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 D*9.1*

Data Management Plan

Deliverable Title	*D9.1 Data Management Plan*	
Brief Description	*Specification of the Inputs and Outputs, data format standards, means of exploitation and data curation policy. This specification will be based on CMIP5 standard diagnostic lists and will include details of file-naming and metadata conventions.*	
WP number	9	
Lead Beneficiary	*Matt Mizielinski Met Office*	
Contributors	*Matt Mizielinski (Met Office)* *Ag Stephens (STFC CEDA)* *Paul van der Linden (Met Office)* *Pierre Antoine Bretonniere (BSC)* *Sandro Fiore (CMCC)* *Jost von Hardenberg (CNR)* *Michael Kolax (SMHI)* *Katja Lohmann (MPG)* *Marie-Pierre Moine (CERFACS)* *Philippe Le Sager (KNMI)* *Tido Semmler (AWI)* *Retish Senan (ECMWF)*	
Creation Date Version Number Version Date	1-4-2016 1 14-6-2016	
Deliverable Due Date Actual Delivery Date	1-5-2016 23-6-2016	
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**
0.1	1st April 2016	Ag Stephens, Matthew Mizielinski	First draft
0.2	5th May 2016	AS & MM	Prototype
1.0	14th June 2016	MM	Final draft

Table of Contents

1. Summary

This plan describes the management of data in the PRIMAVERA project. It specifies the required data inputs and outputs, the data format standards, the means of exploitation and the data curation policy. The specification of the required inputs and outputs is a pre-requisite for all Stream 1 simulations in WP6. This data specification is based on CMIP5 standard diagnostic lists, but input was gathered from all modelling groups for completeness. The specification includes details of file-naming and metadata conventions building on the Data Reference Syntax (DRS) developed in CMIP5 and used for MIP-style projects in the Earth System Grid Federation (ESGF). This plan will be updated during the life of the project as necessary.

2. Project Objectives

With this deliverable, 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)*	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

3. Detailed Report

Introduction and scope

The purpose of this data management plan (DMP) is to set up a coherent approach to data issues pertaining to the PRIMAVERA project. The data management objectives are to ensure that:

- A high quality documented data archive is created.
- Appropriate data support is provided to the data users and creators.
- Data are made available to users in a timely fashion.
- Academic credit for data creation is given.
- Conditions of use, access and deposit are clearly stated and do not infringe on the data creators' rights.
- Potentially scientifically valuable data are kept for reuse in the long-term and by other disciplines.
- Results can be checked and validated.

This document is an agreed record of the data management needs and issues within the project. It defines who is responsible for data management activities both within data centres and by the data creators. It lists the expected data products and provides a mechanism for recording and agreeing changes. Other data needs and issues are also laid out so that problems can be identified early. It includes conditions of use and deposit to clearly express the ownership, responsibilities and rights associated with the data.

The scope of this document is data management of the data used or generated by the PRIMAVERA project.

This data management plan has been agreed between the Data Centre (STFC Centre for Environmental Data Analysis (CEDA)) and the project PIs (Pier Luigi Vidale and Malcolm Roberts). In the event of dispute relating to the data the final decision rests with the PIs and the Executive Management Board.

About the PRIMAVERA dataset

The dataset encompasses all data produced by the project which is to be archived for posterity within the CEDA archive. This may include validated raw data, processed data and model results, where they have long term importance and/or use to the wider scientific community. The individual data sources are summarised in Appendix A.

Roles and responsibilities

Data archival

The roles and responsibilities for PRIMAVERA data management are as follows:

1. It is the responsibility of the grant holder to ensure that all appropriate data and supporting metadata has been submitted to the archive before the end of the grant period.
2. Data archival:

1. When relevant, preliminary data should be made available to project collaborators as soon as possible. Any corrections or amendments to the preliminary data should be announced as soon as possible.
2. All validated processed data (i.e. data sets in their final form) will be archived at the data centre. Archival must take place no later than the end of the project funding period (31st October 2019).

3. Ownership of the data lies with the data creator.
4. Data creators are required to agree to the data deposit conditions before the data are added to the archive.

Deposit Conditions

1. The depositor confirms that he/she is the owner of the data and/or has the right to deposit the data in the CEDA archive.
2. Ownership of the data remains with the data creator.
3. CEDA reserves the right to store the data, and make the data available under the Conditions of Use.
4. The depositor grants CEDA permission to, without changing content, translate the data to any medium or format for the purpose of future preservation and accessibility

Data distribution

Access to all data submitted to the data centre will be restricted to project participants whilst it is held in the PRIMAVERA Group Workspaces on JASMIN. Once data has been ingested into the CEDA archive it will then be will be released into the public domain. All project partners are encouraged to make their data available for archival as soon as possible. Potential users of the archive will be required to agree to the Conditions of Use.

Conditions of Use

1. Access to all data submitted to the archive will be restricted to project participants whilst the data is held in the JASMIN Group Workspace, after which they will be released into the public domain under the CMIP licence.
2. Whilst the data are restricted from the public domain, no data should be transferred to a third party without the originator's consent.
3. If measurements or model results from other groups within the project are used in a project participant's publication during or after the project, joint authorship must be offered.
4. In all cases where the data are used in a presentation or publication, a citation must be given. Users should refer to the appropriate dataset page in the CEDA Data Catalogue (http://catalogue.ceda.ac.uk) for the specific citation to use. For example, "British Antarctic Survey (2008): British Antarctic Survey: high resolution radiosonde data from Halley and Rothera stations. NCAS British Atmospheric Data Centre, *date of citation.* http://catalogue.ceda.ac.uk/uuid/37f2bef57e28bcd780a5cbfe077f4bf8 " Further details on data citation are available at: http://www.ceda.ac.uk/help/users-

guide/citing-data/http://www.ceda.ac.uk/help/users-guide/citing-data/

5. These Conditions of Use should be read alongside the project Consortium Agreement (sections 4.1.5 and 8.2.1) and Article 29 of the Grant Agreement which further detail partner responsibilities surrounding project data.
6. Information submitted in application for access to the data will be made available to the Science and Technology Facilities Council (STFC) Centre for Environmental Data Analysis (CEDA) for the purposes of tracking data usage and of improving the service.

Publication

Results coming out of the project will be published in the usual way. It is each investigator's responsibility to ensure that the data used in publications are the best available at that time. During the embargo period (before data has been published) each investigator will have the right to refuse the use of his/her results in a publication or a presentation prior to the investigator's own publication of that work. If measurements or model results from other groups within the project are used in a project participant's publication during or after the project, joint authorship must be offered. This will not necessarily have to be accepted, particularly in cases where due credit and acknowledgement can be given in other, possibly more appropriate, ways. References of publications should be communicated to the data centre so that they can be integrated into the archive documentation.

Access to third party data

Third-party data required for the development of the project activities and held by CEDA, such as Envisat, ECMWF and Met Office data sets, will be made available to the participants, subject to current access conditions.

If required and depending on available resources, the data centre will endeavour to retrieve data sets from other sources at no cost or will negotiate their acquisition at the best possible cost.

A summary of third party data required by the project is given in Appendix B.

Communication between the data centre and project participants

A dataset catalogue record will be set up at the data centre. This will be the gateway to all project data and metadata, and to all relevant information and links. The web page will be the primary source of information regarding the data archive and will be updated as new information is available.

Support to scientists

The data centre will provide assistance with format and metadata issues, as well as a helpdesk to handle queries from data users and providers.

Archive Location, Security and Backup

The archive will be located at the STFC Rutherford Appleton Laboratory (UK) where CEDA is based. Data will be stored online and accessible to users as per the conditions of use stated above. Data will be curated and backed up according to current data centre practices.

Data Generation Activities

The individual data sources are summarised in Appendix A.

In-Project Data Management Approach

Until the data are supplied to the data centre, the project team are responsible for storing the collected data securely. General advice includes:

- Keep updated anti-virus protection on every computer.
- Use appropriate descriptive file names and directories to avoid over-writing or mixing up results.
- Record all relevant metadata at the same time as the data - it can be hard to remember specifics after the event.
- Record the original creation date and time for files on your systems.
- Version control - keep track of authorship and changes made to data files.
- Lab notebooks should be stored in a safe place.
- Data files should be backed up regularly and the backup data stored in a secure place physically removed from the original data.
- Samples should be appropriately saved so they will not degrade over time.

Metadata and Documentation

Information on the data

Metadata (i.e. information on the data) are a crucial part of any data archive since they ensure the discoverability, accessibility and readability of the data. It is therefore essential that metadata be submitted at the same time as the data sets to which they pertain. Metadata describing any project data not archived at the data centre should also be supplied to the data centre. Metadata for NetCDF files should comply with the Climate Forecasting (CF) convention as far as possible. Metadata requirements are documented here: http://badc.nerc.ac.uk/help/metadata/.

In addition to these metadata, investigators are encouraged to archive all relevant information, including model and experiment descriptions, references, papers, reports, etc. These will be held at CEDA or provided as links to external repositories.

Data Quality

Data should normally be submitted to the data centre ready to archive and the quality level indicated within the data or accompanying metadata.

In order to facilitate future data use, and unless this is inappropriate for particular cases (such as the widespread use of a specific format by a specialised community), NetCDF will be adopted for all the processed data generated by the project. Documentation on this format is available here: http://www.badc.rl.ac.uk/help/formats/.

Additional Services

Alongside the main data archival procedures the following services are provided to PRIMAVERA:

1. **PRIMAVERA Group Workspace**

The PRIMAVERA Group Workspaces are large disk partitions made available to the project on the JASMIN platform. At the time of writing 480 TB is available to PRIMAVERA distributed across 5 separate workspaces.

The purpose of this disk is to be used to:

- Bring preliminary data to JASMIN
- Share data with PRIMAVERA partners
- Prepare data for ingestion into the CEDA Archive

Additional disk may be made requested during the course of the project should it be required.

2. Tape access (for overflow from disk)

In order to manage the large volumes of data that will be produced by PRIMAVERA we will need to migrate much of it onto tape on JASMIN. A tape service, known as “Elastic Tape” will be used to manage this process.

3. Access to Data Transfer servers and services

JASMIN has a set of data transfer nodes that are connected to the UK academic network JANET with access to a 10 Gb/s network connection (shared with other STFC services). Previous tests have shown that data transfer rates of 6-8 TB per day are achievable with data transfer tools such as the Globus toolkit (gridFTP) or bbcp.

4. Access to Data Analysis servers

There are a set of communal servers that are available for interactive scientific data analysis on JASMIN. The specification of these servers is documented at http://help.ceda.ac.uk/category/108-interactive-computing . It is recommended that computationally demanding tasks are performed using the LOTUS cluster.

5. Access to the LOTUS parallel processing cluster

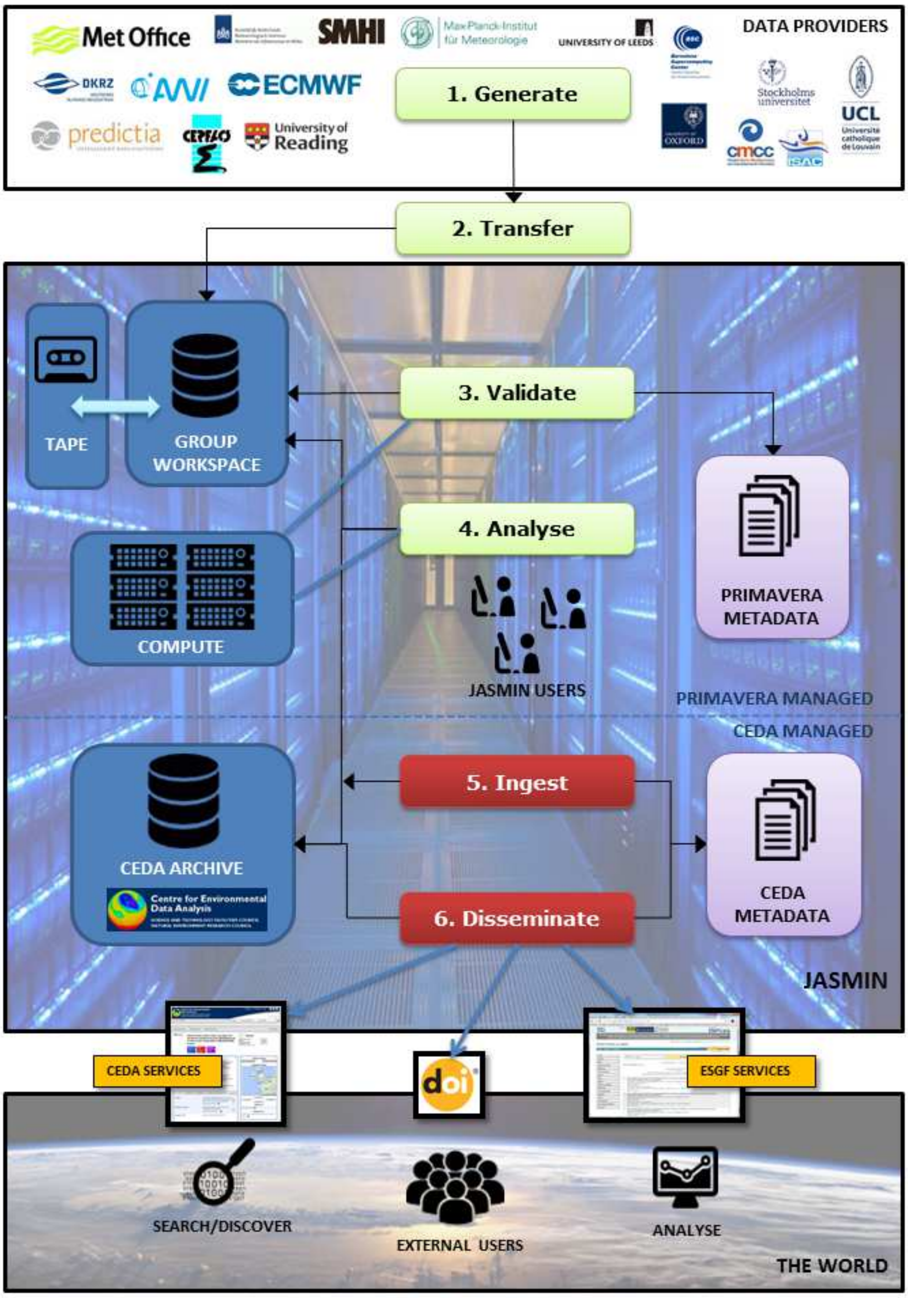
The LOTUS cluster has around 4,000 cores in a cluster attached to JASMIN. It is accessible from all science data analysis systems and users are encouraged to submit sets of batch jobs rather than run long serial processes on the analysis servers. LOTUS is described in detail at http://help.ceda.ac.uk/article/110-lotus-overview

6. PRIMAVERA wiki, project management tools and subversion repository

The PRIMAVERA wiki is a Trac wiki with attached SVN source code repository. Project milestones and tasks can be managed through the ticketing system, and work package leads are encouraged to use this functionality. Information for public dissemination will be made accessible via the primavera web site https://www.primavera-h2020.eu

The PRIMAVERA Data Workflow

The following diagram and table provide an overview of how the data will travel through PRIMAVERA from initial creation (by the Data Providers) to final archival and publication (by CEDA).



This table provides information about each step shown in the diagram above. It explains what each step consists of, provides links to more information and specifies who is responsible.

#	Step	Details	Useful links	Owner
1	**Generate**	Data Provider (DP) generates model output.		DP
2	**Transfer**	DP transfers output to the PRIMAVERA Group Workspace (GWS) using the JASMIN transfer servers. Data is written to "incoming" directory.	http://jasmin.ac.uk/how-to-use-jasmin/group-workspaces/working-	DP

			with-data/ http://jasmin.ac.uk/how-to-use-jasmin/data-transfer/	
3	**Validate**	Validation scripts are run by DP on JASMIN servers. Validation includes (i) identification of ESGF data sets, (ii) restructuring and reformatting of outputs in agreed format, (iii) raising warning and errors about data problems. Data is moved into "validated" directory.		DP
4	**Analyse**	Once in the "validated" directory, data can be used by PRIMAVERA researchers for analysis and intercomparison. Additionally, other data in GWSs or the CEDA archive can be processed on the same platform. The project team is encouraged to use the LOTUS cluster for this work.	http://jasmin.ac.uk/how-to-use-jasmin/lotus-documentation/ http://jasmin.ac.uk/how-to-use-jasmin/virtual-machines/	PRIMAVERA researchers
5	**Ingest**	Once data has been prepared for the archive the DP informs CEDA staff that a set of files are ready for ingestion. The CEDA team runs the ingestion process and informs the DP when completed. The version in the GWS can then be deleted to save disk space. Additionally, information will be derived from the data files (and discussion with the DP) in order to generate metadata for the CEDA Data Catalogue.		CEDA
6	**Disseminate**	On archival, the data will be accessible through the CEDA catalogue and download services. Additionally, where agreed, data sets will also be published through the Earth System Grid Federation (ESGF) services and will have Digital Object Identifiers (DOIs) assigned.		CEDA

The PRIMAVERA Data Management Tool

Within PRIMAVERA, we acknowledge that there are petabytes of data being generated and transferred through different parts of the workflow. Tracking this process manually is not feasible so we are developing a Data Management Tool to manage the process. This will have a web front-end, a database and a library that can be connected to various scripts running in the JASMIN environment. The intention is that project partners can run code to perform data conversions, conformance checking and metadata scanning and the project database will be automatically updated. This will allow data to be tracked, documented and updated in a dynamic way throughout the project.

Draft schema for the database

The following tables outline the proposed schema for the PRIMAVERA Data Management Tool. A number of tables include a field called "dreq_uid" which represents the specific Id used in the CMIP6 Data Request (see: https://earthsystemcog.org/projects/wip/CMIP6DataRequest). In the tables below foreign key relationships are indicated in the "Type" column with the prefix "FK:".

1. Model

Column	Type	Comment
id	String	Short Id for model used in ESGF identifiers
dreq_uid	String	The unique identifier used in the CMIP6 Data Request (if relevant)
long_name	String	Full long name of the model
institute	FK: Party	Institute where model was run

2. Party

A Party is a person or organisation.

Column	Type	Comment
id	String	Short Id for party (if type is "organisation") used in ESGF identifiers
dreq_uid	String	The unique identifier used in the CMIP6 Data Request (if relevant)
first_name	String	First name of party (if type is "individual")
last_name	String	Last name of individual or full name of organisation
type	String	One of: "individual", "organisation"

3. Experiment

Column	Type	Comment
id	String	Short Id for experiment used in ESGF identifiers
dreq_uid	String	The unique identifier used in the CMIP6 Data Request (if relevant)
long_name	String	Full long name of the experiment

4. Project

Column	Type	Comment
id	String	Short Id for project used in ESGF identifiers
dreq_uid	String	The unique identifier used in the CMIP6 Data Request (if relevant)
long_name	String	Full long name of the project

5. Variable

Column	Type	Comment
id	String	Short Id for variable used in ESGF identifiers
dreq_uid	String	The unique identifier used in the CMIP6 Data Request (if relevant)
long_name	String	Full long name of the variable
standard_name	String	CF standard name
units	String	Units of measure

6. Data Issue

A Data Issue can be related to various other objects (Tables) in the schema.

Column	Type	Comment
reason	String	e.g. "2 missing years - will not be provided."
reporter	FK: Party	Person who reported the issue.

date_time	DateTime	When the data issue was reported

7. Holding

A Holding represents a collection of files that are delivered to JASMIN together. They are treated as a single unit in order to track their processing through the system.

Column	Type	Comment
datasets	FK: ESGF Dataset(s)	
status	String	One of: "expected", "validated", "archived", "published"
files	FK: File(s)	
variables	FK: Variable(s)	
incoming_directory	String	The directory where the data was first delivered to on JASMIN
main_directory	String	The directory where the data is currently held
data_issues	FK: Data Issue(s)	
start_time	DateTime	Valid data start time derived from data files
end_time	DateTime	Valid data end time derived from data files

8. File

Column	Type	Comment
incoming_directory	String	The directory where the data was first delivered to on JASMIN
start_time	DateTime	Valid data start time derived from data file
end_time	DateTime	Valid data end time derived from data file
directory	String	The current directory containing the file

name	String	The file name
opendap_url	URL	URL where the data is accessible by OpenDAP (if available)
download_url	URL	URL where the data is accessible to download (if available)
esgf_download_url	URL	URL where the data is accessible to download via ESGF (if available)
checksum	String	The checksum of the file
checksum_type	String	The checksum type, e.g. "MD5" or "SHA256".
variable	FK: Variable	
project	FK: Project	
frequency	String	Time frequency of data in file (using DRS vocabulary)
tape_url	URL	URL to location on tape (if available)
on_line	Boolean	Boolean to indicate if online or on tape

9. ESGF Dataset

This is only relevant when in the archive and published to ESGF. However, the ESGF Dataset, and its DRS identifier, are the common unit across the entire Data Management Tool.

Column	**Type**	**Comment**
id	String	The ESGF DRS Identifier for this dataset
files	FK: File(s)	
variables	FK: Variable(s)	List of variables found in the files
directory	String	Directory path containing the dataset
version - ESGF version	String	ESGF version identifier (including the prefix "v"). E.g. "v20160822"

10. CEDA Dataset

This is only relevant when the data has been

Column	Type	Comment
doi	String	Digital Object Identifier for the CEDA Dataset (if applicable)
esgf_datasets	FK: ESGF Dataset(s)	
catalogue_url	URL	Location of CEDA Catalogue entry
data_issues	FK: Data Issue(s)	
directory	String	Directory path containing the dataset

11. Data Request

This information is derived from the official CMIP6 Data Request but is cached locally for convenience.

Column	Type	Comment
institute	FK: Party	
model	FK: Model	
experiment	FK: Experiment	
variable	FK: Variable	
frequency	String	Time frequency of data in file (using DRS vocabulary)
start_time	DateTime	Valid data start time required
end_time	DateTime	Valid data end time required

More information about this tool will be available in future versions of this Data Management Plan.

Appendix A: Output data

Total HPC usage and Data generated

HPC	Groups involved	Total HPC time (M Core hours)	Data volume generated (TB)
BSC:Marenostrum	BSC	8.9	218

CMCC:IBM iDataPlex	CMCC	17.3	460
DKRZ:ATOS/BULL X	MPG & AWI	25.0	283
KNMI:BULLX B500	KNMI	10.3	293
Met Office:Cray XC40	METOFFICE	56.8	453
Meteo France:Bullx	CERFACS	3.4	163
Munich:Supermuc	CNR	7.9	293
SMHI:Beskow (Cray)	SMHI	6.5	293
		Total:	**2,456**

Source [Internal wiki]: http://proj.badc.rl.ac.uk/primavera-private/wiki/WP9/HPC_Plan_stream_1 (05/04/2016)

Appendix B: Services and third party data

Third party/existing datasets

Data required by the project.

3rd Party Dataset Name	Contact	Location and contents	Responsibility	Comments
ECMWF ERA-Interim re-analysis	Ag Stephens	This is held in the CEDA archive. See catalogue record for more details: http://catalogue.ceda.ac.uk/uuid/00f58d1d7b6c8f38993e77c79e72da92	CEDA - data already exists in the archive.	Data is held on an open licence and is available to all. Users can access the data on JASMIN directly under: /badc/ecmwf-era-interim/data

4. Lessons Learnt

Producing this plan was a complex task involving input from different areas, from people with technical skills in different areas. It covers the evolution of the project from data standards agreement, production, processing, movement, and archival. It is anticipated that the next iteration will be the time to ‘look back’ and consider what the lessons are.

5. Links Built

This report involved interactions with all the WPs.