Data Management Plan

Deliverable 5.2

Author: 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	01/06/2019
RELATED WORK PACKAGE	WP5 – Project Management
RELATED TASK(S)	T5.1
LEAD ORGANIZATION	UNEXE
AUTHORS	Pierre Friedlingstein
SUBMISSION DATE	31/03/2020
DISSEMINATION LEVEL	PU

History

DATE	SUBMITTED BY	REVIEWED BY	VISION (NOTES)
13.1.2020	Pierre Friedlingstein (UNEXE)	Claire Banwell (UNEXE)	Initial version
16.1.2020	Pierre Friedlingstein (UNEXE)		Second version with inputs from the Exeter Research Data Officer
30.1.2020	Pierre Friedlingstein (UNEXE)	All 4C PIs	Version shared with the consortium
18.2.2020	Pierre Friedlingstein (UNEXE)	All 4C PIs	First revised version shared with the consortium
10.3.2020	Pierre Friedlingstein (UNEXE)	All 4C PIs	Second revised version shared with the consortium
31.3.2020	Pierre Friedlingstein (UNEXE)	All 4C Pis Leo de Sousa-Webb (UNEXE)	Final version with input from PO. Proofreading.

Please cite this report as: Friedlingstein, P. (2020), Data Management, D5.2 of the 4C project

Table of Contents

List of tables

About 4C

Climate-Carbon Interactions in the Coming 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 DMP will be updated in November 2020 (Deliverable 5.2) and in June 2022 (Deliverable 5.3).

Keywords

Data Management Plan, 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

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. Data will be made available to all project partners

as soon as they are produced, quality checked, and meta-described. Data will be made publicly available no later than 12 months after that time.

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.

As a result of evolving science, technologies, best practice and new regulations, this DMP will be updated in November 2020 (Deliverable 5.2) and in June 2022 (Deliverable 5.3).

Table 1. Observation-based data - Summary

Type of Data/Format	*Reason for Collection*	*WP*	*Lead Partner*	*Expected Size of Data (per year)*	*Origin of Data*	*Level of Access (Public/Confidential)*	*How will data be disseminated during the project*	*How will data be available after the project (re-use)*	*Data Utility*
Observation based datasets									
Satellite XCO2	Land/ocean carbon sinks	WP1	UBREMEN	~10Gb	XCO2 concentrations retrieved from satellite data (SCIAMACHY, GOSAT, OCO-2).	Publicly available	Shared on project and partners servers, with meta-data provided	Shared on project and partners servers, with meta-data provided	Project partners and scientific community
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	As above
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	As above
Terrestrial Water Storage	Land water-carbon interactions	WP1	ETHZ	<1Gb	Land water mass retrieved from from GRACE satelllite	As above	As above	As above	As above
Land-Flux EVAL dataset	Land carbon sink	WP1	ETHZ	<1Gb	Land evaporation derived from reanalysis/land surface models	As above	As above	As above	As above
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	As above

Table 2. Model simulations data - Summary

Type of Data/Format	*Reason for Collection*	*WP*	*Partners involved*	*Expected Size of Data (per model)*	*Origin of Data*	*Level of Access*	*Dissemination during the project*	*Availability after the project*	*Data Utility*
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 within 1 year of production	Shared publicly on project and partners servers	Shared publicly on project and partners servers	Project partners and scientific community
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	As above
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	As above
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	As above
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	As above
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	As above
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	As above
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	As above
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	As above
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	As above
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	As above

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 (Public/Confidential)*	*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

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.

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 will have a "4C data" entry with a description of available datasets and information on data access.

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 will also 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).

3.2 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.

All observation-based datasets and model simulation datasets will be either stored on the partner's institution servers or an open-access public repository which guarantee long term archival and public open access. More specifically:

- The XCO2 concentrations from satellite data will be made publicly available through the Copernicus Climate Change Service (C3S, https://climate.copernicus.eu/ where they will receive a formal DOI and be citable;
- The neural network-based air-sea CO2 flux estimates will be instantly made publicly available through the National Center of Environmental Information (NCEI) Ocean Carbon Data System (OCADS) where they will receive a formal DOI and be citable;
- The Machine learning Forest NBP data will be made publicly available through the sharebox server hosted by LSCE

- The ocean interior C* based estimates will be made publicly available through the National Center of Environmental Information (NCEI) Ocean Carbon Data System (OCADS) where they will receive a formal DOI and be citable;
- Terrestrial Water Storage and Land-Flux EVAL dataset (TBD).

All model simulation datasets 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 could also be uploaded on a common public repository (to be defined in the near future).
- UNEXE will store their own model output on the TRENDY server hosted at the University of Exeter.
- BSC will store their own primary model output on institute servers, but data could also be uploaded on a common public repository (to be defined in the near future).
- 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 their institute server (CICLAD).

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.

No restriction will apply for the project generated datasets. Data will be made available to all project partners as soon as they are produced, quality checked, and meta-described. Data will be made publicly available no later than 12 months after that time.

When data will be publicly accessible (within 12 months so of being produced, 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).

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.

All observation-based datasets and model simulation datasets in the project will be made openly available no later than 12 months after being produced, quality checked, formatted and meta-described.

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.

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

MPG: Quality of the air-sea CO2 flux 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.

All land-only model data will have a quality check performed by UNEXE via the use of the ILAMB benchmarking tool (https://www.ilamb.org/).

4 Allocation of Resources

All partners have in-kind resources for long term storage of the data generated by the project. In particular:

- 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 long-term storage and accessibility, free of charge via the partner dods server;

- 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.

There are no additional costs anticipated for long-term storage 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 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.