. S., Stephens, A., van der Linden, P., Bretonniere, P. A., Fiore, S., von Hardenberg, J., Kolax, M., Lohmann, K., Moine, M.-P., Le Sager, P., Semmler, T. and Senan, R.: Milestone 22 HPC Plan for Stream 1, , doi:10.5281/ZENODO.3929743, 2016.

Nadeau, D., Seddon, J., Vegas-Regidor, J., Kettleborough, J. and Hogan, E.: PRIMAVERA-H2020/cmip6-cmor-tables: Version 01.00.23, , doi:10.5281/ZENODO.1245672, 2018.

Nadeau, D., Doutriaux, C., mauzey1, Hogan, E., Kettleborough, J., kjoti, TobiasWeigel, Durack, P. J., Nicholls, Z., jmrgonza, wachsylon, taylor13, Seddon, J. and Betts, E.: PCMDI/cmor: 3.5.0, , doi:10.5281/ZENODO.3355583, 2019.

PRIMAVERA: Grant Agreement number: 641727 - PRocess-based climate sIMulation: AdVances in high resolution modelling and European climate Risk Assessment (PRIMAVERA), , doi:10.5281/ZENODO.3874429, 2015.

Seddon, J.: PRIMAVERA-H2020/dreq_tools: Upgrade PRIMAVERA 01.00.07 to HighResMIP 01.00.13, , doi:10.5281/ZENODO.3903546, 2020a.

Seddon, J.: PRIMAVERA-H2020/pre-proc: Initial Zenodo release, , doi:10.5281/ZENODO.3904597, 2020b.

Seddon, J.: PRIMAVERA-H2020/primavera-val: Initial release, , doi:10.5281/ZENODO.3596772, 2020c.

Seddon, J.: PRIMAVERA-H2020/stream2-planning: Data Request Summary, , doi:10.5281/zenodo.3921887, 2020d.

Seddon, J. and Stephens, A.: PRIMAVERA-H2020/primavera-dmt: Updated REAMDE.md, , doi:10.5281/zenodo.3596017, 2020.

Seddon, J., Stephens, A. and Mizielinski, M. S.: PRIMAVERA multi climate model analysis at the JASMIN Super Data Cluster, Submitt. to Geosci. Model Dev., 2020.

Stevens, B., Fiedler, S., Kinne, S., Peters, K., Rast, S., Müsse, J., Smith, S. J. and Mauritsen, T.: MACv2-SP: a parameterization of anthropogenic aerosol optical properties and an associated Twomey effect for use in CMIP6, Geosci. Model Dev., 10(1), 433–452, doi:10.5194/gmd-10-433-2017, 2017.

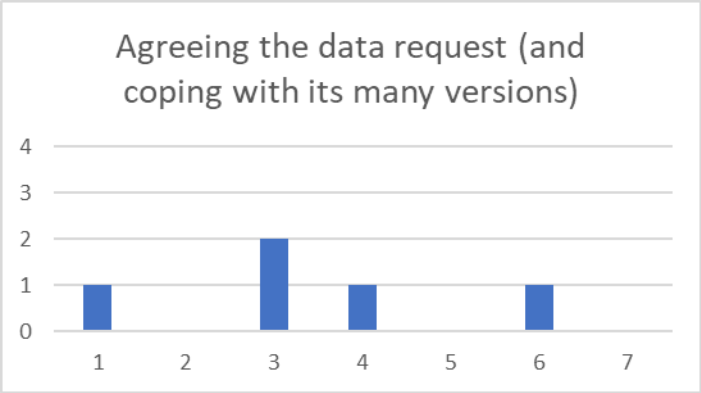
Appendix A – Data Management Plan User Survey

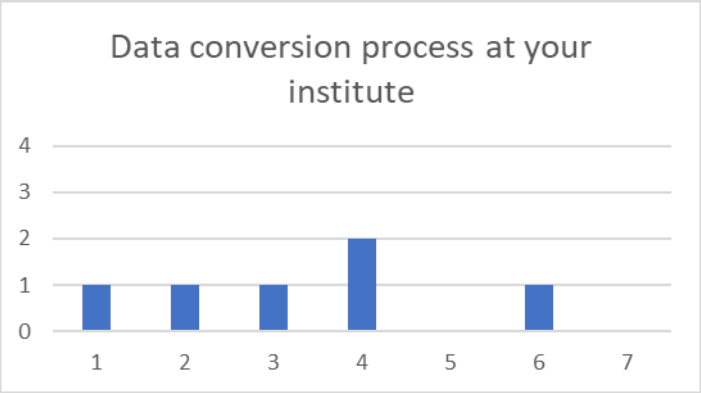
Data Providers

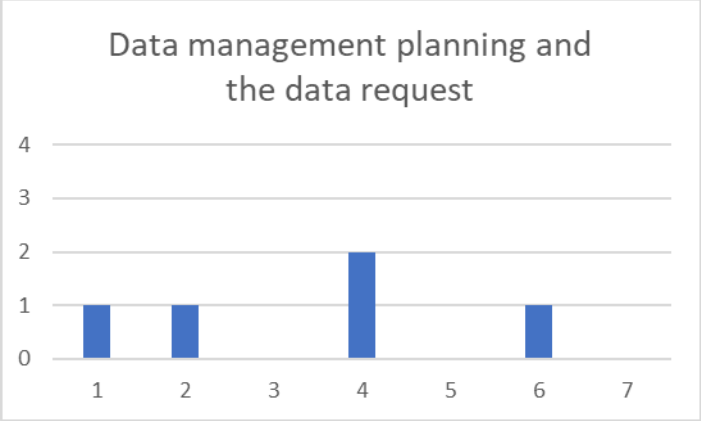
Comments received:

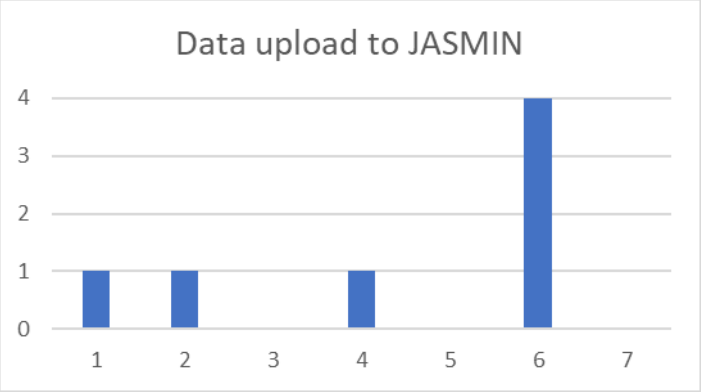
JASMIN/LOTUS system is great. The DMT was sometimes slow or unresponsive, but otherwise worked well.
list of all available models in the DMT filter (sorry, if this is there by now, I have not used the DMT for some time)
The DMT setup has been a great success and I'm very pleased with it.
The method and services deployed for provision of data to PRIMAVERA (DMP, DMT, personal support) were really performing, efficient and well suited/dimensioned. It really helped to make the life of data providers easier. DMT is a super tool! The ways for improvement depend mostly of external (upstream) factors: a pharaonic CMIP6 Data Request, the difficulty to anticipate data volumes 4 years in advance in the DMP (i.e. WP6/Stream 2) and to have a clear idea of all human workload it will implies. We maybe should pay ever strong attention, give more weight to Data aspect when writing the initial project proposal, systematically over-estimating the disk space and human resources needed.

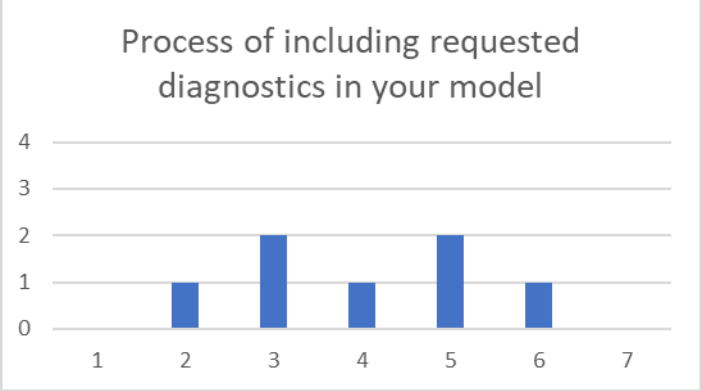
		Agreeing the data request (and coping with its many versions)	Data management planning and the data request	Process of including requested diagnostics in your model	Data conversion process at your institute	Data upload to JASMIN
Score (1 is very difficult and 7 is very	**1**	1	1	0	1	1
	2	0	1	1	1	1
	3	2	0	2	1	0
	4	1	2	1	2	1
	5	0	0	2	0	0
	6	1	1	1	1	4
	7	0	0	0	0	0
# Responses		5	5	7	6	7
Mean		**3.4**	**3.4**	**4.0**	**3.3**	**4.4**


Agreeing the data request (and coping with its many versions)
4
3
2
1
0
1
2
3
4
5
6
7


Data conversion process at your institute
4
3
2
1
0
1
2
3
4
5
6
7


Data management planning and the data request
4
3
2
1
0
1
2
3
4
5
6
7


Data upload to JASMIN
4
3
2
1
0
1
2
3
4
5
6
7


Process of including requested diagnostics in your model
4
3
2
1
0
1
2
3
4
5
6
7

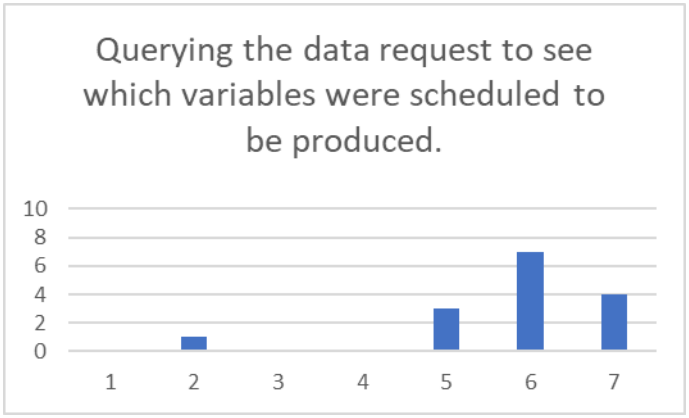
Data Users

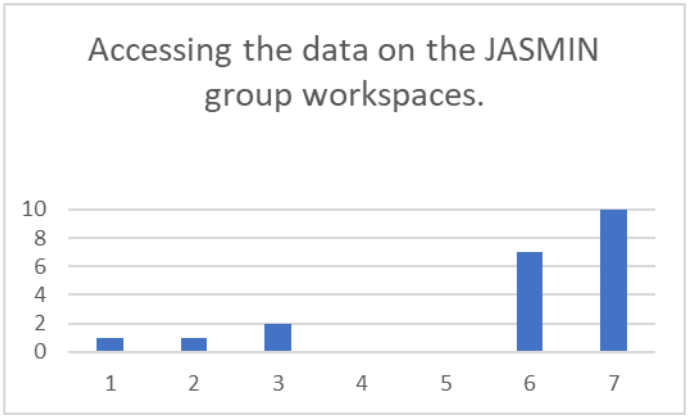
Comments received:

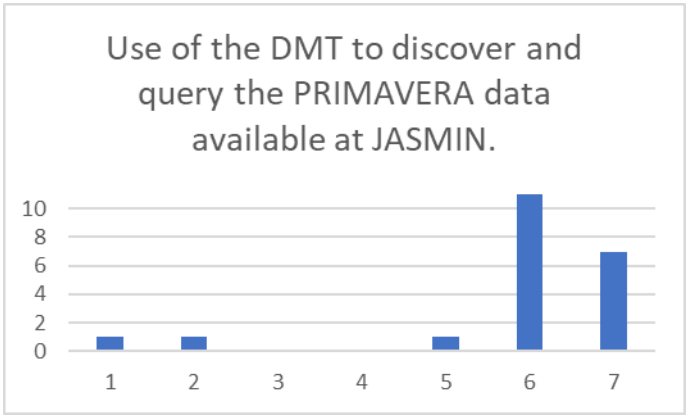
The only two inconveniences I had when using the data were: 1) intermittence in the DMT working to retrieve data 2) being asked to limit the storage space used to analyse the data. I worked with high-frequency data (hourly to 6hr) but only over Europe, so I tended to retrieve and 'crop' and store those files to read multiple times and for multiple things. I do most of the with python and cdo but it takes a long time, so re-doing it every time I needed to use the data was not an option. In my view, two ways to improve this would be to either have more storage available, or to implement a more complex and stable DMT that allows for regional selection. Otherwise, both the CEDA jasmin helpdesk and Jon were very helpful and fast to reply when I needed anything. Thanks!
A well-visible summary of model changes prior to publication on ESGF would be useful to inform users which simulations they should use for particular purposes. For example, the physics changes in coupled EC-Earth3P simulations (p1 to p2) will impact various analyses and discouraging use of the p1 simulations might be advisable in many cases.
I found a little bit annoying asking for several files to be retrieved from the tape. This is because I used to loose the checked boxes when going ahead to another page. However, it can be easily fixed by completing the process for each page.
After taking part to many EU projects, PRIMAVERA was the only one which provided a very efficient data management infrastructure (JASMIN). The DMT led to produce an impressive amount of scientific works, which I doubt would have been delivered without it. I have no specific advice on how to improve JASMIN, but I strongly recommend that other EU future projects follow the PRIMAVERA example.
Searching by text string would be useful when there are multiple similar parameters. (E.g “humidity” would match on specific and relative humidity).
keep the PRIMAVERA DMT alive after July ;-)
it would have been nice to have more reanalysis data side to side with the models
Don't know if possible, but being given a ballpark estimate of how long a given data request would take to restore to disk would be very neat!

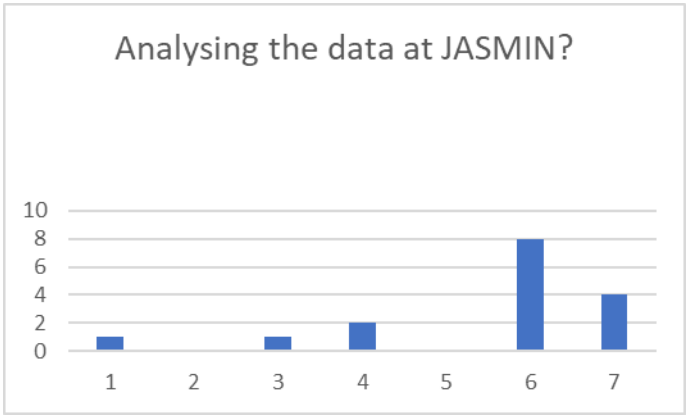
1) A common platform across Europe was big step forward in enabling collaboration. Hopefully we can capitalize on this even more next time. 2) There is some spin-up, in finding new ways to analyze/ process data on this remote platform. 3) the DMT worked well and Jon was v. swift and helpful with any queries. 4) Full ocean grids take up a lot of space. I used CDFtools to calculate ocean heat transport. This needed full ocean grids T, S, V and created large intermediate files as part of the process. So, having heat transport and AMOC output as standard could streamline effort and space. 5) Probably, speaking for myself, scientist could be better at deleting data in their group work space. Typically, data is saved as part of the analysis, but not all of it needs to be kept. I think, what is important is that once scientists have confidence in the robustness/ reliability of the DMT, - and that data can relatively easily re-retrieved and a calculation could ultimately be repeated, they will more readily delete unnecessary data. In this regard I think the success of the DMT will hopefully help the scientist to develop better habits of delete unneeded data more swiftly. 6) thanks to Jon, Great work.
In some cases there were different paths for different models/experiments, which made it difficult to find all data. standardising this would help a lot when running scripts on the original data

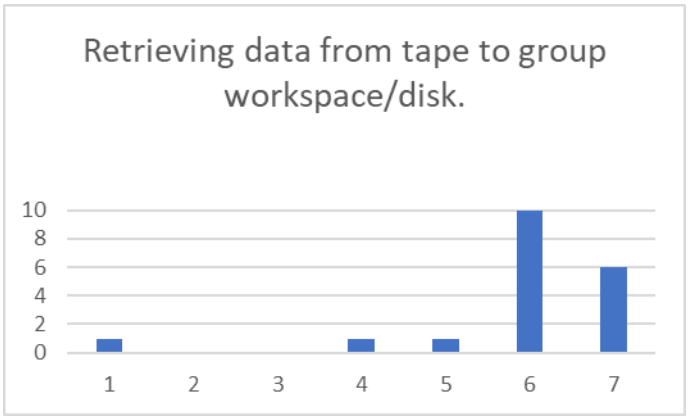
		Querying the data request to see which variables were scheduled to be produced.	Use of the DMT to discover and query the PRIMAVERA data available at JASMIN.	Retrieving data from tape to group workspace/disk.	Accessing the data on the JASMIN group workspaces.	Analysing the data at JASMIN?
Score (1 is very difficult and 7 is very easy)	**1**	0	1	1	1	1
	2	1	1	0	1	0
	3	0	0	0	2	1
	4	0	0	1	0	2
	5	3	1	1	0	0
	6	7	11	10	7	8
	7	4	7	6	10	4
# Responses		15	21	19	21	16
Mean		**5.8**	**5.9**	**5.9**	**5.8**	**5.5**


Querying the data request to see which variables were scheduled to be produced.
10
8
6
4
2
0
1
2
3
4
5
6
7

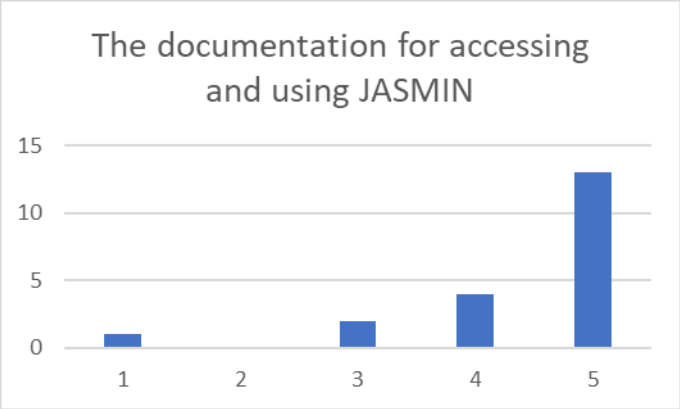

Accessing the data on the JASMIN group workspaces.
10
8
6
4
2
0
1
2
3
4
5
6
7

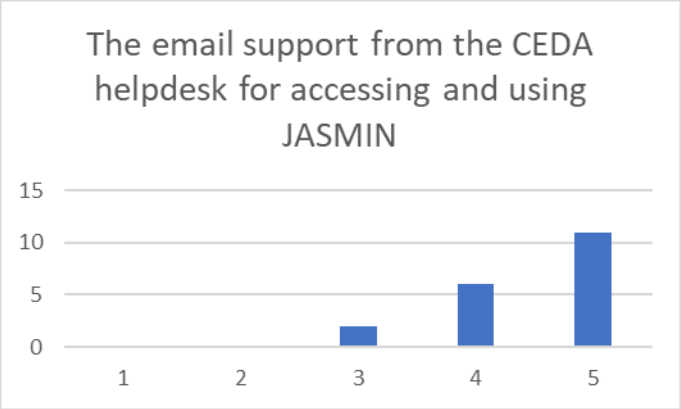

Use of the DMT to discover and query the PRIMAVERA data available at JASMIN.
10
8
6
4
2
0
1
2
3
4
5
6
7


Analysing the data at JASMIN?
10
8
6
4
2
0
1
2
3
4
5
6
7


Retrieving data from tape to group workspace/disk.
10
8
6
4
2
0
1
2
3
4
5
6
7

		The documentation for accessing and using JASMIN	The email support from the CEDA helpdesk for accessing and using JASMIN
Score (1 is not at all useful and 5 is very useful)	1	1	0
	2	0	0
	3	2	2
	4	4	6
	5	13	11
# Responses		20	19
Mean		**4.4**	**4.5**





Appendix B – Variables Retrieved at JASMIN

The following list shows the variables that have been requested for retrieval from tape to disk at JASMIN by the PRIMAVERA project sorted by frequency and table name. The number is the number of times that this variable has been requested. A value of 1 means that this variable has been requested once from a single model for a single experiment. 7 could mean that this variable had been requested from all models for a single experiment, or that the same variable from one model and experiment has been requested by seven different people. The name is in the form "`var-name_table-name`" and the table name includes the frequency of that variable. The variable's standard name is included in the third column.

The full dataset, including a list of variables that were never retrieved is shown in vars_retrieved_20200626.ipynb in (Seddon, 2020d).

pr_E1hr	28	precipitation_flux
clt_Prim1hr	4	cloud_area_fraction_in_atmosphere_layer
rsds_Prim1hr	8	surface_downwelling_shortwave_flux_in_air
rsdsdiff_Prim1hr	2	surface_diffuse_downwelling_shortwave_flux_in_air
tas_Prim1hr	18	air_temperature
ua50m_Prim1hr	4	eastward_wind
uas_Prim1hr	29	eastward_wind
va50m_Prim1hr	4	northward_wind
vas_Prim1hr	24	northward_wind
clt_3hr	2	cloud_area_fraction
hfls_3hr	43	surface_upward_latent_heat_flux
hfss_3hr	34	surface_upward_sensible_heat_flux
huss_3hr	48	specific_humidity
mrro_3hr	22	runoff_flux
mrsos_3hr	8	moisture_content_of_soil_layer
pr_3hr	146	precipitation_flux
prc_3hr	11	convective_precipitation_flux
prsn_3hr	9	snowfall_flux
ps_3hr	18	surface_air_pressure
rlds_3hr	45	surface_downwelling_longwave_flux_in_air
rldscs_3hr	30	surface_downwelling_longwave_flux_in_air_assuming_clear_sky
rlus_3hr	38	surface_upwelling_longwave_flux_in_air
rsds_3hr	15	surface_downwelling_shortwave_flux_in_air
rsdscs_3hr	3	surface_downwelling_shortwave_flux_in_air_assuming_clear_sky
rsus_3hr	33	surface_upwelling_shortwave_flux_in_air
tas_3hr	13	air_temperature
tslsi_3hr	1	surface_temperature
tso_3hr	1	sea_surface_temperature
uas_3hr	72	eastward_wind
vas_3hr	70	northward_wind
prw_E3hr	47	atmosphere_water_vapor_content
psl_E3hr	27	air_pressure_at_sea_level
rsdt_E3hr	2	toa_incoming_shortwave_flux
hus7h_E3hrPt	24	specific_humidity
psl_E3hrPt	8	air_pressure_at_sea_level
ta7h_E3hrPt	29	air_temperature
ua7h_E3hrPt	33	eastward_wind
ua850_E3hrPt	1	eastward_wind
va7h_E3hrPt	30	northward_wind
va850_E3hrPt	1	northward_wind
wap7h_E3hrPt	4	lagrangian_tendency_of_air_pressure
sfcWind_Prim3hr	17	wind_speed
sfcWindmax_Prim3hr	1	wind_speed
ua100m_Prim3hrPt	3	eastward_wind
ua50m_Prim3hrPt	1	eastward_wind
va100m_Prim3hrPt	3	northward_wind
va50m_Prim3hrPt	1	northward_wind
zg7h_Prim3hrPt	111	geopotential_height
tos_3hr	1	sea_surface_temperature
psl_CF3hr	1	air_pressure_at_sea_level
psl_6hrPlev	14	air_pressure_at_sea_level

hus7h_6hrPlevPt	37	specific_humidity
huss_6hrPlevPt	21	specific_humidity
psl_6hrPlevPt	100	air_pressure_at_sea_level
sfcWind_6hrPlevPt	38	wind_speed
ta_6hrPlevPt	6	air_temperature
ta7h_6hrPlevPt	6	air_temperature
tas_6hrPlevPt	8	air_temperature
ts_6hrPlevPt	1	surface_temperature
ua_6hrPlevPt	67	eastward_wind
ua7h_6hrPlevPt	52	eastward_wind
uas_6hrPlevPt	110	eastward_wind
va_6hrPlevPt	65	northward_wind
va7h_6hrPlevPt	48	northward_wind
vas_6hrPlevPt	112	northward_wind
vortmean_6hrPlevPt	1	atmosphere_relative_vorticity
zg27_6hrPlevPt	27	geopotential_height
zg7h_6hrPlevPt	107	geopotential_height
clt_Prim6hr	2	cloud_area_fraction
hfls_Prim6hr	2	surface_upward_latent_heat_flux
hus4_Prim6hr	2	specific_humidity
pr_Prim6hr	96	precipitation_flux
prc_Prim6hr	52	convective_precipitation_flux
ps_Prim6hr	15	surface_air_pressure
rsds_Prim6hr	1	surface_downwelling_shortwave_flux_in_air
sfcWindmax_Prim6hr	1	wind_speed_of_gust
ua4_Prim6hr	2	eastward_wind
va4_Prim6hr	3	northward_wind
wsgmax_Prim6hr	7	wind_speed_of_gust
clt_Prim6hrPt	2	cloud_area_fraction
ps_Prim6hrPt	4	surface_air_pressure
ua1000_Prim6hrPt	1	eastward_wind
clivi_CFday	6	atmosphere_cloud_ice_content
clw_CFday	3	mass_fraction_of_cloud_liquid_water_in_air
clwvi_CFday	8	atmosphere_cloud_condensed_water_content
phalf_CFday	3	air_pressure
ps_CFday	10	surface_air_pressure
rlutcs_CFday	2	toa_outgoing_longwave_flux_assuming_clear_sky
rsdt_CFday	5	toa_incoming_shortwave_flux
rsut_CFday	5	toa_outgoing_shortwave_flux
rsutcs_CFday	2	toa_outgoing_shortwave_flux_assuming_clear_sky
ua_CFday	4	eastward_wind
va_CFday	4	northward_wind
tauu_Eday	4	surface_downward_eastward_stress
ua_Eday	2	eastward_wind
va_Eday	1	northward_wind
zg_EdayZ	8	geopotential_height
tos_Oday	22	sea_surface_temperature
uo_PrimOday	1	sea_water_x_velocity
vo_PrimOday	2	sea_water_y_velocity
zos_PrimOday	6	sea_surface_height_above_geoid
evspsbl_Primday	9	water_evaporation_flux
hus10_Primday	1	specific_humidity
ta23_Primday	8	air_temperature
ts_Primday	46	surface_temperature
ua10_Primday	1	eastward_wind
ua23_Primday	9	eastward_wind
va10_Primday	1	northward_wind
va23_Primday	8	northward_wind
zg10_Primday	14	geopotential_height
zg23_Primday	121	geopotential_height
ua_PrimdayPt	2	eastward_wind
siconc_SIday	5	sea_ice_area_fraction
siu_SIday	17	sea_ice_x_velocity
siv_SIday	18	sea_ice_y_velocity
clt_day	18	cloud_area_fraction
hfls_day	31	surface_upward_latent_heat_flux
hfss_day	19	surface_upward_sensible_heat_flux
hus_day	61	specific_humidity
huss_day	25	specific_humidity
mrro_day	26	runoff_flux
mrsos_day	18	moisture_content_of_soil_layer
pr_day	367	precipitation_flux

psl_day	72	air_pressure_at_sea_level
rlds_day	15	surface_downwelling_longwave_flux_in_air
rlut_day	24	toa_outgoing_longwave_flux
sfcWind_day	31	wind_speed
sfcWindmax_day	62	wind_speed
snw_day	4	surface_snow_amount
ta_day	85	air_temperature
tas_day	74	air_temperature
tasmax_day	152	air_temperature
tasmin_day	64	air_temperature
ua_day	194	eastward_wind
uas_day	81	eastward_wind
va_day	161	northward_wind
vas_day	84	northward_wind
wap_day	18	lagrangian_tendency_of_air_pressure
zg_day	544	geopotential_height
sftlf_fx	1	land_area_fraction
cl_Amon	20	cloud_area_fraction_in_atmosphere_layer
cli_Amon	4	mass_fraction_of_cloud_ice_in_air
clt_Amon	314	cloud_area_fraction
clw_Amon	4	mass_fraction_of_cloud_liquid_water_in_air
evspsbl_Amon	329	water_evaporation_flux
hfls_Amon	236	surface_upward_latent_heat_flux
hfss_Amon	234	surface_upward_sensible_heat_flux
hur_Amon	34	relative_humidity
hurs_Amon	3	relative_humidity
hus_Amon	333	specific_humidity
huss_Amon	323	specific_humidity
pfull_Amon	6	air_pressure
phalf_Amon	5	air_pressure
pr_Amon	1261	precipitation_flux
prc_Amon	68	convective_precipitation_flux
prsn_Amon	32	snowfall_flux
prw_Amon	87	atmosphere_water_vapor_content
ps_Amon	322	surface_air_pressure
psl_Amon	828	air_pressure_at_sea_level
rlds_Amon	202	surface_downwelling_longwave_flux_in_air
rldscs_Amon	47	surface_downwelling_longwave_flux_in_air_assuming_clear_sky
rlus_Amon	131	surface_upwelling_longwave_flux_in_air
rlut_Amon	83	toa_outgoing_longwave_flux
rlutcs_Amon	42	toa_outgoing_longwave_flux_assuming_clear_sky
rsds_Amon	178	surface_downwelling_shortwave_flux_in_air
rsdscs_Amon	45	surface_downwelling_shortwave_flux_in_air_assuming_clear_sky
rsdt_Amon	82	toa_incoming_shortwave_flux
rsus_Amon	172	surface_upwelling_shortwave_flux_in_air
rsuscs_Amon	33	surface_upwelling_shortwave_flux_in_air_assuming_clear_sky
rsut_Amon	88	toa_outgoing_shortwave_flux
rsutcs_Amon	43	toa_outgoing_shortwave_flux_assuming_clear_sky
sfcWind_Amon	3	wind_speed
ta_Amon	354	air_temperature
tas_Amon	692	air_temperature
tasmax_Amon	25	air_temperature
tasmin_Amon	23	air_temperature
tauu_Amon	26	surface_downward_eastward_stress
tauv_Amon	26	surface_downward_northward_stress
ts_Amon	834	surface_temperature
ua_Amon	513	eastward_wind
uas_Amon	29	eastward_wind
va_Amon	417	northward_wind
vas_Amon	32	northward_wind
wap_Amon	144	lagrangian_tendency_of_air_pressure
zg_Amon	361	geopotential_height
hur_CFmon	1	relative_humidity
hus_CFmon	1	specific_humidity
rld_CFmon	1	downwelling_longwave_flux_in_air
rldcs_CFmon	1	downwelling_longwave_flux_in_air_assuming_clear_sky
rlu_CFmon	1	upwelling_longwave_flux_in_air
rlucs_CFmon	4	upwelling_longwave_flux_in_air_assuming_clear_sky
rsd_CFmon	1	downwelling_shortwave_flux_in_air
rsdcs_CFmon	1	downwelling_shortwave_flux_in_air_assuming_clear_sky
rsu_CFmon	1	upwelling_shortwave_flux_in_air
rsucs_CFmon	1	upwelling_shortwave_flux_in_air_assuming_clear_sky

ta_CFmon	1	air_temperature
uqint_Emon	7	integral_of_product_of_eastward_wind_and_specific_humidity_wrt_height
vqint_Emon	7	integral_of_product_of_northward_wind_and_specific_humidity_wrt_height
epfz_EmonZ	28	upward_eliassen_palm_flux_in_air
snc_LImon	5	surface_snow_area_fraction
snd_LImon	34	surface_snow_thickness
snw_LImon	7	surface_snow_amount
tsn_LImon	1	temperature_in_surface_snow
mrlsl_Lmon	34	moisture_content_of_soil_layer
mrro_Lmon	38	runoff_flux
mrros_Lmon	5	surface_runoff_flux
mrso_Lmon	72	soil_moisture_content
mrsos_Lmon	7	moisture_content_of_soil_layer
tran_Lmon	1	transpiration_flux
ficeberg_Omon	1	water_flux_into_sea_water_from_icebergs
friver_Omon	7	water_flux_into_sea_water_from_rivers
hfbasin_Omon	10	northward_ocean_heat_transport
hfcorr_Omon	14	heat_flux_correction
hfds_Omon	66	surface_downward_heat_flux_in_sea_water
hfx_Omon	1	ocean_heat_x_transport
hfy_Omon	1	ocean_heat_y_transport
htovgyre_Omon	2	northward_ocean_heat_transport_due_to_gyre
htovovrt_Omon	2	northward_ocean_heat_transport_due_to_overturning
mlotst_Omon	217	ocean_mixed_layer_thickness_defined_by_sigma_t
msftbarot_Omon	4	ocean_barotropic_mass_streamfunction
msftmyz_Omon	11	ocean_meridional_overturning_mass_streamfunction
msftmz_Omon	5	ocean_meridional_overturning_mass_streamfunction
msftyz_Omon	2	ocean_y_overturning_mass_streamfunction
so_Omon	228	sea_water_salinity
soga_Omon	11	sea_water_salinity
sos_Omon	22	sea_surface_salinity
tauuo_Omon	150	surface_downward_x_stress
tauvo_Omon	4	surface_downward_y_stress
thetao_Omon	472	sea_water_potential_temperature
tos_Omon	228	sea_surface_temperature
umo_Omon	6	ocean_mass_x_transport
uo_Omon	318	sea_water_x_velocity
vmo_Omon	29	ocean_mass_y_transport
vo_Omon	443	sea_water_y_velocity
wmo_Omon	1	upward_ocean_mass_transport
zos_Omon	97	sea_surface_height_above_geoid
uso_PrimOmon	4	product_of_xward_sea_water_velocity_and_salinity
uto_PrimOmon	4	product_of_xward_sea_water_velocity_and_temperature
vso_PrimOmon	4	product_of_yward_sea_water_velocity_and_salinity
vto_PrimOmon	15	product_of_yward_sea_water_velocity_and_temperature
wo_PrimOmon	2	upward_sea_water_velocity
zg_PrimmonZ	2	geopotential_height
siconc_SImon	152	sea_ice_area_fraction
sispeed_SImon	9	sea_ice_speed
sithick_SImon	35	sea_ice_thickness
sitimefrac_SImon	2	sea_ice_time_fraction
siu_SImon	25	sea_ice_x_velocity
siv_SImon	98	sea_ice_y_velocity
sivol_SImon	104	sea_ice_thickness

Appendix C – Variables Retrieved at JASMIN in Popularity Order

The following list shows the same variables as in Appendix B that have been requested for retrieval from tape to disk at JASMIN by the PRIMAVERA project sorted by the number of retrievals.

pr_Amon	1261	precipitation_flux
ts_Amon	834	surface_temperature
psl_Amon	828	air_pressure_at_sea_level
tas_Amon	692	air_temperature
zg_day	544	geopotential_height
ua_Amon	513	eastward_wind
thetao_Omon	472	sea_water_potential_temperature
vo_Omon	443	sea_water_y_velocity
va_Amon	417	northward_wind
pr_day	367	precipitation_flux
zg_Amon	361	geopotential_height
ta_Amon	354	air_temperature
hus_Amon	333	specific_humidity
evspsbl_Amon	329	water_evaporation_flux
huss_Amon	323	specific_humidity
ps_Amon	322	surface_air_pressure
uo_Omon	318	sea_water_x_velocity
clt_Amon	314	cloud_area_fraction
hfls_Amon	236	surface_upward_latent_heat_flux
hfss_Amon	234	surface_upward_sensible_heat_flux
tos_Omon	228	sea_surface_temperature
so_Omon	228	sea_water_salinity
mlotst_Omon	217	ocean_mixed_layer_thickness_defined_by_sigma_t
rlds_Amon	202	surface_downwelling_longwave_flux_in_air
ua_day	194	eastward_wind
rsds_Amon	178	surface_downwelling_shortwave_flux_in_air
rsus_Amon	172	surface_upwelling_shortwave_flux_in_air
va_day	161	northward_wind
siconc_SImon	152	sea_ice_area_fraction
tasmax_day	152	air_temperature
tauuo_Omon	150	surface_downward_x_stress
pr_3hr	146	precipitation_flux
wap_Amon	144	lagrangian_tendency_of_air_pressure
rlus_Amon	131	surface_upwelling_longwave_flux_in_air
zg23_Primday	121	geopotential_height
vas_6hrPlevPt	112	northward_wind
zg7h_Prim3hrPt	111	geopotential_height
uas_6hrPlevPt	110	eastward_wind
zg7h_6hrPlevPt	107	geopotential_height
sivol_SImon	104	sea_ice_thickness
psl_6hrPlevPt	100	air_pressure_at_sea_level
siv_SImon	98	sea_ice_y_velocity
zos_Omon	97	sea_surface_height_above_geoid
pr_Prim6hr	96	precipitation_flux
rsut_Amon	88	toa_outgoing_shortwave_flux
prw_Amon	87	atmosphere_water_vapor_content
ta_day	85	air_temperature
vas_day	84	northward_wind
rlut_Amon	83	toa_outgoing_longwave_flux
rsdt_Amon	82	toa_incoming_shortwave_flux
uas_day	81	eastward_wind
tas_day	74	air_temperature
mrso_Lmon	72	soil_moisture_content
uas_3hr	72	eastward_wind
psl_day	72	air_pressure_at_sea_level
vas_3hr	70	northward_wind
prc_Amon	68	convective_precipitation_flux
ua_6hrPlevPt	67	eastward_wind
hfds_Omon	66	surface_downward_heat_flux_in_sea_water
va_6hrPlevPt	65	northward_wind
tasmin_day	64	air_temperature
sfcWindmax_day	62	wind_speed
hus_day	61	specific_humidity

ua7h_6hrPlevPt	52	eastward_wind
prc_Prim6hr	52	convective_precipitation_flux
va7h_6hrPlevPt	48	northward_wind
huss_3hr	48	specific_humidity
rldscs_Amon	47	surface_downwelling_longwave_flux_in_air_assuming_clear_sky
prw_E3hr	47	atmosphere_water_vapor_content
ts_Primday	46	surface_temperature
rsdscs_Amon	45	surface_downwelling_shortwave_flux_in_air_assuming_clear_sky
rlds_3hr	45	surface_downwelling_longwave_flux_in_air
rsutcs_Amon	43	toa_outgoing_shortwave_flux_assuming_clear_sky
hfls_3hr	43	surface_upward_latent_heat_flux
rlutcs_Amon	42	toa_outgoing_longwave_flux_assuming_clear_sky
sfcWind_6hrPlevPt	38	wind_speed
mrro_Lmon	38	runoff_flux
rlus_3hr	38	surface_upwelling_longwave_flux_in_air
hus7h_6hrPlevPt	37	specific_humidity
sithick_SImon	35	sea_ice_thickness
hur_Amon	34	relative_humidity
snd_LImon	34	surface_snow_thickness
mrlsl_Lmon	34	moisture_content_of_soil_layer
hfss_3hr	34	surface_upward_sensible_heat_flux
rsuscs_Amon	33	surface_upwelling_shortwave_flux_in_air_assuming_clear_sky
ua7h_E3hrPt	33	eastward_wind
rsus_3hr	33	surface_upwelling_shortwave_flux_in_air
prsn_Amon	32	snowfall_flux
vas_Amon	32	northward_wind
sfcWind_day	31	wind_speed
hfls_day	31	surface_upward_latent_heat_flux
va7h_E3hrPt	30	northward_wind
rldscs_3hr	30	surface_downwelling_longwave_flux_in_air_assuming_clear_sky
uas_Prim1hr	29	eastward_wind
uas_Amon	29	eastward_wind
ta7h_E3hrPt	29	air_temperature
vmo_Omon	29	ocean_mass_y_transport
pr_E1hr	28	precipitation_flux
epfz_EmonZ	28	upward_eliassen_palm_flux_in_air
psl_E3hr	27	air_pressure_at_sea_level
zg27_6hrPlevPt	27	geopotential_height
tauu_Amon	26	surface_downward_eastward_stress
tauv_Amon	26	surface_downward_northward_stress
mrro_day	26	runoff_flux
siu_SImon	25	sea_ice_x_velocity
huss_day	25	specific_humidity
tasmax_Amon	25	air_temperature
vas_Prim1hr	24	northward_wind
rlut_day	24	toa_outgoing_longwave_flux
hus7h_E3hrPt	24	specific_humidity
tasmin_Amon	23	air_temperature
tos_Oday	22	sea_surface_temperature
sos_Omon	22	sea_surface_salinity
mrro_3hr	22	runoff_flux
huss_6hrPlevPt	21	specific_humidity
cl_Amon	20	cloud_area_fraction_in_atmosphere_layer
hfss_day	19	surface_upward_sensible_heat_flux
tas_Prim1hr	18	air_temperature
siv_SIday	18	sea_ice_y_velocity
ps_3hr	18	surface_air_pressure
wap_day	18	lagrangian_tendency_of_air_pressure
mrsos_day	18	moisture_content_of_soil_layer
clt_day	18	cloud_area_fraction
siu_SIday	17	sea_ice_x_velocity
sfcWind_Prim3hr	17	wind_speed
rsds_3hr	15	surface_downwelling_shortwave_flux_in_air
ps_Prim6hr	15	surface_air_pressure
vto_PrimOmon	15	product_of_yward_sea_water_velocity_and_temperature
rlds_day	15	surface_downwelling_longwave_flux_in_air
psl_6hrPlev	14	air_pressure_at_sea_level
zg10_Primday	14	geopotential_height
hfcorr_Omon	14	heat_flux_correction
tas_3hr	13	air_temperature
prc_3hr	11	convective_precipitation_flux
msftmyz_Omon	11	ocean_meridional_overturning_mass_streamfunction

```
soga_Omon              11   sea_water_salinity
ps_CFday               10   surface_air_pressure
hfbasin_Omon           10   northward_ocean_heat_transport
sispeed_SImon           9   sea_ice_speed
evspsbl_Primday         9   water_evaporation_flux
ua23_Primday            9   eastward_wind
prsn_3hr                9   snowfall_flux
rsds_Prim1hr            8   surface_downwelling_shortwave_flux_in_air
clwvi_CFday             8   atmosphere_cloud_condensed_water_content
psl_E3hrPt              8   air_pressure_at_sea_level
zg_EdayZ                8   geopotential_height
tas_6hrPlevPt           8   air_temperature
mrsos_3hr               8   moisture_content_of_soil_layer
va23_Primday            8   northward_wind
ta23_Primday            8   air_temperature
snw_LImon               7   surface_snow_amount
friver_Omon             7   water_flux_into_sea_water_from_rivers
wsgmax_Prim6hr          7   wind_speed_of_gust
mrsos_Lmon              7   moisture_content_of_soil_layer
uqint_Emon              7
integral_of_product_of_eastward_wind_and_specific_humidity_wrt_height
vqint_Emon              7
integral_of_product_of_northward_wind_and_specific_humidity_wrt_height
ta_6hrPlevPt            6   air_temperature
ta7h_6hrPlevPt          6   air_temperature
pfull_Amon              6   air_pressure
clivi_CFday             6   atmosphere_cloud_ice_content
zos_PrimOday            6   sea_surface_height_above_geoid
umo_Omon                6   ocean_mass_x_transport
phalf_Amon              5   air_pressure
rsut_CFday              5   toa_outgoing_shortwave_flux
siconc_SIday            5   sea_ice_area_fraction
mrros_Lmon              5   surface_runoff_flux
rsdt_CFday              5   toa_incoming_shortwave_flux
msftmz_Omon             5   ocean_meridional_overturning_mass_streamfunction
snc_LImon               5   surface_snow_area_fraction
va50m_Prim1hr           4   northward_wind
ua50m_Prim1hr           4   eastward_wind
clt_Prim1hr             4   cloud_area_fraction_in_atmosphere_layer
clw_Amon                4   mass_fraction_of_cloud_liquid_water_in_air
cli_Amon                4   mass_fraction_of_cloud_ice_in_air
rlucs_CFmon             4   upwelling_longwave_flux_in_air_assuming_clear_sky
ua_CFday                4   eastward_wind
va_CFday                4   northward_wind
ps_Prim6hrPt            4   surface_air_pressure
wap7h_E3hrPt            4   lagrangian_tendency_of_air_pressure
tauvo_Omon              4   surface_downward_y_stress
snw_day                 4   surface_snow_amount
uto_PrimOmon            4   product_of_xward_sea_water_velocity_and_temperature
uso_PrimOmon            4   product_of_xward_sea_water_velocity_and_salinity
vso_PrimOmon            4   product_of_yward_sea_water_velocity_and_salinity
msftbarot_Omon          4   ocean_barotropic_mass_streamfunction
tauu_Eday               4   surface_downward_eastward_stress
ua100m_Prim3hrPt        3   eastward_wind
va100m_Prim3hrPt        3   northward_wind
rsdscs_3hr              3   surface_downwelling_shortwave_flux_in_air_assuming_clear_sky
hurs_Amon               3   relative_humidity
sfcWind_Amon            3   wind_speed
clw_CFday               3   mass_fraction_of_cloud_liquid_water_in_air
phalf_CFday             3   air_pressure
va4_Prim6hr             3   northward_wind
rsdsdiff_Prim1hr        2   surface_diffuse_downwelling_shortwave_flux_in_air
clt_Prim6hr             2   cloud_area_fraction
clt_Prim6hrPt           2   cloud_area_fraction
zg_PrimmonZ             2   geopotential_height
rlutcs_CFday            2   toa_outgoing_longwave_flux_assuming_clear_sky
rsutcs_CFday            2   toa_outgoing_shortwave_flux_assuming_clear_sky
hus4_Prim6hr            2   specific_humidity
ua4_Prim6hr             2   eastward_wind
clt_3hr                 2   cloud_area_fraction
htovgyre_Omon           2   northward_ocean_heat_transport_due_to_gyre
htovovrt_Omon           2   northward_ocean_heat_transport_due_to_overturning
```

hfls_Prim6hr	2	surface_upward_latent_heat_flux
msftyz_Omon	2	ocean_y_overturning_mass_streamfunction
vo_PrimOday	2	sea_water_y_velocity
ua_Eday	2	eastward_wind
sitimefrac_SImon	2	sea_ice_time_fraction
ua_PrimdayPt	2	eastward_wind
rsdt_E3hr	2	toa_incoming_shortwave_flux
wo_PrimOmon	2	upward_sea_water_velocity
ua50m_Prim3hrPt	1	eastward_wind
va50m_Prim3hrPt	1	northward_wind
tso_3hr	1	sea_surface_temperature
ts_6hrPlevPt	1	surface_temperature
ua850_E3hrPt	1	eastward_wind
va850_E3hrPt	1	northward_wind
ficeberg_Omon	1	water_flux_into_sea_water_from_icebergs
tran_Lmon	1	transpiration_flux
hus_CFmon	1	specific_humidity
hur_CFmon	1	relative_humidity
ta_CFmon	1	air_temperature
rsdcs_CFmon	1	downwelling_shortwave_flux_in_air_assuming_clear_sky
rldcs_CFmon	1	downwelling_longwave_flux_in_air_assuming_clear_sky
rsucs_CFmon	1	upwelling_shortwave_flux_in_air_assuming_clear_sky
rsd_CFmon	1	downwelling_shortwave_flux_in_air
rld_CFmon	1	downwelling_longwave_flux_in_air
rsu_CFmon	1	upwelling_shortwave_flux_in_air
rlu_CFmon	1	upwelling_longwave_flux_in_air
hfx_Omon	1	ocean_heat_x_transport
hfy_Omon	1	ocean_heat_y_transport
uo_PrimOday	1	sea_water_x_velocity
tsn_LImon	1	temperature_in_surface_snow
va_Eday	1	northward_wind
wmo_Omon	1	upward_ocean_mass_transport
va10_Primday	1	northward_wind
ua10_Primday	1	eastward_wind
hus10_Primday	1	specific_humidity
tos_3hr	1	sea_surface_temperature
tslsi_3hr	1	surface_temperature
ua1000_Prim6hrPt	1	eastward_wind
psl_CF3hr	1	air_pressure_at_sea_level
sftlf_fx	1	land_area_fraction
sfcWindmax_Prim3hr	1	wind_speed
sfcWindmax_Prim6hr	1	wind_speed_of_gust
rsds_Prim6hr	1	surface_downwelling_shortwave_flux_in_air
vortmean_6hrPlevPt	1	atmosphere_relative_vorticity

Appendix D – ESGF Variables Downloaded from JASMIN

The following list shows the CMIP6.HighResMIP variables that have been downloaded from CEDA' s ESGF node via HTTP between 25th March 2019 and 14th May 2020. The variables have been sorted by frequency and table name. The number shows the number of times that unique datasets (e.g. a unique combination of institute, model, experiment, variant label, table and variable) have been downloaded from a unique IP address.

pr_E1hr	48
prc_E1hr	38
clt_3hr	7
hfls_3hr	4
hfss_3hr	6
huss_3hr	38
mrsos_3hr	1
pr_3hr	107
prc_3hr	7
prsn_3hr	5
ps_3hr	3
rlds_3hr	13
rldscs_3hr	1
rlus_3hr	14
rsds_3hr	13
rsdscs_3hr	12
rsus_3hr	12
rsuscs_3hr	14
tas_3hr	77
tos_3hr	5
tslsi_3hr	4
uas_3hr	63
vas_3hr	60
psl_CF3hr	3
prcsh_E3hr	2
prw_E3hr	14
psl_E3hr	36
rlut_E3hr	1
rlutcs_E3hr	1
rsut_E3hr	1
rsutcs_E3hr	9
hus_E3hrPt	7
ta_E3hrPt	7
ua_E3hrPt	11
va_E3hrPt	9
psl_6hrPlev	40
wap_6hrPlev	2
hus_6hrPlevPt	25
huss_6hrPlevPt	2
psl_6hrPlevPt	95
rv850_6hrPlevPt	5
sfcWind_6hrPlevPt	2
ta_6hrPlevPt	63
tas_6hrPlevPt	72
ts_6hrPlevPt	3
ua_6hrPlevPt	79
uas_6hrPlevPt	90
va_6hrPlevPt	84
vas_6hrPlevPt	94
wbptemp_6hrPlevPt	16
zg_6hrPlevPt	31
zg500_6hrPlevPt	7
albisccp_CFday	9
ccb_CFday	9
cct_CFday	9
cl_CFday	15
clcalipso_CFday	3
clhcalipso_CFday	3

cli_CFday	9
clisccp_CFday	4
clivi_CFday	4
cllcalipso_CFday	2
clmcalipso_CFday	3
cltcalipso_CFday	7
cltisccp_CFday	10
clw_CFday	12
clwvi_CFday	4
hur_CFday	3
hus_CFday	4
mc_CFday	5
pctisccp_CFday	9
pfull_CFday	18
phalf_CFday	16
ps_CFday	40
rldscs_CFday	10
rlutcs_CFday	15
rsdscs_CFday	15
rsdt_CFday	15
rsuscs_CFday	14
rsut_CFday	18
rsutcs_CFday	17
ta_CFday	5
ta700_CFday	19
ua_CFday	5
va_CFday	3
wap_CFday	3
wap500_CFday	9
zg_CFday	5
parasolRefl_Eday	6
rivo_Eday	10
ta_Eday	1
tauu_Eday	18
tauv_Eday	16
ts_Eday	15
ua_Eday	4
va_Eday	4
epfz_EdayZ	4
utendnogw_EdayZ	1
utendogw_EdayZ	2
vtem_EdayZ	4
wtem_EdayZ	4
zg_EdayZ	2
omldamax_Oday	5
sos_Oday	9
tos_Oday	54
tossq_Oday	7
siconc_SIday	26
siconca_SIday	1
sisnthick_SIday	6
sitemptop_SIday	7
sithick_SIday	20
siu_SIday	9
siv_SIday	10
clt_day	29
hfls_day	40
hfss_day	39
hur_day	42
hurs_day	48
hursmax_day	12
hursmin_day	10
hus_day	158
huss_day	98
mrro_day	43
mrso_day	13
mrsos_day	17
pr_day	545
prc_day	34
prsn_day	37
psl_day	155
rlds_day	55

rlus_day	54
rlut_day	44
rsds_day	81
rsus_day	47
sfcWind_day	68
sfcWindmax_day	20
snc_day	6
snw_day	9
ta_day	168
tas_day	235
tasmax_day	218
tasmin_day	203
tslsi_day	21
ua_day	163
uas_day	152
va_day	145
vas_day	139
wap_day	38
zg_day	96
abs550aer_AERmon	1
cltc_AERmon	1
cod_AERmon	1
lwp_AERmon	1
od550aer_AERmon	2
od550dust_AERmon	1
od550oa_AERmon	3
od550so4_AERmon	1
od550ss_AERmon	1
ptp_AERmon	7
reffclwtop_AERmon	25
rlutaf_AERmon	6
rlutcsaf_AERmon	1
rsutaf_AERmon	5
rsutcsaf_AERmon	5
tatp_AERmon	1
toz_AERmon	5
ttop_AERmon	1
ua_AERmon	2
va_AERmon	1
ztp_AERmon	1
ccb_Amon	30
cct_Amon	28
ch4global_Amon	4
ci_Amon	26
cl_Amon	61
cli_Amon	44
clivi_Amon	50
clt_Amon	82
clw_Amon	44
clwvi_Amon	47
co2mass_Amon	6
evspsbl_Amon	133
hfls_Amon	109
hfss_Amon	109
hur_Amon	91
hurs_Amon	85
hus_Amon	223
huss_Amon	144
mc_Amon	5
n2oglobal_Amon	4
pfull_Amon	14
phalf_Amon	14
pr_Amon	716
prc_Amon	88
prsn_Amon	75
prw_Amon	121
ps_Amon	214
psl_Amon	342
rlds_Amon	119
rldscs_Amon	70
rlus_Amon	98
rlut_Amon	94

rlutcs_Amon	59
rsds_Amon	119
rsdscs_Amon	69
rsdt_Amon	108
rsus_Amon	97
rsuscs_Amon	70
rsut_Amon	96
rsutcs_Amon	58
rtmt_Amon	8
sbl_Amon	29
sci_Amon	21
sfcWind_Amon	80
ta_Amon	303
tas_Amon	627
tasmax_Amon	104
tasmin_Amon	105
tauu_Amon	111
tauv_Amon	109
ts_Amon	294
ua_Amon	317
uas_Amon	291
va_Amon	289
vas_Amon	281
wap_Amon	136
zg_Amon	259
albisccp_CFmon	22
clis_CFmon	1
clisccp_CFmon	1
cllcalipso_CFmon	1
cltcalipso_CFmon	22
cltisccp_CFmon	24
hur_CFmon	27
hus_CFmon	39
pctisccp_CFmon	23
rld_CFmon	26
rldcs_CFmon	23
rlu_CFmon	23
rlucs_CFmon	23
rlucs4co2_CFmon	1
rlut4co2_CFmon	5
rsd_CFmon	23
rsdcs_CFmon	21
rsdcs4co2_CFmon	2
rsu_CFmon	23
rsu4co2_CFmon	2
rsucs_CFmon	24
rsut4co2_CFmon	4
ta_CFmon	37
tnhus_CFmon	2
tnhusa_CFmon	1
tnhusc_CFmon	1
tnhusmp_CFmon	1
tnhusscpbl_CFmon	1
tntscpbl_CFmon	2
evspsblpot_Emon	3
hcont300_Emon	1
hus_Emon	9
intuadse_Emon	2
intuaw_Emon	1
intvadse_Emon	3
intvaw_Emon	1
mrsol_Emon	2
mrtws_Emon	4
nep_Emon	3
parasolRefl_Emon	20
sfcWindmax_Emon	2
t20d_Emon	1
ta_Emon	8
ua_Emon	9
uqint_Emon	25
va_Emon	9
vqint_Emon	22

wtd_Emon	1
zg_Emon	8
epfz_EmonZ	17
vtem_EmonZ	16
wtem_EmonZ	17
agesno_LImon	1
hfdsn_LImon	1
lwsnl_LImon	3
sbl_LImon	3
snc_LImon	31
snd_LImon	66
snm_LImon	48
snw_LImon	53
tsn_LImon	34
baresoilFrac_Lmon	2
c4PftFrac_Lmon	1
cropFrac_Lmon	2
evspsblsoi_Lmon	21
evspsblveg_Lmon	34
gpp_Lmon	37
grassFrac_Lmon	1
lai_Lmon	37
mrfso_Lmon	4
mrro_Lmon	130
mrros_Lmon	78
mrso_Lmon	92
mrsos_Lmon	35
npp_Lmon	43
ra_Lmon	34
rh_Lmon	45
tran_Lmon	10
treeFrac_Lmon	1
tsl_Lmon	95
agessc_Omon	1
bigthetao_Omon	2
bigthetaoga_Omon	1
hfbasinpmadv_Omon	1
hfds_Omon	30
hfx_Omon	1
hfy_Omon	2
htovovrt_Omon	1
masso_Omon	1
mlotst_Omon	28
mlotstsq_Omon	4
msftmz_Omon	2
obvfsq_Omon	1
pbo_Omon	3
rsdo_Omon	1
rsntds_Omon	8
sltovgyre_Omon	2
sltovovrt_Omon	1
so_Omon	28
soga_Omon	1
sos_Omon	20
tauuo_Omon	27
tauvo_Omon	23
thetao_Omon	55
thetaoga_Omon	10
thkcello_Omon	55
tos_Omon	82
tossq_Omon	6
umo_Omon	4
uo_Omon	26
vmo_Omon	4
vo_Omon	27
volo_Omon	2
wfo_Omon	5
wmo_Omon	7
wo_Omon	19
zfullo_Omon	2
zos_Omon	25
zossq_Omon	3

```
zostoga_Omon            2
siage_SImon             1
siarean_SImon           8
siareas_SImon           7
sicompstren_SImon       1
siconc_SImon           47
siconca_SImon           7
sidconcdyn_SImon        4
sidconcth_SImon         4
sidmassdyn_SImon        5
sidmassevapsubl_SImon   1
sidmasslat_SImon        1
sidmassth_SImon         4
sidmasstranx_SImon      4
sidmasstrany_SImon      5
siextentn_SImon        10
siextents_SImon        10
sifb_SImon              4
siflcondbot_SImon       4
siflcondtop_SImon       4
siflfwbot_SImon         4
siflswdtop_SImon        2
sihc_SImon              5
simass_SImon            8
simassacrossline_SImon  4
sisali_SImon            1
sisaltmass_SImon        4
sisnconc_SImon          5
sisnhc_SImon            4
sisnmass_SImon         11
sisnthick_SImon        22
sispeed_SImon           7
sistrxdtop_SImon       22
sistrxubot_SImon        4
sistrydtop_SImon       26
sistryubot_SImon        4
sitemptop_SImon        19
sithick_SImon          29
sitimefrac_SImon        8
siu_SImon              23
siv_SImon              24
sivol_SImon            28
sivoln_SImon            4
sivols_SImon            7
sndmassdyn_SImon        4
sndmasssnf_SImon        4
```

Appendix E – ESGF Variables Downloaded from JASMIN in Popularity Order

The following list shows the CMIP6.HighResMIP variables that have been downloaded from CEDA' s ESGF node via HTTP between 25th March 2019 and 14th May 2020. The variables have been sorted by the number of downloads. The number shows the number of times that unique datasets (e.g. a unique combination of institute, model, experiment, variant label, table and variable) have been downloaded from a unique IP address.

pr_Amon	716
tas_Amon	627
pr_day	545
psl_Amon	342
ua_Amon	317
ta_Amon	303
ts_Amon	294
uas_Amon	291
va_Amon	289
vas_Amon	281
zg_Amon	259
tas_day	235
hus_Amon	223
tasmax_day	218
ps_Amon	214
tasmin_day	203
ta_day	168
ua_day	163
hus_day	158
psl_day	155
uas_day	152
va_day	145
huss_Amon	144
vas_day	139
wap_Amon	136
evspsbl_Amon	133
mrro_Lmon	130
prw_Amon	121
rsds_Amon	119
rlds_Amon	119
tauu_Amon	111
hfls_Amon	109
hfss_Amon	109
tauv_Amon	109
rsdt_Amon	108
pr_3hr	107
tasmin_Amon	105
tasmax_Amon	104
rlus_Amon	98
huss_day	98
rsus_Amon	97
rsut_Amon	96
zg_day	96
tsl_Lmon	95
psl_6hrPlevPt	95
rlut_Amon	94
vas_6hrPlevPt	94
mrso_Lmon	92
hur_Amon	91
uas_6hrPlevPt	90
prc_Amon	88
hurs_Amon	85
va_6hrPlevPt	84
clt_Amon	82
tos_Omon	82
rsds_day	81
sfcWind_Amon	80
ua_6hrPlevPt	79

mrros_Lmon	78
tas_3hr	77
prsn_Amon	75
tas_6hrPlevPt	72
rldscs_Amon	70
rsuscs_Amon	70
rsdscs_Amon	69
sfcWind_day	68
snd_LImon	66
ta_6hrPlevPt	63
uas_3hr	63
cl_Amon	61
vas_3hr	60
rlutcs_Amon	59
rsutcs_Amon	58
thetao_Omon	55
rlds_day	55
thkcello_Omon	55
tos_Oday	54
rlus_day	54
snw_LImon	53
clivi_Amon	50
pr_E1hr	48
hurs_day	48
snm_LImon	48
clwvi_Amon	47
rsus_day	47
siconc_SImon	47
rh_Lmon	45
clw_Amon	44
rlut_day	44
cli_Amon	44
npp_Lmon	43
mrro_day	43
hur_day	42
psl_6hrPlev	40
hfls_day	40
ps_CFday	40
hfss_day	39
hus_CFmon	39
prc_E1hr	38
huss_3hr	38
wap_day	38
ta_CFmon	37
gpp_Lmon	37
lai_Lmon	37
prsn_day	37
psl_E3hr	36
mrsos_Lmon	35
prc_day	34
ra_Lmon	34
evspsblveg_Lmon	34
tsn_LImon	34
zg_6hrPlevPt	31
snc_LImon	31
hfds_Omon	30
ccb_Amon	30
clt_day	29
sbl_Amon	29
sithick_SImon	29
cct_Amon	28
mlotst_Omon	28
so_Omon	28
sivol_SImon	28
hur_CFmon	27
vo_Omon	27
tauuo_Omon	27
ci_Amon	26
siconc_SIday	26
uo_Omon	26
rld_CFmon	26
sistrydtop_SImon	26

hus_6hrPlevPt	25
zos_Omon	25
uqint_Emon	25
reffclwtop_AERmon	25
siv_SImon	24
rsucs_CFmon	24
cltisccp_CFmon	24
rlu_CFmon	23
tauvo_Omon	23
rsu_CFmon	23
pctisccp_CFmon	23
rldcs_CFmon	23
rlucs_CFmon	23
rsd_CFmon	23
siu_SImon	23
vqint_Emon	22
cltcalipso_CFmon	22
albisccp_CFmon	22
sisnthick_SImon	22
sistrxdtop_SImon	22
evspsblsoi_Lmon	21
sci_Amon	21
tslsi_day	21
rsdcs_CFmon	21
sos_Omon	20
sfcWindmax_day	20
sithick_SIday	20
parasolRefl_Emon	20
wo_Omon	19
ta700_CFday	19
sitemptop_SImon	19
tauu_Eday	18
pfull_CFday	18
rsut_CFday	18
rsutcs_CFday	17
mrsos_day	17
epfz_EmonZ	17
wtem_EmonZ	17
wbptemp_6hrPlevPt	16
phalf_CFday	16
tauv_Eday	16
vtem_EmonZ	16
rsdt_CFday	15
cl_CFday	15
ts_Eday	15
rlutcs_CFday	15
rsdscs_CFday	15
rlus_3hr	14
rsuscs_CFday	14
rsuscs_3hr	14
prw_E3hr	14
phalf_Amon	14
pfull_Amon	14
rlds_3hr	13
rsds_3hr	13
mrso_day	13
rsdscs_3hr	12
clw_CFday	12
hursmax_day	12
rsus_3hr	12
ua_E3hrPt	11
sisnmass_SImon	11
tran_Lmon	10
siextents_SImon	10
rldscs_CFday	10
siextentn_SImon	10
hursmin_day	10
siv_SIday	10
cltisccp_CFday	10
rivo_Eday	10
thetaoga_Omon	10
wap500_CFday	9

albisccp_CFday	9
hus_Emon	9
pctisccp_CFday	9
cli_CFday	9
cct_CFday	9
ccb_CFday	9
siu_SIday	9
snw_day	9
rsutcs_E3hr	9
sos_Oday	9
va_E3hrPt	9
va_Emon	9
ua_Emon	9
rtmt_Amon	8
siarean_SImon	8
sitimefrac_SImon	8
simass_SImon	8
ta_Emon	8
zg_Emon	8
rsntds_Omon	8
ta_E3hrPt	7
clt_3hr	7
hus_E3hrPt	7
sispeed_SImon	7
siconca_SImon	7
zg500_6hrPlevPt	7
tossq_Oday	7
sitemptop_SIday	7
cltcalipso_CFday	7
ptp_AERmon	7
prc_3hr	7
siareas_SImon	7
sivols_SImon	7
wmo_Omon	7
hfss_3hr	6
co2mass_Amon	6
rlutaf_AERmon	6
parasolRefl_Eday	6
snc_day	6
sisnthick_SIday	6
tossq_Omon	6
mc_CFday	5
rv850_6hrPlevPt	5
ua_CFday	5
sidmasstrany_SImon	5
prsn_3hr	5
zg_CFday	5
ta_CFday	5
omldamax_Oday	5
rlut4co2_CFmon	5
toz_AERmon	5
rsutaf_AERmon	5
rsutcsaf_AERmon	5
wfo_Omon	5
sihc_SImon	5
sidmassdyn_SImon	5
sisnconc_SImon	5
tos_3hr	5
mc_Amon	5
hfls_3hr	4
mrtws_Emon	4
hus_CFday	4
vmo_Omon	4
mlotstsq_Omon	4
ua_Eday	4
clwvi_CFday	4
clivi_CFday	4
clisccp_CFday	4
rsut4co2_CFmon	4
tslsi_3hr	4
vtem_EdayZ	4
epfz_EdayZ	4

wtem_EdayZ	4
sistrxubot_SImon	4
sistryubot_SImon	4
sisaltmass_SImon	4
sidconcth_SImon	4
siflcondtop_SImon	4
sisnhc_SImon	4
sifb_SImon	4
simassacrossline_SImon	4
sidconcdyn_SImon	4
sndmasssnf_SImon	4
sidmassth_SImon	4
siflcondbot_SImon	4
sndmassdyn_SImon	4
sidmasstranx_SImon	4
sivoln_SImon	4
siflfwbot_SImon	4
va_Eday	4
n2oglobal_Amon	4
ch4global_Amon	4
mrfso_Lmon	4
umo_Omon	4
od550oa_AERmon	3
pbo_Omon	3
sbl_LImon	3
evspsblpot_Emon	3
intvadse_Emon	3
wap_CFday	3
va_CFday	3
hur_CFday	3
clcalipso_CFday	3
clhcalipso_CFday	3
clmcalipso_CFday	3
nep_Emon	3
psl_CF3hr	3
zossq_Omon	3
lwsnl_LImon	3
ts_6hrPlevPt	3
ps_3hr	3
od550aer_AERmon	2
utendogw_EdayZ	2
bigthetao_Omon	2
tntscpbl_CFmon	2
volo_Omon	2
tnhus_CFmon	2
intuadse_Emon	2
cropFrac_Lmon	2
rsu4co2_CFmon	2
cllcalipso_CFday	2
siflswdtop_SImon	2
zg_EdayZ	2
sfcWindmax_Emon	2
wap_6hrPlev	2
baresoilFrac_Lmon	2
sfcWind_6hrPlevPt	2
huss_6hrPlevPt	2
ua_AERmon	2
rsdcs4co2_CFmon	2
mrsol_Emon	2
hfy_Omon	2
sltovgyre_Omon	2
zostoga_Omon	2
msftmz_Omon	2
prcsh_E3hr	2
zfullo_Omon	2
rldscs_3hr	1
mrsos_3hr	1
tnhusc_CFmon	1
rlucs4co2_CFmon	1
bigthetaoga_Omon	1
c4PftFrac_Lmon	1
hfbasinpmadv_Omon	1

```
clis_CFmon              1
obvfsq_Omon             1
masso_Omon              1
utendnogw_EdayZ         1
rlutcs_E3hr             1
rsut_E3hr               1
siconca_SIday           1
treeFrac_Lmon           1
grassFrac_Lmon          1
ta_Eday                 1
hfx_Omon                1
soga_Omon               1
rsdo_Omon               1
t20d_Emon               1
hcont300_Emon           1
tatp_AERmon             1
tnhusmp_CFmon           1
tnhusscpbl_CFmon        1
clisccp_CFmon           1
od550ss_AERmon          1
lwp_AERmon              1
ztp_AERmon              1
rlutcsaf_AERmon         1
intvaw_Emon             1
wtd_Emon                1
od550dust_AERmon        1
agesno_LImon            1
hfdsn_LImon             1
sidmasslat_SImon        1
abs550aer_AERmon        1
ttop_AERmon             1
cod_AERmon              1
od550so4_AERmon         1
intuaw_Emon             1
cltc_AERmon             1
tnhusa_CFmon            1
cllcalipso_CFmon        1
htovovrt_Omon           1
sltovovrt_Omon          1
rlut_E3hr               1
agessc_Omon             1
sicompstren_SImon       1
siage_SImon             1
sidmassevapsubl_SImon   1
sisali_SImon            1
va_AERmon               1
```

Appendix F – CMORization techniques

AWI

The procedure used by AWI to post-process model output data and submit it to JASMIN is:

```
1. each experiment
2.   each variable
3.     highest frequency table
4.       set time axis unit to days
5.       join (merge) files into 10-yearly
6.       adjust fesom timestamps
7.       rename variable and set description if required
8.       apply local attributes from data request
9.       native fesom 3d data is stored in an 1d netcdf vector, convert to
a 2d matrix with depth information
10.        apply fesom mesh description
11.
12.      each other table
13.        (monthly) mean or copy
14.
15.      CMIP6 file name (depends on time range)
16.      global attributes (activity_id, parent branch, uuid, url, ...)
17.      compress netcdf (Met Office requirement)
18.      regrid to lat lon grid to provide JASMIN snapshot
19.      upload to JASMIN storage at the Centre for Environmental Data
Analysis (CEDA)
```

Due to the many output variables, especially the ones with higher frequency and full 3D ocean data (e.g. so, thetao), AWI developed a novel mechanism to greatly improve output writing during the computation. With the developed technique, the data is written asynchronously and the simulation continues to compute while the file is being saved. This results in a 30% better overall performance of the coupled AWI model and enabled the AWI team to finish simulations more quickly and deliver the results in time.

The next big bottleneck in our pipeline from simulation to actual data upload (i.e. to JASMIN) was the post-processing that had to be done to get CMOR compliant data (shown in the listing above). Although the postprocessing was done in parallel, there were two very time-consuming steps, which took several weeks to finish:

a) adjust timesteps (line 6 in the listing)
b) FESOM 3D data conversion (line 9)
c) netCDF compression (line 17)
d) upload to JASMIN (line 19)

The upload to JASMIN d) and probably also not the compression c) could be avoided and so the remaining two bottlenecks were improved:

In the initial approach, AWI had been using the cdo utility to adjust all timestamps to be CMOR compliant. As the cdo command seems to always rewrite the whole bulk data as well (i.e. the

variable data), this procedure is depending on the disk I/O performance and also of the size of the files.

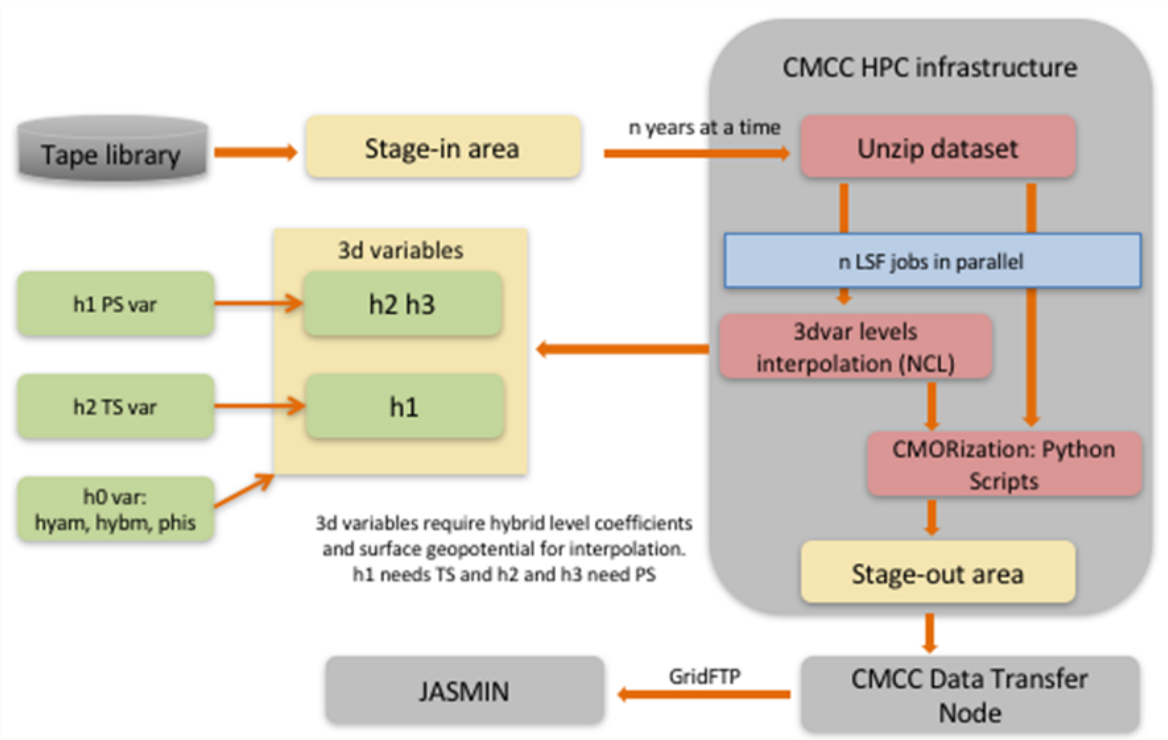
The AWI team developed a dedicated C++ program which injects the CMOR timestamps without having to rewrite any other part of the netCDF file. This procedure can now be applied to each file almost instantaneously and removes the bottleneck from a).

FESOM stores the output for 3D variables in a highly storage efficient manner to save disk space. To be understandable by general ocean-mesh reading tools like cdo, the whole 3D data has to be converted and thus rewritten. This is the reason for bottleneck b). To avoid this, AWI developed an optional mechanism for FESOM which allows to directly write the generalised 3D files when storing the data. This produces about 30% bigger files during the computation. Due to the highly effective asynchronous output writing (see above), this does not lead to a reduced overall computation time for the coupled model. For the CMOR post-processing, AWI could spare the time for reading all 3D data, converting it and storing the general 3D data. For the HR simulations, 10 years of daily 3D data use about 900 GB of disc space (uncompressed). netCDF compression brings this down to about 50% for deflation level 6 (data shuffling enabled). As for the netCDF compression c), AWI is still exploring means to speed up the procedure as well. One could for example directly compress the variable data when writing the files from FESOM. As with the 3D data restructuring, this will take some time during the computation but save the time for reading and writing the file in post-processing. Whether this will be faster depends on how well the HPC hardware handles the asynchronous writing and also how fast the disc I/O is during postprocessing.

CMCC

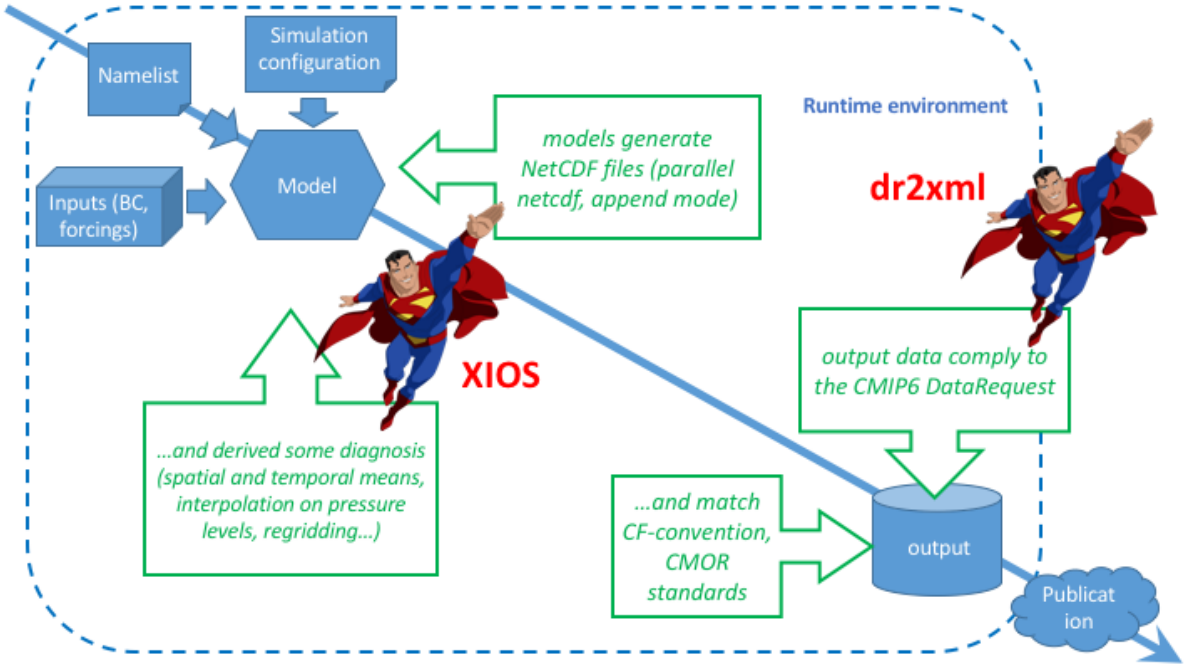
CMCC used the system shown below to CMORize the output from their model. The system runs on their HPC facilities and is designed to reduce the volume of disk space required. The interpolation phase, used to extract the atmosphere 3D variables over predefined CMIP6 pressure levels, requires inputs from various time resolution datasets. For each year, one year at a time, datasets are unzipped on the HPC infrastructure working directory: at the end, all the unzipped files have to be removed to free working disk space. An enhanced version of the implemented scripts allows processing several years at once, according to the resources available on the HPC cluster.

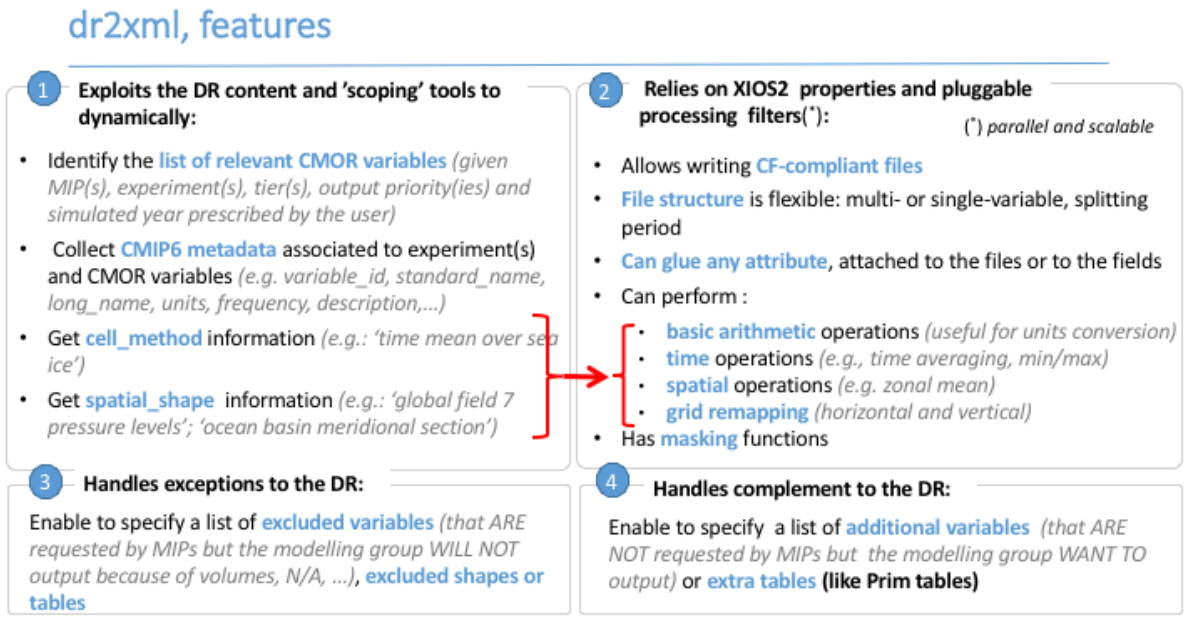
For each model component (atm, ocn, lnd, ice), the corresponding variables are organized into groups and many serial (not MPI) jobs are used to process and CMORize (in parallel) each variable. 2D variables only need to be CMORized (or else, a variable may be calculated as the sum of two raw variables), while 3D variables need to be interpolated before the CMORization, by taking as input other variables available at different time resolutions. CMORized outputs are spot checked by using the PrePARE tool and then moved on to a stage out area mounted via NFS on a data transfer node enabled for remote transferring to JASMIN. Considering several experiments (highresSST-present, control-1950, hist-1950, highres-future, highresSST-future) about the CMCC-CM2 model run at both HR4 and VHR4 resolutions, 400 TiB of uncompressed input data produced around 75 TiB of output data.



CERFACS

CNRM-CERFACS do not use CMOR and instead the model running on their HPC directly produces CMIP6-compliant netCDF files from the model's XIOS I/O server. XIOS is configured directly from the CMIP6 data request using the Dr2xml software that was developed for CMIP6.





EC-Earth

The EC-Earth3 model produces output data in a native format, which is not compatible with CMIP6 data requirements. The individual components of EC-Earth (atmosphere, ocean, etc.) use, moreover, different native formats for the output. Thus, the output needs to be processed on a per-component base and be converted into CMOR compatible data.

The structure of the EC-Earth consortium implies also a geographically distributed use of computational resources and workflow. The EC-Earth member institutions contributing to

PRIMAVERA performed their own sets of experiments and post-processed the output individually. This added another layer of coordination for the data processing.

The tools and workflow for production and dissemination of CMIP6 data was still under development at the time when PRIMAVERA output was processed and uploaded to JASMIN. Thus, early versions of the CMORization tool chain were used and part of the workflow was not yet in place. In fact, the development of the main post-processing tool for EC-Earth data (`ece2cmor3`, see below) benefited from the PRIMAVERA exercise as a first production test case.

The output configuration for EC-Earth, i.e. which variables are produced at which frequency in time and space, is handled by the `genecec` (GENerate EC-Earth Control output files) tool, which produces output control files (e.g. FORTRAN namelists) for a given data request. However, this tool was not yet available for the PRIMAVERA experiments, and the control files have therefore been created manually.

The actual conversion from native model output to CMOR compliant data is managed by `ece2cmor3`, another tool developed for the EC-Earth CMIP6 workflow. At the time of PRIMAVERA data processing, `ece2cmor3` was available in an early version. This version relied on a manual implementation of the data request in `ece2cmor3`. In later `ece2cmor3` releases, this has become more dynamic in the sense that each CMIP experiment is CMORized based on its own data request file (either the raw Excel data request files or the JSON data request files as provided by `genecec`, in which the EC-Earth3 component preferences and the EC-Earth3 availability are taken into account). More recent versions of `ece2cmor3` thus allow a more direct and comfortable control over the CMORized output by table/variable lists in JSON format for each model domain.

The `ece2cmor3` tool was run for each model component for the PRIMAVERA/HighResMIP data request, processing the output of one simulated year at a time. The runs for each model year have been parallelised within the resource limits for each institution.

At BSC, the CMORization was orchestrated with the Autosubmit work flow manager, which handled the whole experiment work flow. It allowed to CMORize the data chunk by chunk, as the simulation runs, directly on the HPC, before transfer to the storage facility and before sending the CMORized data to JASMIN.

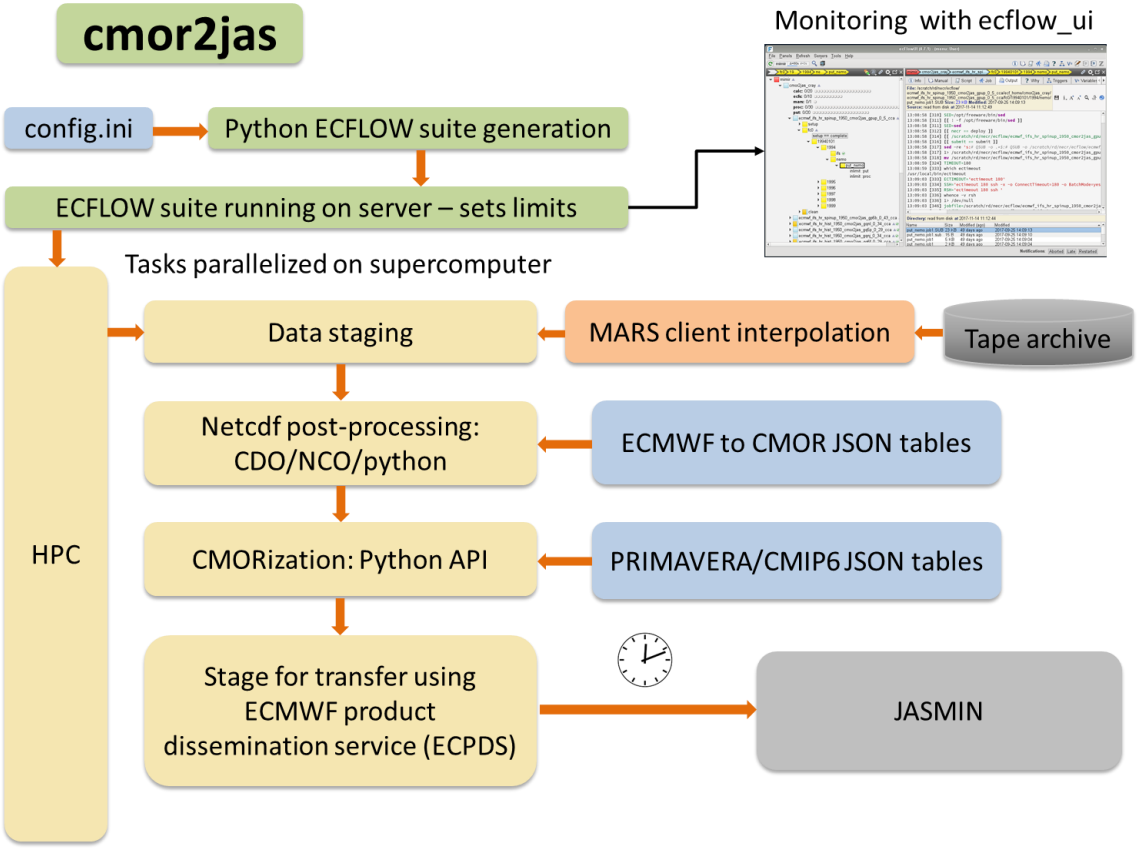
Quality control was not yet well developed for the PRIMAVERA data processing. A basic completeness test was performed to make sure all variables were produced for all years. For later processing of EC-Earth data for CMIP6 the workflow includes comprehensive data checking with QA-DKRZ.

The uploading of CMORized data to the JASMIN system required two stages: The data had to be packaged into chunks of less than ca. 0.5 TiB due to a limitation of the software that validated the data on the JASMIN side. Once a data chunk was prepared, the actual transfer was initiated with rsync. The group workspace on JASMIN allowed for 5-10 chunks of data to be processed in a pipeline-like manner.

EC-Earth3 produced about 1.4 TiB output data per simulated year in the high-resolution configuration for the PRIMAVERA experiments (ca. 0.4 TiB/year for standard resolution). The transfer speed to JASMIN varied between the participating EC-Earth groups, resulting in rates between 1 and 2 simulated years per day.

ECMWF

An automated suite called “cmor2jas” was developed for post-processing CMORization and transfer to JASMIN. An ECFLOW front-end drives shell scripts and a Python backend, which is all based on tools used for the ECMWF operational suites. The data volumes to be processed required the CMORization to be parallelized on the HPC. Use of the HPC was necessary to get the required throughput. The ECMWF’s linux cluster did not have the required scratch space. The throughput on the HPC was limited by the number of parallel retrievals (n < 20) and the temporary disk space available (the scratch quota is 50 TiB). The absolute bottleneck was the transfer of the data to JASMIN, with a rate of 1 to 2 TiB/day being possible due to congestion on ECMWF’s Internet connection. The data was pushed to JASMIN using the ECMWF Product Dissemination System. To overcome the data transfer bottleneck the data on the atmospheric grid was provided at a reduced resolution of 0.5° × 0.5° grid for the 25 km high-resolution model and on a 1.0° × 1.0° grid for the lower resolution 50 km model. The cmor2jas system is summarised below.



MPI

MPI developed a custom workflow that involved the use of NCO and CDO tools.

MOHC

Data from the MOHC and NERC HadGEM3-GC31 models is produced in a proprietary PP format and requires post-processing before submission. A suite of software has been generated to extract data from the Met Office tape archive, CMORize that data, perform quality control checks and submit the data to CEDA's ESGF publication system. This suite was still under development when the PRIMAVERA simulations began. A beta release of the component of the suite that performs the CMORization, named mip_convert, was taken and adapted for PRIMAVERA. mip_convert uses CMOR for writing the output netCDF files and PRIMAVERA's testing of it was useful allowing several bug fixes to be contributed to it. A Rose suite was wrapped around mip_convert and software written to configure the Rose suite directly from the PRIMAVERA data request.

The post-processing was run on the science batch processing cluster at the Met Office, which is similar to the LOTUS cluster at JASMIN. The PP files were retrieved from tape and then CMORized in parallel. Before the netCDF files are written back to the tape archive, the PRIMAVERA validation code (Seddon, 2020c; Seddon and Stephens, 2020) is run on the files and the output is saved in a JSON file. The validations from the JSON file are copied to JASMIN and loaded into the DMT. JASMIN has a direct connection to the Met Office tape archive, allowing PRIMAVERA users to request the MOHC and NERC data directly from the archive. The processing suite and mip_convert code are held in the Met Office Science Repository Service under a proprietary license.

This approach typically required less than 5 TiB of scratch storage per model. The throughput of the processing was limited by the rate at which data could be restored from tape and the time taken to read the PP format files. Future upgrades to the system will look at running the CMORization on the HPC before the PP files are written to tape.