CoCO2

Prototype system for a Copernicus CO_2 service

Data Management Plan

Daniel Thiemert

coco2-project.eu

Co-ordinated by ECMWF

CoCO2

Prototype system for a Copernicus CO_2 service

D9.5 Data Management Plan

Dissemination Level:	Public
Author(s):	Daniel Thiemert (ECMWF)
Date:	30/06/2021
Version:	1.0
Contractual Delivery Date:	30/06/2021
Work Package/ Task:	WP9/ T9.4
Document Owner:	ECMWF
Contributors:	WP Leaders
Status:	Final

CoCO2: Prototype system for a Copernicus CO_2 service

Coordination and Support Action (CSA)
H2020-IBA-SPACE-CHE2-2019 Copernicus evolution – Research activities in support of a European operational monitoring support capacity for fossil CO2 emissions

Project Coordinator: Dr Richard Engelen (ECMWF)
Project Start Date: 01/01/2021
Project Duration: 36 months

Published by the CoCO2 Consortium

Contact:
ECMWF, Shinfield Park, Reading, RG2 9AX,
richard.engelen@ecmwf.int

The CoCO2 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 958927.

Table of Contents

1 Executive Summary 5
2 Introduction 5
2.1 Background 5
2.2 Scope of this deliverable 5
2.2.1 Objectives of this deliverables 5
2.2.2 Work performed in this deliverable 5
2.2.3 Deviations and counter measures 5
3 Open Research Data Objectives 6
3.1 Open Research Data Pilot 6
3.2 CoCO2 Research Data 6
3.3 Data Management Plan Questionnaire 6
4 CoCO2 Data Sets 8
4.1 Work Package 2 8
4.2 Work Package 3 10
4.3 Work Package 4 15
4.4 Work Package 5 17
4.5 Work Package 6 19
4.6 Work Package 7 21
5 Conclusion 23

Tables

Table 1: Data Management Plan Questionnaire 7

1 Executive Summary

The CoCO2 Data Management Plan responds to the requirements of the H2020 Open Research Data Pilot to document which research data is being produced by the CoCO2 project, in which format, and how it will be made available.

It has already identified data sets for work packages 2 to 7, but is only to be seen as an initial version which requires periodic updates to provide the necessary detail as it emerges.

2 Introduction

2.1 Background

To support EU countries in assessing their progress for reaching their targets agreed in the Paris Agreement, the European Commission has clearly stated that a way to monitor anthropogenic CO_2 emissions is needed. Such a capacity would deliver consistent and reliable information to support policy- and decision-making processes.

To maintain Europe's independence in this domain, it is imperative that the EU establishes an observation-based operational anthropogenic CO_2 emissions Monitoring and Verification Support (MVS) capacity as part of its Copernicus programme.

The CoCO2 Coordination and Support Action is intended as a continuation of the CO_2 Human Emissions (CHE) project, led by ECMWF. In the Work Programme, ECMWF is identified as the predefined beneficiary tasked to further develop the prototype system for the foreseen MVS capacity together with partners principally based on the CHE consortium. In addition, ECMWF will continue some of the work initiated in the VERIFY project as well.

The main objective of CoCO2 is to perform R&D activities identified as a need in the CHE project and strongly recommended by the European Commission's CO_2 monitoring Task Force. The activities shall sustain the development of a European capacity for monitoring anthropogenic CO_2 emissions. The activities will address all components of the system, such as atmospheric transport models, re-analysis, data assimilation techniques, bottom-up estimation, in-situ networks and ancillary measurements needed to address the attribution of CO2 emissions. The aim is to have prototype systems at the required spatial scales ready by the end of the project as input for the foreseen Copernicus CO_2 service element.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverables

D9.5 Data Management Plan provides the initial outline of the data management plan including information on which data sets will be created in the project and how they will be made available. This document represents only the initial version where details may not be available yet, and it will be further developed over the course of the project.

2.2.2 Work performed in this deliverable

The work performed included, as per the DoA, the collection of the available descriptions of data sets to be produced by the project, through a questionnaire.

2.2.3 Deviations and counter measures

No deviations have been encountered.

3 Open Research Data Objectives

3.1 Open Research Data Pilot

As per the Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020[1], Research Data

> "Refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation.
>
> In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form."

The Open Research Data Pilot

> "aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects[2]"

and applies to data sets that are

> "needed to validate the results presented in scientific publications[2]".

The Data Management Plan is expected to

> "specify what data will be open: detailing what data the project will generate, whether and how "it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved[2]".

3.2 CoCO2 Research Data

As per the CoCO2 Description of Action, the products of CoCO2 will comprise reports, graphical displays, datasets and improved methods, algorithms and code. All these elements have their own important role. Reports are mostly targeted at informing the Commission and its Task Force on assessments, innovation progress and future directions. Graphical displays, where applicable, are targeted at all users as supportive information for the various model runs, method comparisons, and input datasets. The datasets will also target a wide user community to support them with parallel or alternative studies. Finally, improved methods, algorithms and code are meant to form the basis for follow-on development after the CoCO2 project has finished.

All mature data products of CoCO2 will be made publicly available to maximize the uptake by the scientific community. It is envisaged to make use of three parallel data portals to ensure full visibility of the datasets. These data portals will be based on the ICOS Carbon portal, the Global Carbon Atlas and the Copernicus Atmosphere Data Store, which is currently being implemented by the Copernicus Atmosphere Monitoring Service (CAMS). The steps undertaken by CoCO2 towards building a European Platform for Monitoring CO2 anthropogenic emissions contribute directly to this operational requirement.

3.3 Data Management Plan Questionnaire

The following questionnaire has been provided to CoCO2 work packages to gather the information for this first version of the Data Management Plan.

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[2] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

Table 1: Data Management Plan Questionnaire

<Data set reference and name>	
Data set description	*Description of the data that will be generated or collected (or is already available to the project), its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.* *Limitations?* *Constraints?*
Standards and metadata	*Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.* *Will you generate proper metadata for your data?* *If yes: how do they look like?* *If no: why?* *Data format?* *Will there be a review process to quality- check the data?*
Data Sharing	*Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).* *In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).* *License?*

	Access URL?
Archiving and preservation (including storage and backup)	*Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.* *At which Data Center do you want to store your data? Is there an established workflow for your requested DOI process in place? According to which standards*

4 CoCO2 Data Sets

The following sections provide the responses by work packages 2 to 7. Work Packages 1, 8 and 9 do not produce any data sets.

4.1 Work Package 2

Global Mosaic CO2, NOx, CO, CH4 for 2015/2018	
Data set description	- Mosaic of regional emission grid maps for NOx and CO, gap-filled with EDGARv5 dataset for the years 2015 and 2018, produced at a coherent grid with resolution of 0.1degx0.1deg. - Mosaic of regional emission grid maps for fossil CO2 and biofuel CO2, gap-filled with EDGARv5, produced at a coherent grid with resolution of 0.1degx0.1deg for the year 2018 - Mosaic of regional emission gridmaps for CH4 for the year 2015, gap-filled with EDGARv5 dataset, produced at a resolution of 0.1degx0.1deg. Users are potentially national inventory agencies. Similar exercise: The Hemispheric Transport of Air Pollution published HTAPv2.2 (2015) and is now working on a release of HTAPv3 Limitations: these gridmaps are collated together and inconsistencies and border artifacts might occur Constraints: the quality is determined by the emission data source: CAMS-REG, US-EPA+Environment Canada, REAS (South/south-East/East Asia) v3.1, MIX-Asia (MEIC, 2010-2017), gap-filled with EDGARv6
Standards and metadata	Reference to HTAPv2.2 methodology, and documentation of the original datasets. A proper report

	and metadata file will be produced (according to JRC/EC standards such as INSPIRE etc.) Data format: global gridmaps in netcdf and txt on 0.1degx0.1deg resolution, for one single year. Will there be a review process to quality- check the data: yes
Data Sharing	Data will be also shared with the colleagues of the Global Emissions Initiative (GEIA) and its data portal ECCAD.
Archiving and preservation (including storage and backup)	Data is archived in the JRC data catalogue, providing also a DOI. In addition, for publishing the dataset zenodo might be used if needed.

1-km resolution biospheric fluxes for Europe from VPRM, in Task 2.1	
Data set description	The dataset consists of GPP and respiration fluxes at 1-km spatial and hourly temporal resolution for the European domain (60 W to 30 E, 30 N to 75 N), generated by the Vegetation Photosynthesis Respiration Model (VPRM, Mahadevan et al., 2008). The model uses the indices EVI and LSWI, calculated from MODIS reflectance data, as well as temperature and shortwave radiation data, which in this case will be taken from ERA5 reanalyses. The first version of the dataset will be produced for the year 2018, and a follow-on product will be produced for 2021. The data are expected to be useful as prior fluxes for regional CO2 flux inversions, but may have other applications. Limitations: The model is known to perform rather poorly for managed croplands.
Standards and metadata	Will you generate proper metadata for your data? Yes. The metadata will include all the information describing the input products used to run the model (including the parameters used) in addition to the standard variable descriptions and attributes. Data format? The data will be prepared in daily netcdf files, divided into sub-regions if required due to file size constraints. Will there be a review process to quality- check the data? The dataset will be used within Task 4.4, and user feedback is expected.
Data Sharing	The exact platform on which the data will be distributed has not yet been determined, and is expected to resolved at the project level. (This is only one of many large datasets that will need to be shared.) In the shorter term, it will be shared through a public link to

	the German supercomputing centre DKRZ (https://swiftbrowser.dkrz.de/ - link to be updated once the dataset is complete). Software in python will be provided to assist users in re-gridding the fluxes to match their (curvilinear or rectilinear) model grid. License: CC-like license, to be defined.
Archiving and preservation (including storage and backup)	The data can be shared and archived through whichever repository is decided upon for the project as a whole. Details will be confirmed at a later point.

4.2 Work Package 3

IFS inversion product (gridded fluxes and concentrations): 2018 and 2021	
Data set description	Inversion of global carbon dioxide and methane emissions associated with anthropogenic and biogenic processes using the IFS 4D-Var system. Emission estimates will be provided globally daily at 25km resolution for 2018 and 2021 together with their posterior uncertainties. Evaluation of the product through in situ comparisons with atmospheric measurements (TCCON network) will also be made available. The posterior emission dataset will be used to assess the potential of the global IFS inversion prototype to provide timely update (i.e., daily) of anthropogenic emissions at high spatial resolution.
Standards and metadata	Will you generate proper metadata for your data? Yes NetCDF/grib header Data format? Grib and NetCDF Will there be a review process to quality- check the data? Yes
Data Sharing	Web API with examples on how to retrieve the data from the mars archive
Archiving and preservation (including storage and backup)	Mars archive

Carbon Tracker inversion product (gridded fluxes and concentrations): 2018 and 2021	
Data set description	Optimized biosphere fluxes (using the gridded state vector of CT, splitting ocean/land (with specific ecoregions). Fluxes are optimized at weekly time resolution, using a 5 weekly assimilation window. Carbon Tracker Europe (CTE) employs a global resolution of 3x2 degrees (lon x lat), with a zoom region over Europe (1x1 degree). Production of CTE fluxes is part of the yearly recurring global carbon project (GCP).

	Data will be published as part of the GCP yearly analysis, and are freely available to the community. Availability of the data is in the fall after the year ended (e.g. 2020 is available in fall 2021
Standards and metadata	As part of the yearly GCP analysis, the data are quality checked and compared to other GCP flux products. All information (fluxes, comparison to assimilated data, aggregated totals) are presented on: https://www.carbontracker.eu Data can be downloaded at netCFD4 flies, with full meta-data descriptions.
Data Sharing	Fully open: https://www.carbontracker.eu
Archiving and preservation (including storage and backup)	Data are stored at the supercomputing system that is used for the CTE runs. A full back-up is made, and results are also copied to https://www.carbontracker.eu DOI is available through the GCP process.

Offline land surface model analysis with VOD	
Data set description	IFS offline land surface model analysis with VOD: volumetric soil moisture for model layers 1,2,3 and vegetation LAI will be provided with and without VOD assimilation. The data will be generated and available at 25km resolution, TCo399. LDAS-Monde offline land surface model analysis with VOD: GPP, volumetric soil moisture for model layers 1,2,3, 4, 5, 6, 7, 8 and vegetation LAI will be provided with and without VOD assimilation. The data will be generated and available at 25km resolution (regular lat/lon grid at 0.25° spatial resolution) on a daily frequency . SIF will be used to assess the added value of the assimilation on GPP. Limitations: preliminary investigation of VOD data assimilation
Standards and metadata	Will you generate proper metadata for your data? Yes, in NetCDF header. Data format? Grib and NetCDF
Data Sharing	Data will be available on MARS, accessible from web API.
Archiving and preservation (including storage and backup)	MARS archive at ECMWF

Offline land surface model simulations	
Data set description	IFS offline land surface model simulations with CHTESSEL at 25km resolution (TCo399) over period 2015-2021.

Standards and metadata	Will you generate proper metadata for your data? Yes, in NetCDF header. Data format? Grib and NetCDF
Data Sharing	Data will be available on MARS, accessible from web API with DOI.
Archiving and preservation (including storage and backup)	MARS archive at ECMWF

SIF-based GPP (ORCHIDEE CCDAS)	
Data set description	Global spatially resolved (0.5 and 2 degree resolution, 2015-2020) Gross Primary Productivity (GPP) at monthly time scale (possibly higher frequency) derived from ORCHIDEE after the assimilation (i.e., parameters optimisation) of optical Solar Induced fluorescence data (SIF from OCO-2 data over the period 2015-2016).
Standards and metadata	The product is based on a state-of-the-art process-based global terrestrial carbon cycle model (ORCHIDEE) and a data assimilation system, which is highly documented in the peer reviewed literature and through a web site (papers and system description at: https://orchidas.lsce.ipsl.fr/). The model parameters were calibrated against satellite SIF observations (and also in a previous step against FluxNet data). The main description paper with all metadata is: Bacour, C., Maignan, F., MacBean, N., Porcar-Castell, A., Flexas, J., Frankenberg, C., Peylin, P., Chevallier, F., Vuichard, N., and Bastrikov, V.: Improving estimates of Gross Primary Productivity by assimilating solar-induced fluorescence satellite retrievals in a terrestrial biosphere model using a process-based SIF model, J. Geophys. Res. Biogeosci., 124, 11, 3281-3306, doi: 10.1029/2019JG005040 The simulated GPP has been evaluated against independent data and observations (FluxCOM data from Jung et al. 2016 and FluxSat data from Joiner et al. 2018) The data will be available in netCDF format
Data Sharing	The data will be publicly available and it will be accessible through the LSCE sharebox system; and the VERIFY data portal (http://verify.lsce.ipsl.fr/index.php/products).
Archiving and preservation (including storage and backup)	The data will be preserved at the LSCE archiving facilities but further long-term preservation can be envisaged (possibly through ICOS).

SIF-based GPP (BETHY-CCDAS)	
Data set description	Global spatially resolved (~0.5 deg or higher, 2016-2021) Gross Primary Productivity (GPP) derived from CCDAS after assimilation of optical EO (such as SIF/FAPAR) and/or other data.
Standards and metadata	The simulations are based on a state-of-the-art terrestrial carbon cycle data assimilation system which is well documented in the peer reviewed literature. The model parameters are calibrated against EO observations. The metadata will also be documented with a peer-reviewed publication, including the evaluation of the simulated GPP data against independent data and observations. The data will be available in netCDF format
Data Sharing	The data will be publicly available and it will be accessible through the ICOS Carbon Portal.
Archiving and preservation (including storage and backup)	The data will be preserved also at the ICOS Carbon Portal following the FAIR principles.

Nature runs	
Data set description	High resolution CO2, CH4 and CO IFS simulations (9km and 137 vertical levels) will be performed for 2018 and 2021. Full chemistry and aerosol simulations will be performed for shorter periods of a few months. This will be useful for boundary conditions of regional models (CoCO2 WP4) and to perform Observation System Simulation Experiments (OSSEs).These IFS simulations will include the latest developments in the numerical advections scheme for tracers, anthropogenic emissions with temporal and vertical profiles and CMEMS ocean fluxes (from CoCO2 WP2), ESA-CCI land use and CGLS LAI, as well as a new photosynthesis model. Lower resolution simulations (25km, L137) covering a wider range of years (2016-2021) will be performed to test all these developments and new emissions/fluxes. This follows the work done in the CHE project (CHE nature run.
Standards and metadata	The simulations are based on the state-of-the-art NWP model which is very well documented. The model parameters are based on the WMO standard meteorological parameters and described in grib parameter database (https://apps.ecmwf.int/codes/grib/param-db). The metadata will also be documented with a peer-

	reviewed publication, including the evaluation simulations with independent data and observations. The data will be available in grib or netCDF formats.
Data Sharing	The data will be publicly available and it will be accessible through the ECMWF API (https://www.ecmwf.int/en/forecasts/access-forecasts/ecmwf-web-api).
Archiving and preservation (including storage and backup)	The IFS simulations will be archived in the mars tapes at ECMWF and they will be classified as "publication datasets" which means they will be preserved for at least 5 years, and after which the dataset preservation will be reviewed. The volume of the data will depend on the number of parameters archived, each 3D tracer parameter takes approximately 7MB on model levels and 1MB on pressure levels for a full year. There is an established workflow in place to request a DOI. This procedure was already followed to request a DOI for the CHE nature run.

Ensemble Simulations	
Data set description	Medium resolution CO2 and CH4 IFS 50-member ensemble forecast simulations (25km and 137 vertical levels) will be performed for periods of 2018 and 2021. These IFS simulations will include the latest developments in the numerical advections scheme for tracers, anthropogenic emissions with temporal and vertical profiles and CMEMS ocean fluxes (from CoCO2 WP2), ESA-CCI land use and CGLS LAI, as well as a new photosynthesis model. This follows the work done in the CHE project (CHE nature run). These simulations will provide an estimate of total model error in atmospheric concentration, including a representation of emission uncertainty. We also aim to provide ensemble-based 4d-var inversions for short periods, although the details of these simulations are not yet clear.
Standards and metadata	The simulations are based on the state-of-the-art NWP model which is very well documented. The model parameters are based on the WMO standard meteorological parameters and described in grib parameter database (https://apps.ecmwf.int/codes/grib/param-db). The metadata will also be documented with a peer-reviewed publication, including the evaluation simulations with independent data and observations. The data will be available in grib or netCDF formats.
Data Sharing	The data will be publicly available and it will be accessible through the ECMWF API (https://www.ecmwf.int/en/forecasts/access-forecasts/ecmwf-web-api).

Archiving and preservation (including storage and backup)	The IFS simulations will be archived in the mars tapes at ECMWF and they will be classified as “publication datasets” which means they will be preserved for at least 5 years, and after which the dataset preservation will be reviewed.

4.3 Work Package 4

Library of plume simulations	
Data set description	Library of atmospheric GHG and pollutant plumes from isolated point sources and cities using simulations at very high resolutions from a few metres to 1 km, over few tens to few hundred km wide areas and for few days (per source). The list of source "test cases" will encompass 4-5 industrial plants and 3 cities. The simulations will include CO2 and co-emitted species NO2 and CO, which will be simulated with full or simplified linear chemistry. All the plumes will be simulated in the frame of T4.1 of WP4. This ensemble of plume simulations will be useful to evaluate model performances and parametrizations, and to produce synthetic data to test local scale inversion techniques (in particular in T4.2 and T4.3 of WP4). It should be documented in deliverables D4.1 and D4.2 of CoCO2 including some content similar to CHE deliverable D2.1 (https://www.che-project.eu/node/124) and a dedicated publication in addition to being referred to in publications exploiting the simulations for specific sources.
Standards and metadata	The product will be available as a collection of netCDF files compliant with CF conventions. The output format will be harmonized as much as possible (names of dimensions, coordinates, variables, units). Metadata information will be included in the global attributes of the netCDF files. Detailed information on the models and simulations and on their evaluation based on comparison to measurements will be given in Deliverables D4.1 and D4.2 of CoCO2.
Data Sharing	The model outputs will first be shared among project partners and then made publicly available together with Deliverable D4.2 of the project by January 2023.
Archiving and preservation (including storage and backup)	Unknown at this stage of the project.

Ensembles of local source and plume estimates	
Data set description	Ensembles of emission estimates and concentration fields for the various local scale inversions in T4.2 (providing emission estimates only) and T4.3 (providing both emission and concentration fields).

	The ensembles encompass (i) a subset of the source "test cases" in T4.1 and of the corresponding days of simulations, (ii) results based on various sets of either pseudo-observations or real observations. For each inversion case, the ensemble gathers samples of the prior (before inversion) and posterior (inverted) distribution of the statistical emission and concentrations estimates for CO2 and/or co-emitted species: - samples of the estimates of the source emissions - samples of the corresponding plume simulations - the pseudo or real observations assimilated by the inversion - samples of the simulated concentrations corresponding to the pseudo or real observation These ensembles will characterize the performance of different inversion techniques and observation systems tested in T4.2 and T4.3 of WP4 and will be taken as example of products that can be assimilated in the multiscale prototype developed in WP6. They correspond to part of D4.7.
Standards and metadata	The product will be available as a collection of netCDF files compliant with CF conventions. The output format will be harmonized as much as possible (names of dimensions, coordinates, variables, units). Metadata information will be included in the global attributes of the netcdf files.
Data Sharing	These ensembles will be shared as soon as possible with WP6 even if potentially updated later, and then finalized and made publicly available by July 2023.
Archiving and preservation (including storage and backup)	Unknown at this stage of the project.

Ensembles of national scale emission and concentration estimates	
Data set description	Ensembles of emission estimates and concentration fields for the various national scale inversions in T4.4. The ensembles encompass (i) the various targeted countries in T4.4 and the various inverse modeling system covering each country but probably temporal windows that are limited compared to the full extent of these inversions (ii) results based on various set of either pseudo-observations or real observations. For each inversion case, the ensemble gathers samples of the prior (before inversion) and posterior (inverted) distribution of statistical emission and concentrations estimates for CO2 and/or co-emitted species:

	- samples of the 3D (2D in space, 1D in time) maps of emissions - samples of the corresponding 4D concentration fields (3D in space, 1D in time) - the pseudo or real observations assimilated by the inversion - samples of the simulated concentrations corresponding to the pseudo or real observation These ensembles will characterize the performance of different inversion techniques and observation systems tested in T4.4 of WP4 and will be taken as example of products that can be assimilated in the multiscale prototype developed in WP6. They correspond to part of D4.7.
Standards and metadata	The product will be available as a collection of netCDF files compliant with CF conventions. The output format will be harmonized as much as possible (names of dimensions, coordinates, variables, units). Metadata information will be included in the global attributes of the netcdf files.
Data Sharing	These ensembles will be shared as soon as possible with WP6 even if potentially updated later, and then finalized and made publicly available by July 2023.
Archiving and preservation (including storage and backup)	Unknown at this stage of the project.

4.4 Work Package 5

EC data for benchmarking	
Data set description	Carbon fluxes and meteorological data from compiled eddy covariance datasets to be used for benchmarking land surface models and quantifying errors in biogenic CO2 fluxes. Data will come primarily from FLUXNET (https://fluxnet.org) and other regional products as available. Data will be subject to further data screening and quality checks. Current limitations are due to data availability and the processing chain, particularly in tropical and arid regions.
Standards and metadata	We currently plan to follow the standards outlined in the Plumber with the potential to integrate into the existing infrastructure at modelevaluation.org. Data will be in NetCDF format.
Data Sharing	The existing eddy covariance data are under a CC-BY-4.0 licence and any derived dataset would be under a similar license. We are exploring the possibility of distributing the data via the ICOS Carbon Portal and/or via modelevaluation.org.

Archiving and preservation (including storage and backup)	Data size is relatively small (on the order of GBs).

CO2/CH4 inversions computed with the Community Inversion Framework (CIF)	
Data set description	An inversion inter-comparison over Europe will be computed within WP5 using the CIF. The standardized inversion outputs and inputs (observations used in the inversion and to evaluate the inversion, and corresponding simulations before and after inversion, prior emissions and posterior emissions) will be generated using different inversion set-ups and transport models. This dataset will be useful in other WP and tasks to better set-up inversions at different scales
Standards and metadata	The data generated by the CIF are classical NetCDF files. Information about the set-up used to generate them will be attached to each individual files as NetCDF file attributes.
Data Sharing	The data will be shared among the project partners. The data will be publicly shared later through a dedicated scientific publication describing results from the inversion inter-comparison exercise.
Archiving and preservation (including storage and backup)	The data will be archived on CEA archiving system for long-term storage for at least 10 years.

CH4 inversion inter-comparison	
Data set description	Model submissions generated by the participating groups, including methane fluxes and model simulated methane mixing ratios at measurements locations that are optimized or used for evaluation. The dataset may have to be updated, in case any problems are encountered during the use of it and since the activity is expected to extend beyond the duration of the COCO2 project. Output from external partners is promised but not guaranteed. Since the ownership lies with those partners, we have to ask their consent on where and how to store their data (which is usually not a problem in inter-comparisons).
Standards and metadata	We follow the example of previous model inter-comparisons (e.g. EuroCom), which will be needed also for visualization of data using software from the ICOS carbon portal. The data will come with metadata as well as documentation of how the results were generated. The data will be in netcdf format.

Data Sharing	Data will be shared among participants using a Dutch service called research drive. After data have been published, they will be made available most likely via the ICOS Carbon Portal and/or the VUA iRODS/Yoda RDM server (with the possibility to assign a DOI).
Archiving and preservation (including storage and backup)	Data are archived in iRODS/Yoda for a default period of 10 years. This service is offered by the Vrije Universiteit Amsterdam without costs for researchers using it. Is there an established workflow for your requested DOI process in place? Yes, this is taken care of by the VUA library.

4.5 Work Package 6

Extension of the VERIFY synthesis	
Data set description	During the course of the VERIFY project, datasets from 40 partners across Europe have been collected, primarily LSCE, University of Aberdeen, IIASA, JRC, TNO, University of East Anglia, University of Munich, University of Brennan, MPI, and Wageningen Environmental Research, as well as publicly-available data from the UNFCCC, FAOSTAT, the TRENDY project, and the Global Carbon Project. All of this underpins the VERIFY synthesis papers, published in ESSD: https://doi.org/10.5194/essd-2020-376 and https://doi.org/10.5194/essd-2020-367 These data are made freely available on the VERIFY database one year after being received by the project, though they are available to partners as soon as they are received. The dataset will be continued for at least one year after the end of VERIFY within CoCO2, and will be updated (extending the temporal coverage).
Standards and metadata	Metadata is included in the name of the file on the database: Method/object _ species _ Variable_ Simulation type_ Institute _ sector _ Region-extent _ TimeStep _ Version _ Time-stamp _ author _ WP.nc The data format is NetCDF files, and outline more in the Milestone 13 of the VERIFY project. We will not generate additional metadata. All metadata is searchable from the Products page of the VERIFY website.

	Data has been subjected to some basic quality checks through the data harmonization procedure, checking for units and to make sure data is present. Additional checks were done in the creation of the synthesis papers for selected datasets by comparing them.
Data Sharing	Data is currently made freely available through the VERIFY website: http://webportals.ipsl.jussieu.fr/VERIFY/CountryTot.html For the raw data, users need to register to get a generic username and password. The database is maintained on the computer network infrastructure of the LSCE for the current time, though discussions are underway to transfer it to another system (for long term storage), in particular for the CoCO2 extension.
Archiving and preservation (including storage and backup)	Discussions are underway to transfer/archive the data to another system (for long term storage and access), in particular for the CoCO2 extension.

Inversion of global carbon dioxide and methane emissions	
Data set description	Inversion of global carbon dioxide and methane emissions associated with anthropogenic and biogenic processes using the IFS 4D-Var system. Emission estimates will be provided per country and per month for 2018 and 2021 together with their posterior uncertainties. Evaluation of the product through in situ comparisons with atmospheric measurements (TCCON network) will also be made available. The posterior emission dataset will provide a tool to quantify the trend in anthropogenic emissions from 2018 to 2021 per country.
Standards and metadata	The dataset will be available in both the netcdf and grib formats. Metadata are provided as part of the interface of the MARS server where the dataset will be stored.
Data Sharing	The emission estimate dataset will be made freely accessible through ECMWF's MARS server located at https://apps.ecmwf.int/mars-catalogue/. The posterior errors and evaluation dataset will be made available on a separate dedicated ftp server set up for the CoCO2 project.
Archiving and preservation (including storage and backup)	Long-term archiving of the dataset will be available on ECMWF's MARS server (https://apps.ecmwf.int/mars-catalogue/) and on a dedicated ftp server set up for the CoCO2 project for the posterior error and evaluation dataset.

4.6 Work Package 7

Krakow dataset from Task 7.5	
Data set description	Data set of one year of observations performed during 2021 in Krakow, Poland including: 1. Regular observations of a. meteorology, b. radiation balance, c. air quality data 2. Measurements collected during campaigns performed on a ca. monthly basis dedicated to the characterization of: a. the diurnal variation of vertical CO2 profiles within the urban boundary layer b. the anthropogenic CO2 contribution to the total emissions based on the carbon isotopic analysis using spot flask samples. The dataset is intended for high-resolution modelling systems to capture the spatiotemporal variability in CO2 concentrations and flux distributions in an urban environment and thus better discriminate between anthropogenic and biogenic CO2 fluxes.
Standards and metadata	Yes, we will provide meta-data. Metadata will be an integral part of the dataset provided as part of the D7.6 deliverable The data will be in txt format Will there be a review process to quality-check the data? yes, users of the data will be asked to provide feedback.
Data Sharing	Data will be initially be made available to partners via ftp server, in any upcoming project data repositories or existing data repositories like the ICOS-carbon portal
Archiving and preservation (including storage and backup)	The dataset will be stored on external memory devices and cloud systems at the AGH University of Science and Technology (dysk.agh.edu.pl) as well as at any future project data repository.

IoT system data from Task 7.5	
Data set description	Data originated by the prototypes of IoT sensors to be used in urban environment. A limitation is high uncertainty in the measurement, which will be documented. The data will come from different sensors in different urban contexts.
Standards and metadata	The data will follow the data structure and metadata generally used for concentration and meteo data in ICOS Will you generate proper metadata for your data? If yes: how do they look like? Similar to the ICOS metadata, adapted to the specific dataset

	Data format? To be decided, based on user requests Will there be a review process to quality- check the data? By the users in the projects
Data Sharing	Data will be shared in the ICOS Carbon Portal under the CC-BY license of ICOS.
Archiving and preservation (including storage and backup)	The ICOS Carbon Portal will ensure archiving, preservation and DOI.

Heraklion flux tower data from Task 7.5	
Data set description	A complete set of meteorological and micrometeorological variables derived by 2 Eddy Covariance (EC) stations in the urban area of Heraklion, Greece. The datasets include the following variables: • Air temperature • Sonic temperature (high frequency) • Relative humidity • Wind speed (3D, high frequency) • CO_2, H_2O density (high frequency) • Wind direction (high frequency) • Barometric pressure • Net radiation components (R_n, L↑, L↓, K↑, K↓) The high frequency (10 Hz) measurements of the EC stations are processed to provide atmospheric turbulence and stability characteristics, as well as heat and gas fluxes at an aggregated time-scale (e.g. 30 min): • CO_2 flux • Sensible heat flux • Latent heat flux • Evaporation • Momentum flux • Friction velocity • Turbulent Kinetic Energy • Monin-Obukhov length • Monin-Obukhov stability parameter • Quality flags
Standards and metadata	Datasets are stored in table format (csv, xlsx)[3] and will be accompanied with appropriate technical documents that would describe in detail the site characteristics, data acquisition attributes and methodologies. All text documents will be in pdf format[4]. The methodology standards for urban Eddy Covariance installation and flux processing follow the standards and

[3] https://www.ecma-international.org/publications/standards/Ecma-376.htm
[4] https://www.iso.org/standard/63534.html

	recommendations of Eddy Covariance applications in urban environments by Feigenwinter et al. (2012)[5]. Data readings are associated with date and time stamp (UTC +2) and appropriate metadata that would include: • Unique code per station • Geographic location (geographic coordinates in WGS84) • Altitude above sea level • Unique code per variable • Variable description • Units of variable • Temporal resolution • Sensor information (model, type, etc) • Site characteristics (to be defined) • Height of the sensor above ground (m) • Data format
Data Sharing	The mechanism for data sharing has not yet been established.
Archiving and preservation (including storage and backup)	The procedures for long-term storage of the data have not yet been established.

5 Conclusion

This initial Data Management Plan has identified a number of data sets for each of the work packages 2 to 7, identifying the required details (where possible) on what data will be open, how it will be made accessible, and how it will be curated.

The Data Management Plan is to be seen as a living document and will be reviewed and revised periodically to ensure that information contained therein is up-to-date and correct.

[5] Feigenwinter, C., Vogt, R. & Christen, A. Eddy Covariance Measurements Over Urban Areas - Eddy Covariance: A Practical Guide to Measurement and Data Analysis. in (eds. Aubinet, M., Vesala, T. & Papale, D.) 377–397 (Springer Netherlands, 2012). doi:10.1007/978-94-007-2351-1_16

Document History

Version	Author(s)	Date	Changes
0.1	Daniel Thiemert (ECMWF)	18/06/2021	initial version for review
1.0	Daniel Thiemert (ECMWF)	30/06/2021	Final Version after review

Internal Review History

Internal Reviewers	Date	Comments
Mirosław Zimnoch (AGH)	29/06/2021	approved with comments

Estimated Effort Contribution per Partner

Partner	Effort
ECMWF	0.1
Total	**0.1**