CARBON ATMOSPHERIC TRACER RESEARCH TO IMPROVE NUMERICAL SCHEMES AND EVALUATION

CATRINE
Carbon Atmospheric Tracer Research to Improve Numerics and Evaluation

D9.3 Data Management Plan

Due date of deliverable	30/6/2024
Submission date	
File Name	D9.3 Data Management Plan
Work Package /Task	WP9/ Task 9.4
Organisation Responsible of Deliverable	ECMWF
Author name(s)	Tanya Warnaars, Rhona Phipps, Anna Agusti-Panareda, Michail Diamantakis and WP partners
Revision number	1
Status	Final
Dissemination Level / location	Public

Funded by the European Union

The CATRINE project (grant agreement No 101135000) is funded by the European Union.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them.

1 Executive Summary

The CATRINE 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 and the ICOS Carbon Portal.

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

This CATRINE 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 CATRINE 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 support EU countries in achieving the targets, the EU and European Commission (EC) recognise the need to support establishing the new European anthropogenic CO_2 emissions Monitoring and Verification Support capacity (CO2MVS). To support the Commission and the CO2 Task Force with designing and ultimately building the CO2MVS, previous projects have been funded such as: the CO2 Human Emissions (CHE) project, the CoCO2 project (https://coco2-project.eu/) and recommendations from the VERIFY project (https://verify.lsce.ipsl.fr/). However, some of the recommendations from the CHE project were not available at the time of the definition of the CoCO2 project and could therefore not be fully taken into account. The EC, supported by the CO2 Task Force, took the various recommendations from the CHE project onboard as input to the Horizon Europe Work Programme and created two Calls. The CATRINE project addresses the requirements from one of those two resultant calls.

The Carbon Atmospheric Tracer Research to Improve Numerical schemes and Evaluation (CATRINE) project aims to evaluate and improve the numerical schemes for tracer transport in the new Copernicus anthropogenic CO2 emissions Monitoring and Verification Support capacity (CO2MVS) and more widely in the Copernicus Atmosphere Monitoring Service (CAMS). The research and development activities in CATRINE will focus on the priorities identified by these previous activities. The CATRINE project will contribute to the further development of the new Copernicus element for the monitoring of anthropogenic CO_2 and CH_4 emissions and sinks.

The main objectives of CATRINE are to improve the methods used to represent resolved tracer transport by the winds, with a particular focus on mass conservation, and to identify other systematic errors associated with unresolved processes represented by parametrizations. The project will define protocols for evaluating tracer transport models at both global and local scales. Test beds based on field campaign case studies will be developed, along with suitable metrics for tracer transport evaluation, utilising a range of tracers and observations at both global and local scales. These metrics will be employed in the operational CO2MVS to evaluate the implementation of new transport model developments, characterise transport accuracy and representativity in data assimilation, and provide a quality control stamp of tracer transport accuracy. Lastly, CATRINE will provide clear recommendations to the CO2MVS and the Carbon Cycle Community which works with atmospheric inversion models for the evaluation and quality assessment of tracer transport models

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

This D9.3 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, D9.3 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.

2.2.1 Reference Documents

[1] 101135000-CATRINE-HORIZON-CL4-2023-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

2.3 Project partners:

Partners	
EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS	ECMWF
COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES	CEA
METEO-FRANCE	METEO-FRANCE

WAGENINGEN UNIVERSITY	WU
KARLSRUHER INSTITUT FUER TECHNOLOGIE	KIT
HELSINGIN YLIOPISTO	UH
UNIVERSITE DE REIMS CHAMPAGNE-ARDENNE	URCA
ALBERT-LUDWIGS-UNIVERSITAET FREIBURG	UFR

3 Data Summary

Our Data Management Plan (DMP) is developed following a standard approach 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 9.4 (WP9) is responsible for this activity.

CATRINE activities are closely coordinated with related activities in the Copernicus Atmosphere Monitoring Service (CAMS) and the CORSO project. Whilst prior emission data sets are already available from CAMS, CoCO2, CORSO and other projects, While the CATRINE activities are very much focused on the development of the CO2MVS, the results will also be of high relevance for the carbon cycle science community in general.

The products of CATRINE 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 CATRINE project has finished.

The types of data that will be used or produced in the CATRINE project are satellite, field campaign and in-situ observations, prior emissions and natural fluxes, and results from tracer transport simulation studies. The data produced by the project will comply with the findable, accessible, interoperable, and reusable (FAIR) data principles, as much as possible. The data will be accessible using existing data portals, such as the Copernicus Atmosphere data Store, designed to support interoperability and include clear licensing information as well as tools to make best use of the data. Some data will also be made available in publications and a public repository (e.g. Zenodo or Pangaea).

Reports will be openly available from the public pages of the central CATRINE website (www.catrine-project.eu). To increase its visibility, the CATRINE website will be linked on the websites of ECMWF, CAMS, C3S, and other partners.

All mature data products of CATRINE will be made publicly available to maximize the uptake by the scientific community. These include the results from the various modelling studies (depending on maturity). It is envisaged to make use of several parallel data portals to ensure full visibility of the datasets.

The topics being addressed in the CATRINE project are active research areas and the CATRINE results will contribute to the scientific discussions and developments around the world. This especially the case for the high resolution simulations that CATRINE will deliver at global scale from sub-degree resolution to 4.5 km with several state-of-the-art European transport models and several very high resolution local models, and a new framework to evaluate the accuracy of the transport combining the use of a range of observed tracers including CO2, Radon 222 (Rn222) and Sulphur hexafluoride (SF6) together with idealised tracers. These will form an important new asset to evaluate and quality control global transport models used in carbon cycle studies in more detail

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

[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

[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

CATRINE has a strong link to the Copernicus Atmosphere Monitoring Service. The close collaboration will ensure that the CATRINE activities are complementary to what is done elsewhere, like in the project CORSO, CoCO2. In addition, CATRINE will also interact with global initiatives that work on the interface of science and policy, such as the Global Carbon Project (GCP), the WMO Integrated Global Greenhouse Gas Information System (IG3IS), the Global Climate Observing System (GCOS), and the World Climate Research Programme (WCRP).

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, 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 reusable.

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, CATRINE 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, 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

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

and used, if possible, are publicly open and available. The CATRINE 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 CATRINE 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 CATRINE 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, CATRINE 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, 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.

The CATRINE 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, CATRINE will incorporate suitable standards and vocabularies for data and metadata creation. However in the case of CATRINE, 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 CATRINE 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, CATRINE 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 CATRINE FAIR have been included in the budget of the project. In general, the CATRINE consortium as a whole will decide and contribute to relevant aspects of the data management cycle during and after the completion of this project.

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 CATRINE budget.

7 Data Security

The CATRINE 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, 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 report, 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 CATRINE. Additionally, the Grant Agreement and the CATRINE 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 Other issues

The project will not make use of other national/funder/sectorial/departmental procedures for data management.

10 Conclusion

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

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

11 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 CATRINE 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*

12 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 CATRINE. 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 WP lead and co-lead with input from WP1 partners

Data set	**Advection method sensitivity tracer experiments**
Data set description	Generated experiments: • Set atmospheric composition experiments to test the sensitivity of different methodologies applied to the global CO2MVs advection scheme and identify the best currently available options. • Likewise test sensitivity of advection methodologies using 'idealised' tracer experiments with artificial tracers having full control on different spatial characteristics and geographical location at initial time. Time span: • Experiments will sample from few days to few months depending on the nature of the test. Horizontal resolution: TCo399 (25km). Few cases can be run at higher resolution (Tco1279, 9km). The data underpins scientific publications related to numerical methods for atmospheric transport
Standards and metadata	IFS data is produced in GRIB-1 and GRIB-2 format and is stored in the ECMWF MARS archive. It can also be downloaded in NetCDF4 format. Results from ICON simulation will be available in NetCDF4 format. Metadata will be included in the GRIB and NetCDF headers. The dataset will be reviewed internally in WP1.
Data Sharing	Data from IFS experiments will be available via the ECMWF MARS archive and CATRINE consortium members can access this using their ECMWF member-state account. NetCDF4 data that may be produced from other participating models and any other non-GRIB format data of relevance will be stored in the CATRINE account of the ftp site: https://myftp.ecmwf.int/login CATRINE consortium members can access this account.

	The data can be made publicly available on request if deemed interesting for the scientific community. A lot of these tests will be of internal nature to tune our developments.
Archiving and preservation (including storage and backup)	The data in the MARS tape library are backed up.

Work Package 2

Completed by WP lead and co-lead with input from WP2 partners

Data set	**IFS data-assimilation experiments to assess linear model and adjoint developments for CO2MVS**
Data set description	Generated experiments: • NWP experiments with the IFS system to test new tangent linear and adjoint developments which will eventually be used by the 4D-VAR minimization algorithm of the CO2MVS. These experiments will use moist tracers as test bed to exploit the large number of available observations for verification and validation of performance. Time span: experiments sampling 2-3 months in a winter or summer season at a recent year post 2022. Horizontal resolution: TCo399 (25km), global domain. The data underpins scientific publications related to 4D-VAR assimilation development.
Standards and metadata	Data is produced in GRIB-1 and GRIB-2 format and is stored in the ECMWF MARS archive. It can also be downloaded in NetCDF4 format. Metadata will be included in the GRIB headers. The dataset will be reviewed internally in WP2.
Data Sharing	Data will be available via the ECMWF MARS archive and CATRINE consortium members can access this using their ECMWF member-state account. The data can be made publicly available on request if deemed interesting for the scientific community. A lot of these tests will be of internal nature to tune our developments.
Archiving and preservation (including storage and backup)	The data in the MARS tape library are backed up

Work Package 2

Completed by WP lead and co-lead with input from WP2 partners.

Data set	**Evaluation of improvements in global CO2MVS IFS advection scheme**
Data set description	Generated experiments: • Improvements implemented in the semi-Lagrangian global CO2MVs advection scheme will be evaluated repeating some of the experiments conducted in WP1 and described in dataset 'Advection method sensitivity tracer experiments'. • Experiments demonstrating capability to couple the global CO2MVS model IFS with a locally conserving advection scheme. • NWP forecast experiments will be conducted to test the implementation of a 'dry mass conserving' continuity equation formulation described in task 2.1 of WP2. Time span: • Atmospheric composition experiments sampling from few days to few months depending on the nature of the test. • NWP forecast experiments sampling part of the winter/summer season at a recent year post 2022. Horizontal resolution: TCo399 (25km). Few cases can be run at higher resolution (Tco1279, 9km). The data underpins scientific publications related to numerical methods for atmospheric transport.
Standards and metadata	IFS data is produced in GRIB-1 and GRIB-2 format and is stored in the ECMWF MARS archive. It can also be downloaded in NetCDF4 format. Data produced from NWP experiments conducted in Meteo-France will be also in GRIB format and they can be made available on the ftp site. Metadata will be included in the GRIB headers. The dataset will be reviewed internally in WP2.
Data Sharing	Data will be available via the ECMWF MARS archive and CATRINE consortium members can access this using their ECMWF member-state account. The data can be made publicly available on request if deemed interesting for the scientific community. A lot of these tests will be of internal nature to tune our developments.
Archiving and preservation (including storage and backup)	The data in the MARS tape library are backed up.

Work Package 3 & 4

Data set	**Emission data sets WP 3/4**
Data set description	Within WP3/4 we will conduct model intercomparisons. Modellers will be asked to use common emission data sets to focus the comparison on transport differences. We will provide emissions of CO2, per sector, and with diurnal time factors. For some cases, we will also address CH4 (Rotterdam), and chemistry simulations. Here we will additionally provide emissions of NOx, CH4, CO, and VOCs. Depending on the case, resolution will be limited to available data (e.g. at 1x1 km2). Models might run a finer resolution, so some downscaling might be needed.
Standards and metadata	We will provide CF-compliant netcdf4 datasets. Metadata will include information about the data source, resolution, geographical locations. Data will be reviewed by all participating models, and potential issues will be resolved by the WP leads.
Data Sharing	Initially, data will be shared among the modellers. Access will be provided through a common data-share at ECMWF. Upon publication, we will deposit the data to a common repository (Zenodo).
Archiving and preservation (including storage and backup)	Apart from sharing in the ECMWF data-share, we will also keep local copies. Moreover, we will deposit and share the software to generate the data (from raw available data sources) to GitHub. Upon publication, Zenodo will provide a DOI.

Data set	**Model output WP3/4**
Data set description	Within WP3/4 we will conduct model intercomparisons. Modellers will provide output according to the model intercomparison protocols. We will ask the modellers to provide output on native model resolution. Moreover, we will ask the modellers to provide time series on locations on which observations are available. Apart from CO2, and other chemical tracers, we will ask for output of meteorology, like, temperature, winds, boundary layer height, moisture fields, etc. This depends on the available observations.
Standards and metadata	We will ask for CF-compliant netcdf4 datasets. Metadata will include information about the resolution, geographical locations, etc. Data will be reviewed by the PostDocs/WP leads that are responsible for the Deliverables in WP3/4.
Data Sharing	Initially, data will be shared among the modellers.

	Access will be provided through a common data-share at ECMWF. Upon publication, we will deposit the data to a common repository (Zenodo).
Archiving and preservation (including storage and backup)	Apart from sharing in the ECMWF data-share, we will also keep local copies. Moreover, we will archive the model versions that were used to produce the output on GitHub (as Github version). Upon publication, Zenodo will provide a DOI of the model output.

Data set	**Observations WP3/4**
Data set description	Within WP3/4 we will conduct model intercomparisons. We will compare the output of several models to a set of observations that are available for the specific case (Rotterdam, Indianapolis, Paris, …). Observations may come in various forms (meteorology, mole fractions, total columns, …).
Standards and metadata	We will ask for CF-compliant netcdf4 datasets. Metadata will include information about the species, provider of the data, time resolution, etc.
Data Sharing	Some of the data may consist of already published observations. We foresee, however, that parts of the data are not freely available. This will be indicated in the netcdf4 files, and these data are (with consent of the data providers) only shared among the participants. Upon publication, we will deposit the observation data on Zenodo. Depending on the data-provider, restrictions in the use of the data will apply.
Archiving and preservation (including storage and backup)	Upon publication of scientific papers of project deliverables, Zenodo will provide a DOI.

Work Package 5 and 6:

Data set	**CloudRoots-Amazon22** (https://cloudroots.wur.nl/)
Data set description	Observations have been collected during the campaign. They were part of the three-week campaign taken in the ATTO and Campina sites. A complete description of the campaign can be found at: https://journals.ametsoc.org/view/journals/bams/aop/BAMS-D-23-0333.1/BAMS-D-23-0333.1.xml
Standards and metadata	All the observations are checked according to the quality data
Data Sharing	CloudRoots-Amazon22 (https://cloudroots.wur.nl/)

Archiving and preservation (including storage and backup)	The data archive is under construction. Data can be obtained by contacting Oscar Hartogensis (Oscar.hartogensis@wur.nl)

Data set	**Ruisdael (https://ruisdael-observatory.nl/cesar/), and the Meteorology and Air quality group (https://maq-observations.nl/)**
Data set description	Observations are collected continuously.Meta-data information is documented, and continuously updated at https://maq-observations.nl/instruments/. The meta-data includes variable names, variable units, instrument type, long names, notes (e.g. calibrated/uncalibrated), inclusion in download packages and measurement interval. The meta-data sheet also contains contact information, (co-)authors, dataset version and a full list of edits to the meta-data sheets. All data are contained in a mySQL database, which can be accessed via a user interface, or using a user-specific API key described on the website. Upon using the user interface, the user will receive the data in a .csv format. All data only have a temporal component, with varying timesteps from 20 seconds to 30 minutes. The MAQ-Observations is subject to automatic data-quality checks, as well as regular manual quality assurance checks. Upon identifying potential troublesome data, the mySQL database will be updated.
Standards and metadata	All the observations are checked according to the quality data
Data Sharing	Data is shared through our web portal www.maq-observations.nl and can be accessed through a user-interface or using a personal API key (e.g. json commands or Python requests). Data is available in near-real time (~ 10 minutes after time of collection) and is available for the full historical record (back to 2011 depending on the station and instrument). MAQ-Observations follows an open-access policy and are licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0). There is no embargo in place for specific periods our user groups.
Archiving and preservation (including storage and backup)	The data is contained in a MySQL database and backed up at a frequency of 1 week. These backups are kept for 4 consecutive backups. The raw data used to fill the database, including the necessary pre-processing scripts are also backed up at Wageningen University (W:/ drive). The current size of the database is 50 GB, but because the data collection is still operational, this will grow with ~3 GB per year. Additionally, the database can be expended with newly planned measurements which are currently not yet in the database. As long as the measurement sites are operational, the data will be stored in our mySQL database, and accessible through www.maq-observations.nl. Long-term archiving of the raw data is in place at Wageningen University.

Data set	**DALES (https://github.com/dalesteam/dales)**
Data set description	Large-eddy simulations; data is generated or will be generated and archived at (https://www.surf.nl/en) > this will include a description of the case (README file) and all the specifications of the numerics. The model description can be found at:

	https://gmd.copernicus.org/articles/3/415/2010/gmd-3-415-2010.html https://link.springer.com/article/10.1007/s10546-016-0182-5
Standards and metadata	Model results are standardized and formatted in netcdf files
Data Sharing	DALES code and results are open to ensure reproducibility of the numerical experiments. DALES code is made available under the terms of the GNU GPL version 3, see the file COPYING for details.
Archiving and preservation (including storage and backup)	Modelling results are archived at SURFSARA (https://www.surf.nl/en)). In cases of publications with some specifics the paper will include the URL of ZENODO

Data Set	**ICON-ART**
Data set description	ICON-ART simulations in different resolutions; data will be generated at the HPC center at KIT. This dataset includes all model output, the model setup and the executable.
Standards and metadata	Model results are saved in netcdf4 format and will use CF-1.6 conventions for metadata. The model setup is saved as ASCII file.
Data Sharing	The ICON code is available in DKRZ gitlab (access only on request). The dataset itself will be made available on a Thredds server that can be accessed by everyone (similar to this CARIBIC dataset: https://thredds.atmohub.kit.edu/thredds/catalog/caribic/IAGOS-CARIBIC_MS_files_collection_20240112/CARIBIC-1/catalog.html)
Archiving and preservation (including storage and backup)	Data will be long-term archived in the bwDataArchive (https://www.rda.kit.edu/english/index.php) Subsets of data (e.g. used in publications) will be published at RADAR4KIT (https://www-bibliothek-kit-edu.translate.goog/english/radar.php?_x_tr_sl=en&_x_tr_tl=de&_x_tr_hl=de&_x_tr_pto=sc) and will get a DOI there

Data Set	**IFS**
Data set description	IFS simulations in different resolutions, using different configurations including coupling of energy/water fluxes with photosynthesis, using different turbulent mixing schemes and perturbing model parameters to produce an ensemble of simulations. The simulations will be used as part of WP5-6 test beds to evaluate the model transport, with emphasis on the parametrizations of turbulent mixing and convection.
Standards and metadata	Model results are saved in grib and netcdf4 formats.
Data Sharing	Data will be available via the ECMWF MARS archive and CATRINE consortium members can access this using their ECMWF member-state account. The data for the domain around the test beds will be archived together with ICON-ART and DALES simulations on a Thredds server hosted by KIT.

Archiving and preservation (including storage and backup)	Data will be long-term archived in the bwDataArchive (https://www.rda.kit.edu/english/index.php) Subsets of data (e.g. used in publications) will be published at RADAR4KIT (https://www-bibliothek-kit-edu.translate.goog/english/radar.php?_x_tr_sl=en&_x_tr_tl=de&_x_tr_hl=de&_x_tr_pto=sc) and will get a DOI there

Work Package 7 and 8:

Data set	**high-resolution atmospheric transport model intercomparison**
Data set description	Global tracer transport simulations at “high resolution” under a common protocol, prepared by CATRINE partners and external voluntary teams. The simulations focus on carbon dioxide (CO2), sulfur hexafluoride (SF6) and the radon isotope 222Rn. In addition, the transport of imposed fossil CO2 emission fields is considered. Optional tracers like water vapour may also be simulated. The high-resolution simulation database will be the first of this sort and may serve various applications beyond the initial motivation on model evaluation, from, e.g., model uncertainty quantification to AI training. It is planned to publish it in the scientific literature.
Standards and metadata	Data standards will follow standards already is use on the Atmospheric Data Store, in particular for the “CAMS global inversion-optimised greenhouse gas fluxes and concentrations”. This includes the metadata. The format will be NetCDF. Data will have been quality-checked in the intercomparison process.
Data Sharing	The submitted data of the transport simulations will be evaluated internally together with the data providers. Quality-controlled data will be distributed freely and publicly via the CATRINE web site (and behind it from https://thredds-su.ipsl.fr/thredds/catalog/tgcc_thredds/work/p24cheva/CMEMS/catalog.html). Users of the data will be encouraged to give fair credit to the contributors and contact them.
Archiving and preservation (including storage and backup)	The data will be archived for long-term storage on https://hpc.cea.fr/fr/TGCC.html It will cover of the order of 100 TB. A doi will be requested, but the originated authority has not been identified yet.

Document History

Version	Author(s)	Date	Changes
0.1 (Initial document created)	T. Warnaars, R.Phipps, A.Agusti-Panareda, M.Diamantakis	June 2024	Initial version
1.0 version		June 2024	Final version, input from internal reviewers included

Internal Review History

Internal Reviewers	Date	Comments
Maarten Krol	24/06/2024	Editorial formatting
Frederic Chevalier	24/06/2024	Editorial formatting