Ref. Ares(2020)7218452 - 30/11/2020

First update on Data Management Plan

Deliverable 5.3

Authors: Leo de Sousa-Webb (UNEXE), Pierre Friedlingstein (UNEXE)

This project received funding from the Horizon 2020 programme under the grant agreement No. 821003.

Document Information

GRANT AGREEMENT	821003
PROJECT TITLE	Climate Carbon Interactions in the Current Century
PROJECT ACRONYM	4C
PROJECT START DATE	2019-06-01
RELATED WORK PACKAGE	W5
RELATED TASK(S)	T5.1
LEAD ORGANIZATION	UNEXE
AUTHORS	Leo de Sousa-Webb (UNEXE), Pierre Friedlingstein (UNEXE)
SUBMISSION DATE	2020-11-30
DISSEMINATION LEVEL	PU

History

DATE	SUBMITTED BY	REVIEWED BY	VISION (NOTES)
2020-04-27	Leo de Sousa-Webb (UNEXE)	Pierre Friedlingstein (UNEXE)	Initial update on PO feedback
2020-11-27	Pierre Friedlingstein (UNEXE)		Update on PO feedbacks and new information from partners

Please cite this report as: de Sousa-Webb, L., Friedlingstein, F, (2020), First update on Data Management Plan, D5.3 of the 4C project

Disclaimer: The content of this deliverable reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

Table of Contents

List of tables

About 4C

Climate-Carbon Interactions in the Current Century (4C) is an EU-funded H2020 project that addresses the crucial knowledge gap in the climate sensitivity to carbon dioxide emissions, by reducing the uncertainty in our quantitative understanding of carbon-climate interactions and feedbacks. This will be achieved through innovative integration of models and observations, providing new constraints on modelled carbon-climate interactions and climate projections, and supporting Intergovernmental Panel on Climate Change (IPCC) assessments and policy objectives.

Executive Summary

The purpose of the Data Management Plan (DMP) is to provide all needed information on the data generated over the course of the project. The DMP follows the FAIR data management, using the Horizon 2020 data management plan template.

The research team will have regular electronic meetings to ensure all members have the same research data management procedures and policies in place adhering to the Data Management Plan. The senior project member at each institution will be responsible for research data management at that institution. The project PI will have overall responsibility for data management.

This first update draws on internal and external feedback and newer best practice to iterate improvements and clarifications across the consortium as a whole and each partner individually.

This initial DMP was submitted in March 2020 (Deliverable 5.2) and will be updated again in June 2022 (Deliverable D5.4).

Keywords

Data, Management, Plan, dataset, DMP, FAIR, Horizon, server, open, access, pilot, observation-based, dataset, Earth, System, Model, simulation, climate-carbon, interactions, carbon, budget, carbon, cycle, near-term, prediction, climate, projections

1 Introduction

The purpose of the Data Management Plan (DMP) is to provide all needed information on the data generated over the course of the project. The DMP follows the FAIR data management, using the Horizon 2020 data management plan template.

2 Data Summary

2.1 Overview

All data generated during the project will serve the objectives of the project. In particular:

all “observation based datasets” (see Table 1. below) will serve the overall objective 1 of the project: “*Better understanding of processes controlling the global carbon cycle* “

“model simulation datasets” from WP1 (see Table 2. below) will serve the overall objective 1 of the project: “*Better understanding of processes controlling the global carbon cycle* “

“model simulation datasets” from WP2 (see Table 2. below) will serve the overall objective 2 of the project: “*Towards a near-term prediction of the climate and carbon cycle* “

“model Simulation datasets” from WP3 (see Table 2. below) will serve the overall objective 3 of the project: “*Reducing uncertainties in climate projections over the 21st century”*

Workshop participants lists, pictures and videos will serve the organisation of the project scientific and communication activities (see Table 3. below).

All observation-based datasets and model simulation datasets will be in netcdf format, compliant with the data standards established in the community. All Earth System Models (ESM) outputs will be following the CMIP6 CMOR standard. Offline land-only and ocean-only models outputs will be following the CMIP5/6 standards.

Other existing datasets will be used over the course of the project in order to provide further constraints on the global carbon cycle, addressing the overall objective 1 of the project: “*Better understanding of processes controlling the global carbon cycle* “. These existing datasets are listed in the Grant Agreement, Table 1.1a “Available existing observations and observation-based data products used in 4C”.

The origin of each data generated during the project is described in the Data Summary tables above. The overall size of the data produced by the project is expected to amount to ~50Gb for observation-based datasets, and to ~10Tb for model simulation datasets. See Data Summary tables above for details.

The observation-based data and the model simulation data generated during the project will be primarily used by the project partners to achieve the objectives of the project. All data will be made available to all project partners as soon as they are produced, quality checked, and meta-described. All data will also be made publicly available, no later than 12 months after that time for model simulations datasets (to protect any 4C scientific publication embargo).

The research team will have regular electronic meetings to ensure all members have the same research data management procedures and policies in place adhering to the Data Management Plan. The senior project

member at each institution will be responsible for research data management at that institution. The project PI will have overall responsibility for data management.

Tables 1 and 2 below summarise the datasets to be produced within 4C and the key information with respect to data production and availability.

Table 1. Observation-based datasets - Summary

Type of Data/Format	*Reason for Collection*	*WP*	*Lead Partner*	*Expected Size of Data (per year)*	*Origin of Data*	*Level of Access, (IPR)*	*How will data be disseminated during the project*	*How will data be available after the project (re-use)*	*Timeline of availability*
Observation based datasets									
Satellite XCO2	Land/ocean carbon sinks	WP1	UBREMEN	~10Gb	XCO2 concentrations retrieved from satellite data (SCIAMACHY, GOSAT, OCO-2).	Publicly available	See section 2.2	See section 2.3	12/2020
Neural network air-sea C fluxes	Ocean carbon sink	WP1	MPG	<1Gb	Air-sea CO2 flux from a 2-step neural network model	As above	As above	As above	08/2020
Ocean interior C change	Ocean carbon sink and storage	WP1	ETHZ	<1Gb	Ocean DIC from a multiple linear regression (eMLR(C*)) model	As above	As above	As above	Autumn 2021
Terrestrial Water Storage	Land water-carbon interactions	WP1	ETHZ	<1Gb	Land water mass retrieved from from GRACE satelllite	As above	As above	As above	Autumn 2021
Land-Flux EVAL dataset	Land carbon sink	WP1	ETHZ	<1Gb	Land evaporation derived from reanalysis/land surface models	As above	As above	As above	Autumn 2021
Machine learning Forest NBP	Land carbon sink	WP1	CEA	<1Gb	Forest net annual CO2 flux from machine learning model	As above	As above	As above	Autumn 2021

Table 2. Model Simulations datasets - Summary

Type of Data/Format	*Reason for Collection*	*WP*	*Partners involved*	*Expected Size of Data (per model)*	*Origin of Data*	*Level of Access (IPR)*	*Dissemination during the project*	*Availability after the project*	*Timeline of availability*
Model Simulation datasets									
Forced historical run land carbon	Understanding processes causing land carbon sinks; model evaluation	WP1	UNEXE, MPG, BSC, UBERN, CEA	<1Gb	Land carbon cycle model simulations	Public no later than 12 months after production	See section 2.2	See section 2.3	Month 36
Forced historical run ocean carbon	Understanding processes causing ocean carbon sinks; model evaluation	WP1	UEA, ENS, MPG, ETHZ, BSC	~10Gb	Ocean carbon cycle model simulations	As above	As above	As above	Month 36
Forced historical run ocean carbon high resolution	Quantifying the effect of small-scale processes on ocean carbon variability	WP1	UEA	~100Gb	Ocean carbon cycle model simulations	As above	As above	As above	Month 36
Historical coupled simulation	Evaluation of global carbon cycle; decadal predictions; emergent constraints	WP1	ENS, MPG, BSC, UBERN, CEA	~10Gb	Earth System model simulations	As above	As above	As above	Month 36
Factorial experiments individual forcings	Attribution of carbon cycle changes to drivers	WP1	UNEXE, UEA, MPG, ETHZ, UBERN, CEA	~100Gb	Land & ocean carbon cycle model simulations	As above	As above	As above	Month 45
Perfect model decadal predictions	Assess potential predictability of climate-carbon system	WP2	BSC, ENS, MPG, CEA	~500Gb	Earth System model simulations	As above	As above	As above	Month 18
Data-assimilated reconstruction	Provide initial conditions for hindcast and future predictions	WP2	BSC, ENS, MPG, CEA	~500Gb	Earth System model simulations	As above	As above	As above	Month 28
Retrospective decadal predictions (Conc. driven)	Assess predictability against observations Bias correction estimate	WP2	BSC, ENS, MPG, CEA	~500Gb	Earth System model simulations	As above	As above	As above	Month 36
Retrospective decadal predictions (Emis. driven)	Access predictability of atmospheric CO_2 against observations	WP2	BSC, ENS, MPG, CEA	~500Gb	Earth System model simulations	As above	As above	As above	Month 36
Future decadal predictions (NDCs and baseline)	Prediction of next decade of atmospheric CO_2, carbon and climate	WP2	BSC, ENS, MPG, CEA	~100Gb	Earth System model simulations	As above	As above	As above	Month 45
Adaptive scenarios projections	Assessment of TCRE, remaining carbon budget, climate response	WP3	ENS, MPG, BSC, UBERN, CEA	~1Tb	Earth System model simulations	As above	As above	As above	Month 45

Table 3. Project participants data - Summary

Type of Data/Format	*Reason for Collection*	*WP*	*Lead Partner*	*Expected Size of Data*	*Origin of Data*	*Level of Access*	*How will data be disseminated during the project*	*How will data be available after the project (re-use)*	*Data Utility*
Project participants data									
Workshop participants list / xls, doc, pdf	Participant names, positions, institutions, email addresses, dates attending, dietary requirements, meal preferences and access requirements for event management purposes.	WP4	UNEXE	<1Mb	Created by emailing potential attendees and collating the responses.	Restricted due to data protection	An agenda with the participant list will be disseminated to workshop participants, but only name, position and institution would be listed.	N/A	Workshops may be attended by members of the consortium, members from the European Commission and policymakers.
Pictures and videos of workshops and other events such as seminars / jpg, mp4, mov	Communication	WP4	UNEXE	<1Gb	Photos and videos of 4C workshops and events	Public as long as consent has been obtained to record events and to share the photos and videos.	Photos and videos may appear on the website or on the websites of the institutions of the Coordinator and partners.	website will be available for 5 years after the project has finished	Public

2.2 Data accessibility

All observation-based datasets and model simulation datasets will be either stored on the partner's institution servers or an open-access public repository, guaranteeing long term archival and public open access. More specifically for each dataset described in Table 1,

- The XCO2 concentrations from satellite data will be made publicly via the partner website: https://www.iup.uni-bremen.de/carbon_ghg/cg_data.html#satellite_XCO2_for_4C

- The neural network-based air-sea CO2 flux estimates is publicly available through the National Center of Environmental Information (NCEI) Ocean Carbon Data System (OCADS) https://www.ncei.noaa.gov/access/ocean-carbon-data-system/oceans/SPCO2_1982_present_ETH_SOM_FFN.html
  DOI : https://doi.org/10.7289/V5Z899N6

- The ocean interior C* based estimates will be made publicly available by Autumn 2021 through the National Center of Environmental Information (NCEI) Ocean Carbon Data System (OCADS) : https://www.ncei.noaa.gov/access/ocean-carbon-data-system/oceans/ndp_100/ndp100.html.

- The Terrestrial Water Storage and the Land-Flux EVAL dataset will be made available via the ETH data server http://iacweb.ethz.ch) with all data available upon registration. The, multi-forcing observation based global runoff reanalysis (GRUN-ENSEMBLE) has already been produced, with the data description paper under review (Ghiggi et al., Water Resources Research, in review). The runoff reconstruction dataset is a key element to estimate evapotranspiration in combination with reconstructions of terrestrial water storage. When the publication is accepted, the datasets will be hosted on the permanent repository https://figshare.com/ with a reserved doi number. For reviewing purposes the data is held on a preliminary repository (https://figshare.com/s/ad6d5cdfbba945d93ad2). Further products for evapotranspiration and terrestrial water storage will be released within 4C in Autumn 2021.

- The Machine learning observation-based Forest NBP dataset will be made publicly available through the ICOS carbon portal after the end of the project and through the LSCE local server (sharebox) during the project duration. The dataset will be publicly available in Autumn 2021.

All model simulation datasets (Table 2) will be stored on the partner's institution servers, with the models' data publicly available from the partners' servers on request. More specifically:

- ETHZ will store their own primary model output on its own servers but will put analysed model data to the digital library of ETH for open access.
- MPG will archive model data on the DKRZ server
- UBERN will store their own model output on institute servers, but data from the Bern3D model will also be uploaded on ZENODO, a common public repository.
- UNEXE will store their own model output on the TRENDY server hosted at the University of Exeter, with data available on request
- BSC will store their own primary model output on institute servers, but a selection of model output will be also uploaded on EUDAT (https://b2share.eudat.eu/) from where it will be publicly accessible.
- UEA will store their full primary model output on its own servers which are securely backed up. It will put a selection of model output accessible from the UEA web site https://www.uea.ac.uk/green-oceanmodel which is being updated, and on the UK BODC data centre (up to 1Gb of data). All analysed model output used in publications will be made available through the journals.
- CEA and ENS will store their own primary model output on the French national computing centres (GENCI). A selection of model output will be accessible through the IPSL institute server (CICLAD, platform http://ciclad-web.ipsl.jussieu.fr).

2.3 Long term storage

All data produced within the lifetime of the project will be available after the end of the project for longer-term storage. For model simulation datasets, stored on institutions data servers, the following will apply:

- MPG will provide long-term archive on tape as part of Good Scientific Practice at MPI-M;
- ETHZ will provide long-term storage and accessibility to high-level data through the digital library (https://www.library.ethz.ch/en/ms/Research-Data-Management-and-Digital-Curation);
- CEA and ENS will provide free long-term storage and accessibility to high level data through local server or IPSL CICLAD platform http://ciclad-web.ipsl.jussieu.fr
- UNEXE will be providing long-term storage and accessibility to high-level data through the ORE university repository;
- UEA will provide long-term storage that is backed up with fast access up to the end of the project. Key model output will be archived at the UK BODC;
- BSC will provide the data on demand initially through a public ftp server. The data provided on the ftp will be a copy of the data hosted on the BSC tapes, ensuring their long-term preservation and accessibility. In a second stage, access through a public THREDDS server will be provided.

2.4 Data Quality control

Observation based data will follow a data quality assurance process, with a quality check done by the data PI. In particular:

- The 4C satellite-derived XCO2 Obs4MIPs format data set (2003-2019, monthly, 5x5 spatial resolution) has been quality assessed by comparisons with ground-based XCO2 retrievals from the Total Carbon Column Observing Network (TCCON).
- Quality of the neural network-based air-sea CO2 flux estimates product is 2-fold. Firstly, via rigorous independent data testing. Secondly, via participation in the Surface Ocean CO2 Mapping intercomparison (SOCOM) project where available estimates are compared.
- Quality of the ocean interior C estimates is assessed by (i) extensive testing of the methodology using synthetic data generated from a global biogeochemical model (Clement and Gruber, 2018), and by comparing the estimates against independent constraints, such as the concentration of other man-made substances, such as chlorofluorocarbons. Finally, the transient steady-state assumption provides also a very good zero order estimate of the expected changes in the ocean storage of anthropogenic CO2 (see discussion in Gruber et al., 2019).
- Data quality control of the forest NBP dataset is performed by the data provider (CEA). Different estimates from ensemble members will be compared for uncertainty estimation. Comparison with independent forest biomass stock change from inventories will be performed to assess the quality of the product against independent estimate at biome scale.

3 FAIR Data description

3.1 Making data findable, including provisions for metadata

Observation-based data produced will be associated with a unique DOI and publicly available on project and partners servers, with meta-data provided. If needed, version numbers will be provided to ensure data traceability. Section 2.2 above described specific repositories for the observation and model based datasets produced within 4C

Key relevant outputs from land-only, ocean-only and Earth System models simulations will be publicly available on project and partners servers, with meta-data provided.

The 4C project website has a “4C data” entry with a description of available datasets and information on data access. Information on the 4C website will be provided as soon as the datasets are being released/

Visibility of the EC funding through the 4C grant will be ensured with proper acknowledgement on the different data distribution components and via a clear description on the 4C web site (https://4c-carbon.eu/resources/datasets).

3.2 Making data openly accessible

All observation-based datasets and model simulation datasets in the project will be made available to all project partners as soon as they are produced, quality checked, formatted and meta-described. All data will be made openly available latest no later than12 months after that time. No restriction will apply for the project generated datasets.

All observation-based datasets and model simulation datasets will be in netcdf format. Usual data transfer tools are sftp or wget; usual netcdf data manipulation software are NCL or NCO; and usual data analysis or visualisation software are python, R, or ferret.

The software listed above are typical of tools used of large Earth science datasets; all of these software are freely available and have extensive online documentation. The source codes of these software need to be installed on the UNIX machine of the data user. Up-to-date, operating system specific, sources are available from the software developers.

When data will be publicly accessible, the identity of persons accessing the data after will not be monitored. Access to models datasets kept on partners servers might be monitored if access credentials need to be provided.

3.3 Making data interoperable

As mentioned before, observation-based datasets and model simulation datasets will be in netcdf format. Deliverables, papers and publications will use Microsoft Office formats (.doc, .xls, .ppt) or PDF from Adobe. Pictures will use .jpg or .tiff and videos will use .mp4 or .mov.

Observation-based datasets and model simulation datasets will use the standard CMIP6 compliant convention, using standard vocabularies (https://pcmdi.llnl.gov/CMIP6/Guide/dataUsers.html#3-accessing-model-output).

The metadata output will follow the standard NetCDF Climate and Forecast (CF) Metadata convention, using the standard variable names, units, dimensions, axis, required 'coordinates' attribute, etc., following the CMIP6 models outputs meta-data requirement: https://pcmdi.llnl.gov/CMIP6/Guide/dataUsers.html#2-model-output-specifications

Whenever possible, file naming convention for model simulation datasets will follow the CMIP6 naming convention, that is :

filename = <*variable name*>_<*model*>_<*experiment*>_[*ensemble member]*_<*temporal subset*>.nc (such as nbp_JULES_ historical_ 185001-201412.nc).

3.4 Increase data re-use (through clarifying licences)

All data generated during the project will be licensed to permit wider re-use, all data providers adopting the Creative Commons CC BY license.

The data produced will be useable by third parties. There will be no restriction on data-use after the end of the project.

There is no time limit on how long data can be used.

4 Allocation of Resources for data storage

All partners have in-kind resources for long term storage of the data generated by the project, see section 2.3 for institutional storage capability of model simulation datasets.

There are no additional costs anticipated for long-term preseravtion of the 4C generated datasets. Long-term preservation of these data will be provided as follow:

- MPG: data stored on tape for long-term archive i.e. >5 years after end of project;
- UBERN: data preserved on partner servers for 5 years after end of project;
- ETHZ: high-level data stored on partner servers and cured for >10 years after end of project;
- BSC: data preserved on partner servers for 5 years after end of project;
- CEA and ENS: data preserved on partner servers for 5 years after end of project;
- UEA will archive model data on local servers for at least 5 years after end of project;
- UNEXE, data preserved for 5 years after end of project.

5 Data Security

All consortium-shared and processed data will be stored in secure environments at the locations of consortium partners with access privileges restricted to the relevant project partners. Data storage will be secured and backed up on a local network.

Observation-based datasets will be archived on public repositories for long term preservation (see section 2.2 above)

Storage of sensitive data such as the Project participants data (see Table 3. above) will comply with the requirements of the General Data Protection Regulation (GDPR) and University policies. All data covered by the GDPR will be password protected and kept on secure University of Exeter filespace. The data produced by the University of Exeter will be stored on the University of Exeter network. Each researcher is allocated up to 20GB of secure, backed up network storage.

6 Ethical Aspects

We have an informed consent form for sharing personal data (Table 3. above) where necessary to organise workshops (please see D6.2). No personal data will be kept long term. The project will comply with the requirements of GDPR and University policies. All data covered by the GDPR will be password protected and kept on secure University filespace.

7 Institutional procedures for data management

Each partner abides and complies with their own institutional data management procedures, policies and systems, including country-specific regulations which are in line or in addition to the European directive on General Data Protection Regulation (GDPR).