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

Milestone MS25

Finalise and agree updated Data Management Plan for Stream 2 simulations

Milestone Title	Finalise and agree updated Data Management Plan for Stream 2 simulations	
Brief Description	Finalise and agree updated Data Management Plan for Stream 2 simulations.	
WP number	9	
Lead Beneficiary	Susanna Corti, CNR	
Contributors	Jon Seddon, Met Office	
Creation Date	14th November 2018	
Version Number	1.0	
Version Date	20th December 2018	
Milestone Due Date	October 2018	
Actual Delivery Date	December 2018	
Nature of the Milestone	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**
0.1	30/11/2018	Jon Seddon	First draft
1.0	20/11/2018	Jon Seddon	Agreed by project

Contents

List of Tables

List of Figures

1. Executive Summary

This report forms the Data Management Plan (DMP) for the PRIMVERA Stream 2 simulations. The DMP describes which simulations will be shared across the project and which can be analysed at the computing facility where they were run. For simulations that need to be shared across the project it is not feasible to upload as much data as was provided in Stream 1. The variables to be shared are described in this DMP along with an estimate of the volume of this data and the time taken to upload it to the central analysis facility at JASMIN. The DMP reinforces the metadata standards from Stream 1 that all PRIMAVERA data should follow.

2. Project Objectives

With this Milestone, the project has contributed to the achievement of the following objectives (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)*		No
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)*	Yes	
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

3. Detailed Report

3.1. Aims

The Data Management Plan for the PRIMAVERA Stream 2 simulations aims to ensure that:

- the Stream 2 data that is required for analysis by multiple PRIMAVERA project members is collected together at the central analysis facility at JASMIN.
- the data is made available to users as quickly as possible after the simulations have finished by only uploading the data that is absolutely needed for analysis to JASMIN.
- the data is safely archived so that it is available to the project members and the wider community after the end of the project.

3.2. Workflow

The workflow used in the Stream 1 simulations is shown in Figure 1. The same workflow will be used in Stream 2. Data is generated on the HPCs at the seven modelling centres. The modelling centres communicate with WP9 and when there is sufficient disk capacity at JASMIN, the model output data is transferred to a PRIMAVERA group workspace at JASMIN. When the upload is complete the uploaded files are validated and the files' metadata is stored in the Data Management Tool's (DMT) database. The files are then moved to tape to create disk space to allow more data to be uploaded.

Users can query the data that has been uploaded using the DMT's web interface. They can use the DMT to request the retrieval of data that they need from tape to disk. When the analysis of the data is complete, users can mark it as being complete, allowing it to be deleted, thus creating space for other users' data retrieval requests.

When all years for a variable have been uploaded then the data will be ingested to the CEDA archives and published to the wider community through the Earth System Grid Federation (ESGF).

This workflow has worked well during Stream 1. Much development and optimisation has occurred in the processes and the software tools that enable the workflow to make it as reliable and as efficient as possible. The modelling centres and the users of the data are familiar with the processes and tools.

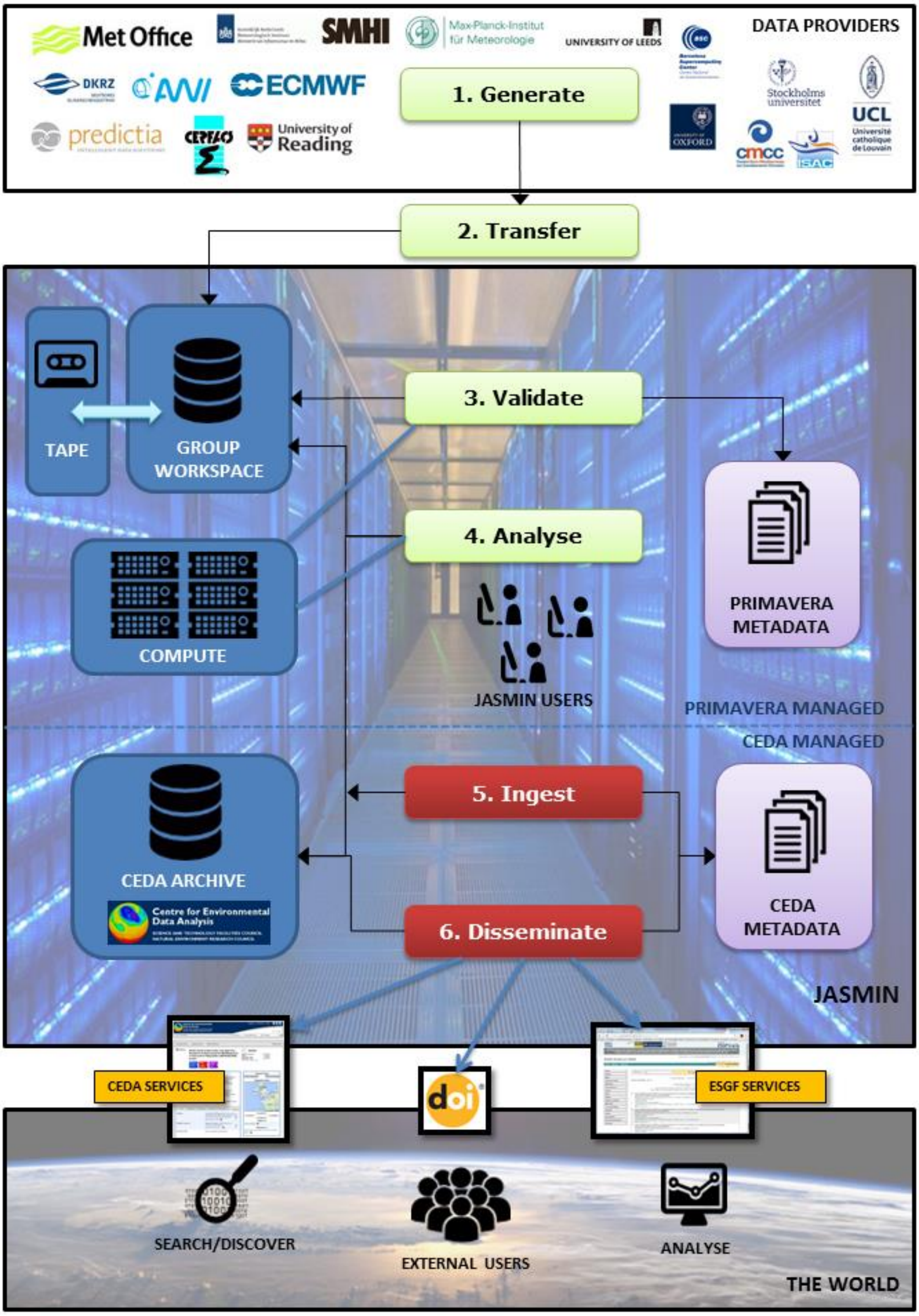

Figure 1 - The data workflow for the PRIMAVERA Stream 1 and Stream 2 simulations.

3.3. Simulations to be run

The simulations that each centre has agreed to run are given in the Stream 2 summary document at:
https://docs.google.com/document/d/11WTQ2lqqTCXfCqcuk2Qh8oAJ29mp83AZcSXCq1nacAk/edit

There will be two types of simulation:

- simulations that test new physics or a specific hypothesis. These simulations will typically be analysed by the group that ran them. These simulations do not need to be uploaded to JASMIN.
- simulations that complement or extend the Stream 1 simulations. These will be analysed by many groups and will be uploaded to JASMIN. These simulations are shown in green in the Stream 2 summary document.

3.4. Data volumes

Some modelling centres such as ECMWF, AWI and MPI either regrid their output onto a reduced resolution regular latitude-longitude grid or output a smaller number of output variables compared with other centres. These three centres' Stream 2 simulations are estimated to produce an output data volume that can be transferred to JASMIN in under two months.

If the remaining modelling centres output the same set of variables as they did in the Stream 1 simulations then the volume of data generated could take over one year to upload to JASMIN, preventing the simulations from being analysed before the end of the project.

At the Stream 2 planning meeting it was agreed that the reduced data request given in Section 3.5 would be produced. An estimate of the volume of data produced by each modelling centre is given in https://docs.google.com/spreadsheets/d/1_nxG5F2jtmZF-WL2XMcmyzJwjg3NBknDqS1BSIYZl-Y/edit#gid=1845230278. The estimates of the reduced data volume per experiment were made in https://github.com/PRIMAVERA-H2020/stream2-planning/blob/master/stream1_high_freq.ipynb. Five of the modelling centres will be able to upload this data in under two months. The remaining centres will be able to upload the data in under four months.

3.5. Agreed Data Request

At the Stream 2 planning meeting it was realised that it was not feasible to upload the full set of Stream 1 variables for all of the planned ensemble members. At the meeting it was agreed that the following variables will be uploaded to JASMIN from each simulation:

- the same monthly variables as in Stream 1.
- the same daily variables as in Stream 1, with the exception of variables in MIP table CFday. From CFday only ps will be uploaded.
- the 6 or 3 hourly variables listed in Table 1 and Table 2.

The full monthly and daily output, with the exception of CFday, can be uploaded in under one week for each ensemble member. Most of the variables in the CFday table are on atmosphere levels, which for the PRIMAVERA models is typically 85 levels, and are the typically 20% of the volume of Stream 1 data. The CFday variables are designed to allow the study of cloud feedbacks, which isn't a priority in PRIMAVERA.

Modelling centres can chose which variables to output from their model as long as only the specified variables are uploaded to JASMIN. For example, for the Met Office simulations all of the Stream 1 variables will be output by the models and stored in the Met Office tape archive. However, only the requested Stream 2 variables will be uploaded to JASMIN.

The 6 or 3 hourly variables requested by WP10/11 are:

Request Description	**CMOR name**	**3hr Table**	**6hr Table**
Surface short-wave (solar) radiation downwelling, diffuse; time-average	rsdsdiff	3hr	Prim6hr
Surface short-wave (solar) radiation downwelling, total; time-average	rsds	3hr	Prim6hr
Surface air temperature (T2M); instantaneous values	tas	3hr	6hrPlevPt
Wind vector (i.e., both U and V); 10m above ground; instantaneous values	uas, vas	3hr	6hrPlevPt
Wind vector (i.e., both U and V); 50m above ground; instantaneous values	ua50m, va50m	Prim3hrPt	Prim6hrPt
Wind vector (i.e., both U and V); 100m above ground; instantaneous values	ua100m, va100m	Prim3hrPt	Prim6hrPt
Wind vector (i.e., both U and V); 850 hPa; instantaneous values	ua850, va850	E3hrPt	
Wind speed; 10m above ground; time-average	sfcWind	Prim3hr	
Wind speed; 10m above ground; maximum value	sfcWindmax	Prim3hr	Prim6hr
Wind gust; 10m above ground; maximum value	wsgmax		Prim6hr
Total precipitation; time-average or accumulation	pr	3hr	Prim6hr
MSLP; time-average	psl	E3hr	6hrPlev

Table 1 - The WP10/11 high-frequency requested data.

Some groups were not able to output ua850 and va850 from Stream 1. In this case then ua or ua7h and va or va7h from E3hrPt or 6hrPlevPt are a suitable equivalent.

There is an additional request from other work packages for a single high-frequency variable to allow tropical cyclones to be tracked:

Request Description	**CMOR name**	**3hr Table**	**6hr Table**
Geopotential height; 500 and 250 hPa; instantaneous values	zg7h	Prim3htPt	6hrPlevPt

Table 2 - The additional high-frequency requested data.

The frequency that Stream 1 variables were requested for restoration from tape to disk was checked to ensure that no essential variables had been omitted from this reduced data request:
https://github.com/PRIMAVERA-H2020/stream2-planning/blob/master/vars_retrieved.ipynb .

The modelling groups that were able to produce the requested high-frequency variables in their Stream 1 output are shown in: https://docs.google.com/spreadsheets/d/1_nxG5F2jtmZF-WL2XMcmyzJwjg3NBknDqS1BSIYZl-Y/edit#gid=545191476.

To reduce Cerfacs' data volumes they will not produce the variables in the tables listed below. These variables are 3D fields, the majority of which have not yet been accessed from the Stream 1 data.

AERmon: ua va
Amon: cl cli clw mc pfull phalf
CFmon: clc clic clis cls clwc clws hur hus mcd mcu rld4co2 rld rldcs4co2 rldcs rlu4co2 rlu rlucs4co2 rlucs rsd4co2 rsd rsdcs4co2 rsdcs rsu4co2 rsu rsucs4co2 rsucs ta tnhus tnhusa tnhusc tnhusmp tnhusscpbl tnt tnta tntc tntmp tntr tntscpbl
Emon: cldicemxrat27 cldwatmxrat27 hus27 t2 ta27 tntmp27 tntrl27 tntrs27 twap u2 ua27 ut uv uwap v2 va27 vt vwap wap2 wap zg27

3.6. Data and metadata standards

As was the case for Stream 1, all data should be netCDF files suitable for submission to the CMIP6 project. Specifically:

- files should comply with "CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CV's", https://goo.gl/v1drZl.
- files should be suitable for submission to the ESGF and so should have a data_specs_version attribute value of 01.00.21 or more recent.
- all files should use netCDF deflation to reduce their size to save storage space and maximise the transfer rates achievable.

3.7. Timescales

The first priority for WP9 is to complete the upload and validation of the Stream 1 simulations. There are small amounts of outstanding coupled simulations to be uploaded from Stream 1 historic and all of the future simulations. The future forcings became available in November 2018 and so the future simulations will soon be ready for upload to JASMIN.

Priority will be given to the Stream 1 simulations but the Stream 2 simulations will be uploaded whenever there is capacity. Slotting the Stream 2 simulations around the outstanding Stream 1 simulations will maximise the use PRIMAVERA's resources at JASMIN.

To optimise the use of resources, users should email the WP9 lead before beginning any uploads. WP9 will advise on where the data should be uploaded to.

It is still estimated that it will take some groups up to four months to upload their data. The initial data analysis will be done on the daily and monthly data. It is recommended that centres upload the daily and monthly data for all ensemble members first, allowing analysis to begin using the most frequently accessed variables [1]. The sub-daily data can then be uploaded.

The estimates of the total data volume are shown in:

https://docs.google.com/spreadsheets/d/1_nxG5F2jtmZF-WL2XMcmyzJwjg3NBknDqS1BSIYZl-Y/edit#gid=1845230278

It is estimated that the Stream 2 data that needs to be uploaded to JASMIN is around 400 terabytes. The longest upload time for a single modelling centre is 109 days. This assumes that uploads run uninterrupted for this time. A rough estimate would be that this centre's uploads would be complete within six months. There is sufficient disk space at JASMIN for three modelling centres to upload simultaneously. Therefore all of the centre's uploads should complete within six months assuming that there aren't any technical problems at the centres. However, by first uploading the monthly and daily, which is used in the majority of analyses, then results should be available from the Stream 2 data before the end of this six month period.

4. Lessons Learnt

This data management plan is a continuation of the data management plan that was developed for and optimised during Stream 1. The data access patterns observed with the Stream 1 simulations have been used to reduce the data request for Stream 2. This will ensure that the planned data can be uploaded to JASMIN and made available for analysis by the whole PRIMAVERA project.

5. Links Built

This DMP is based on the decisions made at a meeting held at Schiphol airport on 12th November 2018. This meeting was attended by the work package leads and representatives from the WP6 institutions.

The DMP will directly allow deliverables D6.5, D6.6, D6.7 and D9.5 to be achieved. The data uploaded will also allow deliverables from other work packages to be completed.

[1] https://github.com/PRIMAVERA-H2020/stream2-planning/blob/master/vars_retrieved.ipynb