CO2MVS RESEARCH ON SUPPLEMENTARY OBSERVATIONS

D5.5 Data Management Plan

Due date of deliverable	30/06/2023
Submission date	
File Name	D5.5 Data Management Plan
Work Package /Task	WP5/ T5.3
Organisation Responsible of Deliverable	ECMWF
Author name(s)	Tanya Warnaars, Richard Engelen, Rhona Phipps and WP partners
Revision number	1
Status	final
Dissemination Level	Public

The CORSO project (grant agreement No 101082194) is funded by the European Union.

1 Executive Summary

The CORSO Data Management Plan describes the specifications for data, quality control, metadata generation, data access, data stewardship and how data will be maintained and preserved. The types of data that will be used or produced in the project are satellite and in-situ observations, prior emissions, and results from inversion studies. The data of the project will comply with the FAIR data principles, adhering to the principle 'as open as possible and as closed as necessary'[1]. The data will be accessible using existing data portals, such as the Copernicus Atmosphere data Store, the ICOS Carbon Portal, and the Centre for Environmental Data Analysis (CEDA) archive.

This document is a living document which will be developed during the lifetime of the project to follow and share the developments of the CORSO project.

[1] European Commission, Directorate-General for Research and Innovation, *Horizon Europe, open science – Early knowledge and data sharing, and open collaboration*, Publications Office of the European Union, 2021, **https://data.europa.eu/doi/10.2777/18252**

Table of Contents

2 Introduction

The following provides the plans for how the project will set up, administer and archive the legacy of data arising from CORSO. This deliverable aims at supporting partners' in their efforts and responsibilities in making project data that is FAIR (Findable, Accessible, Interoperable, Reusable) and 'as open as possible, as closed as necessary'. It will also ensure consistency across the project.

This deliverable is primarily targeted at the consortium partners and should serve as a reference for the management of data products in the relevant deliverables. It also serves to support the cross-cutting activity on data integration and data products, which will interact with all WPs throughout the duration of the project to maximize benefits of the data generated by CORSO.

This CORSO data management plan describes the data management life cycle for all datasets to be collected, processed and generated in the project. It constitutes the first version of the DMP and provides the baseline of the policy that will be followed by the CORSO consortium with respect to the data management related activities. More specifically, it covers the following activities:

- What types of data will be collected and/or generated?
- What standards will be used?
- How will this data be exploited, shared, processed and made accessible?
- How will this data be curated, stored and preserved?
- Which tools and methodologies will be used to store this data and for how long?
- How are data restriction levels managed?

This DMP outlines how research data will be handled throughout the life cycle of the project.

2.1 Background

To enable the European Union (EU) to move towards a low-carbon economy and implement its commitments under the Paris Agreement, a binding target was set to cut emissions in the EU by at least 40% below 1990 levels by 2030. European Commission (EC) President von der Leyen committed to deepen this target to at least 55% reduction by 2030. This was further consolidated with the release of the Commission's European Green Deal on the 11th of December 2019, setting the targets for the European environment, economy, and society to reach zero net emissions of greenhouse gases in 2050, outlining all needed technological and societal transformations that are aiming at combining prosperity and sustainability. To support EU countries in achieving the targets, the EU and European Commission (EC) recognised the need for an objective way to monitor anthropogenic CO_2 emissions and their evolution over time.

Such a monitoring capacity will deliver consistent and reliable information to support informed policy- and decision-making processes, both at national and European level. To maintain independence in this domain, it is seen as critical that the EU establishes an observation-based operational anthropogenic CO_2 emissions Monitoring and Verification Support (MVS) (CO2MVS) capacity as part of its Copernicus Earth Observation programme.

The CORSO research and innovation project will build on and complement the work of previous projects such as CHE (the CO2 Human Emissions), and CoCO2 (Copernicus CO2 service) projects, both led by ECMWF. These projects have already started the ramping-up of the CO2MVS prototype systems, so it can be implemented within the Copernicus Atmosphere Monitoring Service (CAMS) with the aim to be operational by 2026. The CORSO project will further support establishing the new CO2MVS addressing specific research & development questions.

The main objectives of CORSO are to deliver further research activities and outcomes with a focus on the use of supplementary observations, i.e., of co-emitted species as well as the use of auxiliary observations to better separate fossil fuel emissions from the other sources of atmospheric CO_2. CORSO will deliver improved estimates of emission factors/ratios and their uncertainties as well as the capabilities at global and local scale to optimally use observations of co-emitted species to better estimate anthropogenic CO_2 emissions. CORSO will also provide clear recommendations to CAMS, ICOS, and WMO about the potential added-value of high-temporal resolution $^{14}CO_2$ and APO observations as tracers for anthropogenic emissions in both global and regional scale inversions and develop coupled land-atmosphere data assimilation in the global CO2MVS system constraining carbon cycle variables with satellite observations of soil moisture, LAI, SIF, and Biomass. Finally, CORSO will provide specific recommendations for the topics above for the operational implementation of the CO2MVS within the Copernicus programme.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverables

This D5.5 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

As per the DoA, D5.5 the work performed includes the collection of the available descriptions of data sets to be produced by the project, through a questionnaire (see Annex i).

2.2.3 Deviations and counter measures

No deviations have been encountered.

2.2.4 Reference Documents

[1] [RD 1] 101082194-CORSO-HORIZON-CL4-2022-SPACE-01 Description of the Action

[2] European Commission, Directorate-General for Research and Innovation, *Horizon Europe, open science – Early knowledge and data sharing, and open collaboration*, Publications Office of the European Union, 2021, https://data.europa.eu/doi/10.2777/18252

2.1 Project partners:

Partners	
EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS	ECMWF
AKADEMIA GORNICZO-HUTNICZA IM. STANISLAWA STASZICA W KRAKOWIE	AGH
BARCELONA SUPERCOMPUTING CENTER - CENTRO NACIONAL DE SUPERCOMPUTACION	BSC
COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES	CEA
KAMINSKI THOMAS HERBERT	iLab
METEO-FRANCE	MF
NEDERLANDSE ORGANISATIE VOOR TOEGEPAST NATUURWETENSCHAPPELIJK ONDERZOEK TNO	TNO
RIJKSUNIVERSITEIT GRONINGEN	RUG
RUPRECHT-KARLS-UNIVERSITAET HEIDELBERG	UHEI
LUNDS UNIVERSITET	ULUND
UNIVERSITE PAUL SABATIER TOULOUSE III	UT3-CNRS
WAGENINGEN UNIVERSITY	WU
EIDGENOSSISCHE MATERIALPRUFUNGS- UND FORSCHUNGSANSTALT	EMPA
EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZUERICH	ETHZ
UNIVERSITY OF BRISTOL	UNIVBRIS
THE UNIVERSITY OF EDINBURGH	UEDIN

3 Data Summary

Our Data Management Plan (DMP) is developed following the standard approach to the European Monitoring and Evaluation Programme (EMP) whereby it sets out the specifications for data, quality control, metadata generation, data access, data stewardship and how data will be maintained and preserved. It is developed to provide guidelines to adhere to article 17 to the Grant Agreement. As with scientific peer-reviewed publications, datasets generated by the project will be deposited in repositories and made Open Access. Data will be made freely available for use where possible. To facilitate the exploitation and monitoring of the Data Management Plan a specific Task 5.4 (WP5) is responsible for this activity.

CORSO is benefiting from previous H2020 projects, and has identified *Prior information* as a building block. CORSO will use a multitude of existing tools and data sets using the heritage of CAMS and the CHE, CoCO2, and VERIFY projects, contributing to the re-use of project outputs. Indeed, the CORSO project will advance the use of observations of co-emitted species (e.g., NO_2 and CO). CORSO activities are closely coordinated with related activities in the Copernicus Atmosphere Monitoring Service (CAMS) and the CoCO2 project. Whilst prior emission data sets are already available from CAMS, CoCO2, and other projects, CORSO will add value by focusing specifically on the definition of the uncertainties and correlations in emission factors and emission ratios for the relevant pollutants, both in space and time.

The products of CORSO will comprise reports, graphical displays, datasets and improved methods, algorithms and code. All these elements have their own important role. 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.

Data products arising from the project:

- Global maps of CO2, CO and NOx emission factors and their uncertainties
- Improved global point source emission dataset
- Data sets of in-situ observations of $^{14}CO_2$ and APO

The development of the Copernicus CO2MVS capacity is a concerted effort involving the main institutional partners (EC, ECMWF, ESA, and EUMETSAT), observations infrastructures (e.g., ICOS), and the science community. The CORSO consortium will join and strengthen this established collaboration, focusing on specific aspects of the development as outlined in this proposal. There is a strong link to the Copernicus Atmosphere Monitoring Service, which is responsible for the implementation of the CO2MVS, through the CORSO project coordinator ECMWF, which is the Entrusted Entity for CAMS. CORSO will therefore directly interact with relevant activities in CAMS. CORSO will also closely collaborate with the H2020 CoCO2 project (2021 – 2023) using data provided and methods developed by CoCO2 as input to the CORSO studies.

CORSO will generate data of relevance to support emission estimates. For example it will compile a database of existing APO and $^{14}CO_2$ observations to assess the potential of global $^{14}CO_2$ and APO observations to provide a robust continental-scale constraint over relatively long timescales.

CORSO is specifically focused on implementing and testing the use of additional observations from satellites and ground-based networks, such as for co-emitted species, radiocarbon and APO, as well as satellite observations that can constrain the fluxes between vegetation and the atmosphere. Ensuring the availability of these observational datasets and providing guidance for their maintenance and extension are therefore key components of the CORSO project.

For the satellite observations, CORSO will directly link with existing frameworks, such as Copernicus and the ESA CCI datasets, to obtain quality retrievals of CO_2 from the Japanese GOSAT and American Orbiting Carbon Observatory 2 (OCO-2) instruments. CORSO will also use observations of CH_4, NO_2 and CO from the Copernicus Sentinel-5p satellite.

The outputs generated by CORSO can be beneficial to a number of target groups, such as CAMS, WMO, ICOS and other international in-situ infrastructures, Carbon cycle science community and Other national, EU, or international research projects.

3.1 Definitions related to the approach to Open Science:

The Horizon Europe programme guide states[2]: "*Open science is an approach based on open cooperative work and systematic sharing of knowledge and tools as early and widely as possible in the process."* In this regard we clarify for CORSO the vocabulary on open access below:

Open Access Data: Open access refers to unrestricted access to research results. Commonly, the open access characterization is given to open-source peer-reviewed publications, datasets, tools and source code. Open access focuses on building a community and enables scientists, researchers, interest groups and individuals to:

- Build and enhance existing research results

[2] Guidelines on FAIR Data Management in Horizon Europe (Version 2.0, 01 April 2022), https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf

- Avoid redundancy
- Participate in Open Innovation activities
- Benefit from the results of the CORSO project

Open Research Data: Open research data refers to the disclosure of the linked research data which are needed to assess, validate and replicate the results presented in research publications. Complementary to the concept of open access, open research data enables the online availability of data resources towards promoting research.

The open research data concept focuses on enabling researchers and individuals to:

- understand, assess, reconstruct and further expand scientific publications
- build innovative concepts on top of existing research data
- establish a continuous improvement mechanism of research

3.2 Approach

The general strategy for data management sets out the specifications for data, quality control, metadata generation, data access, data stewardship and how data will be maintained and preserved. The types of data that will be used or produced in the project are satellite and in-situ[3] observations, prior emissions, and results from inversion studies.

CORSO has a strong link to the Copernicus Atmosphere Monitoring Service. The close collaboration will ensure that the CORSO activities are complementary to what is done elsewhere, like in the project CoCO2. In addition, CORSO will interact with the H2020 ICOS-Cities project, where relevant.

4 FAIR Data

The data of the project will comply with the FAIR data principles, as much as possible. The data will be accessible using existing data portals, such as the Copernicus Atmosphere data Store, the ICOS Carbon Portal, and the Centre for Environmental Data Analysis (CEDA) archive. All these data portals have been designed to support interoperability and include clear licensing information as well as tools to make best use of the data.

Each participating organization will examine whether open access can be granted without affecting any legal and ethical requirements, including the Intellectual Property Rights as per the dissemination access level of each dataset produced.

This DMP follows the EU guidelines[1] and describes the data management procedures according to the FAIR principles[4]. The acronym FAIR identifies the main features that the project research data must have in order be findable, accessible, interoperable and

[3] In the current EU Space Regulation, in-situ observations are defined as follows: ‘Copernicus in-situ data’ means observation data from ground-based, seaborne or airborne sensors, as well as reference and ancillary data licensed or provided for use in Copernicus

[4] The FAIR data principles (GO FAIR), https://www.go-fair.org/fair-principles

4.1 Making data findable, including provisions for metadata

Importance is placed on enhancing the discoverability of the collected and generated data. Metadata links information and data across the web and constitutes a powerful tool that helps individuals (researchers, developers, citizens, etc.) to discover, identify, and manage digital resources. Metadata refers to information about the data collected and/or generated. It is usually structured as textual information that describes the creation, content, or context of a digital resource. The most notably known types of metadata are names, dates, location, data types, relations and interdependencies to other data sets.

Datasets that will be uploaded to open access repositories will be deposited in a searchable resource and listed on our project website. The naming conventions for the project's data files can significantly increase their searchability. Towards this, CORSO will design consistent data file names that properly describe their content, status and versioning, with a view on increasing their discoverability.

During the course of the project, and at least at the moment of publication of the project results, each research team will deposit and describe the relative underlying data sets. Trusted data repositories can attribute persistent unique identifiers (PIDs) to the deposited items (e.g. Zenodo, Copernicus Atmosphere Data Store, the ICOS Carbon Portal, and the Centre for Environmental Data Analysis (CEDA) archive).

4.2 Making data accessible

FAIR open access to the data guide refers to making data accessible to all project partners, researchers and the public, following the privacy and anonymity guidelines of the EU and National regulations. Accessibility for the Horizon Europe, which states that all data generated and used, if possible, are publicly open and available. The CORSO partnership will ensure the integrity of personal data and sensitive information prior to the dissemination of the datasets.

The project does not aim to replicate any data and will maintain a list of data sets it accesses for the purposes of CORSO activities on the project website. The accessibility of the data will be ensured at two levels: internally to the project, and to the general public. The strong connection to the CAMS community strengthens the use and accessibility of CORSO outputs.

During the execution of the project, each partner will provide detailed information on privacy/confidentiality and the procedures that will be implemented for data collection, storage, access, sharing policies (especially when third party countries are concerned), protection, retention and destruction. The consortium will confirm that the project complies with national and EU legislation throughout its lifetime and after its completion.

As a guiding principle, CORSO seeks to ensure open access to research data, via repositories, as soon as possible and within the limits and deadlines set out in the DMP, in order to allow dissemination, validation and re-use of research results. During the project, trusted repositories will be chosen such as Zenodo, Copernicus Atmosphere data Store, the ICOS Carbon Portal, and the Centre for Environmental Data Analysis (CEDA) archive. The project data sets will be visible via the OpenAIRE portal, facilitating project reporting procedures. Data deposition in repositories will guarantee long time preservation and accessibility to datasets.

Restrictions to access are applied only in the following cases:

- when collected data belongs to third party which have denied permission for sharing them;

- on account of confidentiality and proprietary issues;
- protection of personal data of subjects involved in the research
- when availability of the data would mean that the project's main aim might not be achieved.

For data that falls under some of the restrictions described above and for which it is not possible to take any action to make them shareable, EU allows complete closure or restricted access to them.

CORSO DMP indicates the versions or parts of the data sets that can(not) be freely shared providing the specific details in Annex II. The specific repositories for data set publication and preservation will be further expanded during the project.

4.3 Making data interoperable

Data interoperability refers to the ability of systems and services to access readable and editable data, in terms of their content, context and meaning. To achieve it, CORSO will incorporate suitable standards and vocabularies for data and metadata creation. However in the case of CORSO, the primary end user of the data is the CAMS community. To this end the level of integration to those existing services is a driver for the project as CORSO products need to be interoperable with the applications and workflows of the CAMS / CO2MVS services.

To allow data exchange and re-use among researchers, institutions, organisations, countries, etc., partners will make them available in well-known and documented open formats, as much as possible compliant with available (open) applications.

4.4 Increase data re-use

The GO FAIR principles state “FAIR is to optimise the reuse of data”. Data availability after the end of the project depends highly on the type and content of data, taking into account sensitivity and specific licences. Data should be available for public reusability after being granted permission from their respective contributors, following the proposed legal and ethics requirements.

Rich metadata will enable proper discovery and identification of the data along with the appropriate licensing schemes facilitating their re-usability. In principle, it is expected that data will become available after the publication of the respective deliverables and will remain available after the completion of the project.

To safeguard the transparency, consistency, quality, completeness and accuracy of the data, CORSO adopts a data quality assurance procedure. Peer-reviews of the data generation methods and/or data summaries are inherent in the work of the project and will be applied to assess the quality of the dataset and identify any need for improvement.

5 Other research outputs

Other research data will be stored and backed up regularly through existing back-up mechanisms in place at Sharepoint and the internal Confluence pages. This is particularly

relevant to project documents, reports, internal data sharing between consortium partners and web content.

6 Allocation of resources

The resources required for making the data generated by CORSO FAIR have been included in the budget of the project. In general, the CORSO consortium as a whole will decide and contribute to relevant aspects of the data management cycle during and after the completion of this project. A specific table summarising the research team leaders responsible for each dataset will be added in the future release of the DMP.

At this state, the chosen repository for long term deposit and preservation of searchable data intended for public use, does not apply fees for archiving and data curation. Peer-reviewed publications costs related to open-access research data are eligible in Horizon Europe and will be covered by the CORSO budget.

7 Data security

The CORSO consortia place a strong emphasis on ensuring the security of all the produced datasets, safeguarding them from unauthorized access and loss. All the information will be stored in a private and secure storage area. The data will be backed up on a regular basis and access will be restricted only to the members of the consortium. In case of personal data collections, it is crucial that this data can only be accessible by those authorized to do so. To make the data publicly accessible in dedicated public repositories, storage environments will investigate in depth options such as Zenodo, CADS, CEDA, ICOS etc..For what concerns ECMWF, a robust and rigorous data security system is available, including backups. The physical security includes 24/7 monitoring, fire suppress and power backup systems.

All the relevant personal protection protocols, such as GDPR, ECMWF's Personally Identifiable Information Protection and relevant national legislation, will be applied on information of an individual and any reference to personal data or sensitive information will be fully masked in any printed materials, project reports or dissemination activities. Personal data, such as personal information from project partners members, will be treated confidentially, taking into consideration all the proper technical means. General and personal data will be stored separately. All personal data not needed for the final repot, will be destroyed at the end of the project and retained after the completion of the final report.

8 Ethics

All details about ethics and legal compliance in terms of current EU legislative initiatives have been considered and are not of relevance at this point for the data arising from CORSO. Additionally, the Grant Agreement and the CORSO Consortium Agreement are to be referred to for further details on the ownership and management of intellectual property and access.

No ethics or legal issues are foreseen in the project apart from the respect of the GDPR rules when gathering the personal information

9 Conclusion

In this deliverable, the CORSO Data Management Plan has been initiated.

Whilst this provides a good starting point for the FAIR data activities of the CORSO project, it nevertheless needs careful further reflection and updating when appropriate to ensure that new developments (technical as well as strategy) within the CORSO project and beyond are well reflected by the Data Management Plan. The CORSO Consortium will ensure that all generated datasets do not infringe either partner IPR rules or regulations related to personal data protection.

10 ANNEX I

Annex I includes the text of the questionnaire that was shared with each WorkPackage to gather, in table format, the data sets of CORSO by WPs. The table below shows what was asked in order to describe if data:

- is available, or
- will be generated, or
- will be collected

Workpackage X

<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 you 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*

11 ANNEX II

Annex II includes an extensive list of the datasets, already available or to be developed in the context of the project's research and implementation activities. The list is defined for each workpackage of CORSO. The table below shows each data set that:

- is available, or
- will be generated, or
- will be collected

(Note that this is a living document and the information included here may be subject to change throughout the lifetime of the project).

Work Package 1:

Completed by Mark Guevara and Thierno Doumbia with input from WP partners:

Data Set	**Global maps of CO_2, CO and NO_x emission factors**
Data set description	The dataset is a compilation of emission factors per country/region for a few sectors (transportation and later for energy/residential). This dataset will be used to estimate the uncertainty in emission factors as well as in emissions. All of this data will be included in a web-based tool. The developed online system will be published.
Standards and metadata	The emission factor data for each country will be provided as an CSV file as well as a gridded global file. The resulting uncertainties on anthropogenic emissions and the associated metadata will also be made available in the same file formats.
Data Sharing	The emission factor dataset will be shared with the consortium. It might not be possible to share the full dataset with the general public, as some of the data we were able to obtain are access-restricted. The dataset of uncertainties that we will calculate will be shared with the consortium throughout the project and with the general public after CORSO is completed.
Archiving and preservation (including storage and backup)	All the datasets developed in CORSO will be archived as part of the ECCAD (Emissions of atmospheric Compounds and Compilation of Ancillary Data) database. ECCAD does not yet support the easy display of non-gridded data, but this feature might be available before the end of the project.

Data Set	**Global point source emission dataset**
Data set description	The database will consist on a global catalogue of CO2 emissions and co-emitted species (i.e., NOx, SOx, CO, CH4) from power plants and other industries (e.g., cement plants) at high spatial (exact geographic location) and temporal resolution (up to the hourly level) for the year 2021. To construct the

	catalogue, several public and commercial databases will be collected and combined, including among others: - The European Pollutant and Transfer Register database - The Large Combustion Plants database - The integrated Industrial Reporting Database - The Global Coal and Gas Plant Tracker databases - The IEA World Energy Balances Users are potentially atmospheric chemistry modellers and emission inventory developers. The dataset will build on the existing CoCO2 global power plant database, as described in Guevara et al., (2023)[5] Limitations: small and medium size industrial facilities may be missing in the catalogue due to lack of information. Constraints: the quality of the final product is determined by the quality of the data sources and emission factors considered.
Standards and metadata	Reference to the CoCO2 point source database, and documentation of the original database. A proper report and metadata file will be produced. The metadata file will consist on a TXT README file provided together with the final datasets, which will include a detailed description of the different fields of information. Data format: Catalogue of point sources in CSV format for one single year There will be a review process to quality check the data.
Data Sharing	The final dataset is expected to be distributed through a dedicated Copernicus CAMS website repository, as done with the previous CoCO2 point source database (https://atmosphere.copernicus.eu/node/865). In the short term, it will be shared among CORSO partners through a public FTP provided by the BSC. A specific DOI will be assigned to the dataset. No software or tools will be needed to enable its re-use. License: The dataset is expected to be licensed under CC BY 4.0 license.
Archiving and preservation (including storage and backup)	The dataset will be archived in the dedicated Copernicus CAMS website repository mentioned above, providing also a DOI that will be assigned by ECMWF. The approximate volume of the final datasets is expected to be between 10-15Mb.

[5] Guevara, M., Enciso, S., Tena, C., Jorba, O., Dellaert, S., Denier van der Gon, H., and Pérez García-Pando, C.: A global catalogue of CO2 emissions and co-emitted species from power plants at a very high spatial and temporal resolution, Earth Syst. Sci. Data Discuss, https://doi.org/10.5194/essd-2023-95, in review, 2023.

Data Set	Emission factors
Data set description	The dataset is a compilation of emission factors per country/region for a few sectors (transportation and later for energy/residential). This dataset will be used to estimate the uncertainty in emission factors as well as in emissions. All of this data will be included in a web-based tool. The developed online system will be published.
Standards and metadata	The emission factor data for each country will be provided as an csv file as well as a gridded global file. The resulting uncertainties on anthropogenic emissions and the associated metadata will also be made available in the same file formats.
Data Sharing	The emission factor dataset will be shared with the consortium. It might not be possible to share the full dataset with the general public, as some of the data we were able to obtain are access-restricted. The dataset of uncertainties that we will calculate will be shared with the consortium throughout the project and with the general public after CORSO is completed.
Archiving and preservation (including storage and backup)	All the datasets developed in CORSO will be archived as part of the ECCAD (Emissions of atmospheric Compounds and Compilation of Ancillary Data) database. ECCAD does not yet support the easy display of non-gridded data, but this feature might be available before the end of the project.

Work Package 2:

Completed by Gerrit Kuhlmann with input from WP partners:

Data Set	**List of NO_2 hot spots**
Data set description	The dataset will consists of a list of the locations of NO_2 hot spots (cities, power plants and industrial facilities) in Europe, Africa and SE Asia derived from TROPOMI and GEMS (only SE Asia) satellite observations.
Standards and metadata	Metadata: A report and metadata file will be produced. The metadata file will consist on a README file provided with the datasets, which will include a detailed description of the different fields of information. Data format: CSV or netCDF There will be a review process to quality check the data.
Data Sharing	The final dataset is expected to be distributed through a dedicated Copernicus CAMS website repository. A specific DOI will be assigned to the dataset. No software or tools will be needed to enable its re-use.

Data Set	**Time series of NO_x and CO emissions of hot spots**
Data set description	Time series of NO_x and CO emission estimates for hot spots in Europe, Africa and SE Asia from TROPOMI and GEMS NO_2 and CO observations (year: 2021).
Standards and metadata	Metadata: The dataset will be available as a collection of netCDF files complaint with CF convention. A proper report and metadata file will be produced. The metadata file will consist on a README file provided with the datasets, which will include a detailed description of the different fields of information. Data format: netCDF There will be a review process to quality check the data.
Data Sharing	The final dataset is expected to be distributed through a dedicated Copernicus CAMS website repository. A specific DOI will be assigned to the dataset. No software or tools will be needed to enable its re-use.

Data set	**Optimized B matrix parameters (i.e., temporal, spatial, cross-species correlations)**
Data set description	Maps of monthly temporal and spatial error correlation length scales, and cross-species error correlations for NO_x, CO and CO_2
Standards and metadata	Metadata: A report and metadata file will be produced. The metadata file will consist on a README file provided with the datasets, which will include a detailed description of the different fields of information. Data format: netCDF

	There will be a review process to quality check the data.
Data Sharing	The final dataset is expected to be distributed through a dedicated Copernicus CAMS website repository. A specific DOI will be assigned to the dataset. No software or tools will be needed to enable its re-use.

Data set	**Multi-scale global IFS inversion outputs (2021) with assimilated posterior emissions from hotspots.**
Data set description	Maps of global monthly posterior CO2 emissions including assimilation of hotspot posterior estimates for year 2021.
Standards and metadata	Metadata: A report and metadata file will be produced. The metadata file will consist on a README file provided with the datasets, which will include a detailed description of the different fields of information. Data format: netCDF There will be a review process to quality check the data.
Data Sharing	The final dataset is expected to be distributed through a dedicated Copernicus CAMS website repository. A specific DOI will be assigned to the dataset. No software or tools will be needed to enable its re-use.

Work Package 3:

Completed by Ingrid Luijkx and Gregoire Broquet with input from WP partners

Data set	**Database of existing 14CO2 measurements**
Data set description	Collection of existing 14CO2 measurements from a global set of background stations. This database will be made available for use by the modelling components in the work package. This includes flask samples as well as integrated samples. This dataset corresponds to D3.1.
Standards and metadata	The data will be presented with proper metadata, following the procedures as in e.g. the ObsPack data sets, including data on the laboratories and methods to analyse the samples. The format will be text files and netcdf, following ObsPack standards.
Data Sharing	We aim to share the dataset through a service as the ICOS Carbon Portal, using the same license CC BY 4.0 (although still needs to be discussed)
Archiving and preservation (including storage and backup)	This is also foreseen through the ICOS Carbon Portal. A full DOI system is in place.

Data set	**Database of existing APO measurements**
Data set description	Collection of existing O2 measurements from a global set of stations. This database will be made available for use by the modelling components in the work package. This includes flask samples as well as continuous observations. The data will be included on the Scripps O2 scale. This dataset corresponds to D3.2.
Standards and metadata	The data will be presented with proper metadata, following the procedures as in e.g. the ObsPack data sets, including data on the laboratories and methods to analyse the samples. The format will be netcdf as well as text files, following ObsPack standards.
Data Sharing	We aim to share the dataset through a service as the ICOS Carbon Portal, using the same license CC BY 4.0 (although still needs to be discussed)

Archiving and preservation (including storage and backup)	This is also foreseen through the ICOS Carbon Portal. A full DOI system is in place.

Data set	**14CO2 measurement dataset for the 1-year intensive observations in Western Europe**
Data set description	Δ14CO2 activity concentrations of flask samples in the calendar year 2024. The data are reported in with respect to the international 14C reference material Oxalic Acid 2. At twelve stations in central-western Europe, 14CO2 flask sampling will be performed approximately every third day in the calendar year 2024. Ten of these stations are part of the ICOS Atmosphere class 1 measurement network. These will be supplemented by one station in Poland and one in the United Kingdom. The flasks are integrated over one hour in the afternoon. Comparable data sets exist for all ICOS Atmosphere Class1 stations but with much lower temporal resolution. The new data to be acquired will be integrated directly into the existing ICOS data sets. Additionally, it includes flask samples from the Cabauw station in the Netherlands, to be analysed by University of Groningen, as part of intercomparison activities and for targeted samples. Limitations: Due to delays caused by transport between the sampling station and the laboratories, as well as delays caused by measurement capacity limitations, the data will only be available with a delay of about 6 months. This dataset corresponds to part of D3.3.
Standards and metadata	The data will be published in Obspack format. The metadata for individual sampling and station locations will comply with ICOS standards. The 14CO2 measurements from the ICOS stations are subject to quality control by the ICOS Central Radiocarbon Laboratory. Sampling is validated by the individual station PIs.
Data Sharing	For the ICOS stations the datasets will be shared through a service at the ICOS Carbon Portal, under the CC BY 4.0 license. For the two non ICOS stations we aim also for a provision by the ICOS carbon portal using the same licence.
Archiving and preservation (including storage and backup)	The ICOS carbon portal will provide the long-term storage. A full DOI system is in place.

Data set	APO measurement dataset for the 1-year intensive observations in Western Europe
Data set description	This includes new O2 measurements in Cabauw (the Netherlands), to be started in fall 2024. It will include half-hourly values for O2 and CO2, from different levels in the tower. The data will be made available after calibration, with a delay of maximum 6 months. This dataset corresponds to part of D3.3.
Standards and metadata	The data will be shared in obspack format. The metadata will be similar as for ICOS stations. The scale will be the Scripps O2 scale.
Data Sharing	We aim to share the dataset through a service as the ICOS Carbon Portal, using the same license CC BY 4.0 (although still needs to be discussed).
Archiving and preservation (including storage and backup)	This is also foreseen through the ICOS Carbon Portal. A full DOI system is in place.

Data set	APO and 14CO2 flux database
Data set description	Ensemble of 3D (2D in space, 1D in time) series of maps of 14CO2 and APO/O2 surface fluxes at global and European scales over 2004-2024, that will be generated or collected in task T3.2. The fluxes include the terrestrial 14CO2 disequilibrium calculated using the isotope enabled LPJ dynamic vegetation model, oceanic 14CO2 disequilibrium, 14CO2 emissions from nuclear reactors and reprocessing plants, 14CO2 signals from biofuels, ocean APO/O2 fluxes, and inventories of the APO/O2 fluxes associated to fossil fuel and biofuel combustion. The database will also include 3D fields of 14CO2 cosmogenic production. This database will be used as prior input for the 14CO2 and APO atmospheric inversions in tasks T3.3 and T3.4. It corresponds to D3.4.
Standards and metadata	The product will be available as a collection of netCDF files compliant with CF conventions.
Data Sharing	The dataset will be made publicly available once the corresponding deliverable D3.4 will be finalized (M24). We aim at using the ICOS-Carbon portal to share it and to provide a visualization interface.

Archiving and preservation (including storage and backup)	Unknown at this stage of the project but we aim at storing the data at the ICOS-Carbon portal.

Data set	**Ensembles of global scale emission and concentration estimates**
Data set description	Ensemble of CO2 flux (emission and absorption) estimates and CO2/14CO2/APO concentration timeseries from the various global scale inversions in task T3.3. The ensemble encompasses results (i) from the different global scale CO2 inverse modeling systems over 2004-2024 (ii) based on different sets of real observations (CO2, 14CO2 and/or APO datasets). For each inversion case, the ensemble gathers the prior (before inversion) and posterior (inverted) fluxes and concentration estimates: - 3D (2D in space, 1D in time) maps of anthropogenic emissions, ocean fluxes and terrestrial ecosystem fluxes - the observations assimilated by the inversion - samples of the simulated concentrations corresponding to the real observations This ensemble corresponds to part of D3.5. Scientific publication on these inversions are foreseen. It can be used as a benchmarking case for global scale CO2 inversion and to provide boundary conditions for regional scale inversions.
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	This ensemble will be made publicly available once finalized at this end of the project (M36). We aim at using a service such as that of the ICOS-Carbon portal to share it and to provide a visualization interface.
Archiving and preservation (including storage and backup)	Unknown at this stage of the project but we aim at storing this ensemble of inversion using a service such as the ICOS-Carbon portal.

Data set	**Ensembles of European scale emission and concentration estimates**
Data set description	Ensemble of CO2 flux (emission and absorption) estimates and CO2/14CO2/APO concentration timeseries from the various European scale inversions in task T3.4. The ensemble encompasses results (i) from the different European scale CO2 inverse modeling systems over 2019-

	2024 (ii) based on different sets of real observations (CO2, 14CO2 and/or APO datasets including or not the extra samplings/measurements in 2024). For each inversion case, the ensemble gathers the prior (before inversion) and posterior (inverted) fluxes and concentration estimates: - 3D (2D in space, 1D in time) maps of anthropogenic emissions, ocean fluxes and terrestrial ecosystem fluxes - the observations assimilated by the inversion - samples of the simulated concentrations corresponding to the real observations This ensemble corresponds to part of D3.6. Scientific publication on these inversions are foreseen. It can be used as a benchmarking case for regional scale CO2 inversions over Europe.
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	This ensemble will be made publicly available once finalized at this end of the project (M36). We aim at using a service such as that of the ICOS-Carbon portal to share it and to provide a visualization interface.
Archiving and preservation (including storage and backup)	Unknown at this stage of the project but we aim at storing this ensemble of inversion using a service such as the ICOS-Carbon portal.

Work Package 4:

Completed by Patricia de Rosnay, Fabienne Maignan and Jean-Christophe Calvet with input from WP partners

Data set	**IFS coupled land-atmosphere assimilation experiments**
Data set description	Generated data: Set NWP experiments to assess the impact of active and passive microwave and Solar Induced Fluorescence data assimilation on carbon fluxes. Input data: ASCAT, SMOS, SMAP, AMSR2, TROPOSIF. System configuration: using ML-based observation operators developed in CORSO. Time span: sampling a few months for summer and winter seasons post-2018. Horizontal resolution: TCo399 (25km), global domain

	The data underpins scientific publications related the CORSO coupled land-atmosphere data assimilation for the CO2MVS
Standards and metadata	Data is produced in GRIB-1 and GRIB-2 format and is stored in the ECMWF MARS archive. The dataset will be reviewed internally in WP4.
Data Sharing	Data will be available via the ECMWF MARS archive and CORSO consortium members can access this using their ECMWF member-state account. The data will be publicly available on request.
Archiving and preservation (including storage and backup)	The data in the MARS tape library are backed up.

Data set	**IFS simulations at 9km and 4km**
Data set description	*Using the new urban tile of the land surface model, updated vegetation cover (adjusted with urban cover), and the latest photosynthesis model to assess the impact on the CO2 biogenic signal from city plume.*
Standards and metadata	*The data will be produced in grib format which can be converted to NetCDF, including relevant metadata.* *The dataset is mostly part of a sensitivity study to assess impact of biogenic fluxes and resolution on the CO2 enhancement, so it will not be evaluated with independent data.*
Data Sharing	*This dataset is not foreseen to be shared with the public as it is part of a model-based sensitivity study. If the results from the case study are deemed to be important and useful to the wider scientific community, they will be shared via a scientific publication, including the simulation data which will be shared using Zenodo or similar public repository.*
Archiving and preservation (including storage and backup)	*The data will be stored in the mars archive at ECMWF*

Data set	**ESA TROPOMI SIF (TROPOSIF)**
Data set description	The TROPOSIF L2B data was made available from May 2018 till end of 2021 thanks to the ESA TROPOSIF project (https://s5p-troposif.noveltis.fr/). The operational production of L2 data files has then been ensured by the ESA S5P-PAL (Product Algorithm Laboratory) system, from 2022 onwards, with a daily update (https://data-portal.s5p-pal.com/browser/). The data was described in Guanter et al. (2021). A similar dataset is produced by the CalTech team (Köehler et al., 2018), without an operational framework. This dataset is available here: ftp://fluo.gps.caltech.edu/data/tropomi/

	Guanter, L., Bacour, C., Schneider, A., Aben, I., van Kempen, T. A., Maignan, F., ... & Zhang, Y. (2021). The TROPOSIF global sun-induced fluorescence dataset from the Sentinel-5P TROPOMI mission. Earth System Science Data, 13(11), 5423-5440. Köhler, P., Frankenberg, C., Magney, T. S., Guanter, L., Joiner, J., & Landgraf, J. (2018). Global retrievals of solar-induced chlorophyll fluorescence with TROPOMI: First results and intersensor comparison to OCO-2. Geophysical Research Letters, 45(19), 10-456.
Standards and metadata	The product user manual is available here: https://s5p-troposif.noveltis.fr/wp-content/uploads/2021/05/NOV-FE-0956-MU-019_v2.0.pdf The metadata is included in the netcdf files. The TROPOSIF products are provided in self-explanatory netCDF-4 files as ungridded data. The L2B data generated in the frame of the TROPOSIF project and available from ftp://fluo.gps.caltech.edu/data/tropomi/ consist in L2B daily files, with only valid retrievals. The recent data files generated by ESA S5P-PAL consist in L2 orbit files, including all retrievals (the selection of only the valid ones requires the exploitation of the quality flag). Data checks and verification was already done within the timeframe of the corresponding ESA project.
Data Sharing	The data is freely available to the scientific community, for non-commercial purposes only, under the Creative Common license BY-NC. See https://s5p-troposif.noveltis.fr/wp-content/uploads/2021/09/TROPOSIF_Data_Use_Policy.pdf http://ftp.sron.nl/open-access-data-2/TROPOMI/tropomi/sif/v2.1/l2b/ https://data-portal.s5p-pal.com/
Archiving and preservation (including storage and backup)	Provided by ESA PAL.

Document History

Version	Author(s)	Date	Changes
1.0	Tanya Warnaars,	12/05/2023	Initial version
1.0	Tanya Warnaars & Rhona Phipps	20/06/2023	Draft version with Annex II

Internal Review History

Internal Reviewers	Date	Comments
Jean-Christophe Calvet (MF)	June 2023	Comments addressed
Marko Scholtze (ULUND)	June 2023	No additional comment