CAMS AERosol Advancement

D11.4 Data Management Plan

Due date of deliverable	July 2024
Submission date	
File Name	D11.4 Data Management Plan
Work Package /Task	Task 11.4
Organisation Responsible of Deliverable	HYGEOS
Author name(s)	Silvia Jacob, Samuel Remy
Revision number	1
Status	Issued
Dissemination Level	PU (Public)

CAMS AERosol Advancement

Horizon Europe RIA (Research and Innovation Action)

Call: HORIZON-CL4-2023-SPACE-01
TOPIC ID: HORIZON-CL4-2023-SPACE-01-31

Project Coordinator:	Dr Samuel REMY (HYGEOS)
Project Start Date:	01/01/2024
Project Duration:	36 months

Published by the CAMAERA Consortium

Contact:
HYGEOS, 165 Avenue de Bretagne, 59000 Lille, France, sr@hygeos.com, sj@hygeos.com

Acknowledgement and Disclaimer

Funded by the European Union

The CAMAERA project (grant agreement No 101134927) 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 CAMAERA Data Management Plan (DMP) 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 gridded analysis data, satellite products and in-situ observations. 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 of the participating organisations, many of them operational centres with a well developed infrastructure for data storage and sharing. Once the CAMAERA data results and products have been adopted by the Copernicus Atmosphere Monitoring Service, they will be distributed by the Copernicus Atmospheric Data Store.

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

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

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 CAMAERA. 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 CAMAERA.

This CAMAERA Data Management Plan (DMP) 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 CAMAERA 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

Monitoring the composition of the atmosphere is a key objective of the Copernicus programme. The Copernicus Atmosphere Monitoring Service (CAMS) combines satellite observations with numerical modelling by means of data assimilation and inversion techniques to turn the content of Earth-Observation data into information products that address some of today's major societal topics.

The HORIZON Europe-funded CAMS AERosol Advancement (CAMAERA) project aims at providing strong improvements of the aerosol modelling capabilities of the regional and global CAMS systems, on the assimilation of new sources of data, and on a better representation of secondary aerosols and their precursor gases.

The overall goal is to enhance the quality of key products of the CAMS service such as PM2.5, PM10 and therefore help CAMS to monitor air pollutants. To achieve this purpose, CAMAERA develops new prototype service elements of CAMS, beyond the current state-of-the-art, and complements research topics addressed in the CAMEO project.

CAMAERA will upgrade and improve the aerosol and precursor gases modelling capabilities of the global and regional CAMS systems, focusing on key topics such as dust emissions and secondary organic aerosols.

CAMAERA will test the assimilation of new streams of data, allowing for a better constraint of the simulated state of the composition of the atmosphere.

CAMAERA will contribute to the medium- to long-term evolution of the CAMS production systems and products.

The transfer of developments from CAMAERA into subsequent improvements of CAMS operational service elements is a main driver for the project and is the main pathway to impact for CAMAERA.

The CAMAERA consortium, led by HYGEOS, includes all of the partners operating the regional and global CAMS systems, as well as most of partners contributing to the development and upgrade of these systems. This allows CAMAERA developments to be carried out directly within the CAMS production systems and facilitates the transition of CAMAERA results to future upgrades of the CAMS service.

This will maximise the impact and outcomes of CAMAERA as it can make full use of the existing CAMS infrastructure for data sharing, data delivery and communication, thus supporting policymakers, business and citizens with enhanced atmospheric environmental information.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverables

This D11.4 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

In this deliverable, the work as planned in the Description of Action (DoA, WP11 T11.4) was performed.

2.2.3 Deviations and counter measures

No deviations have been encountered.

2.2.4 Reference Documents

[1] Project: 101134927 — CAMAERA — HORIZON-CL4-2023-SPACE-01

2.2.5 CAMAERA Project Partners:

(Participant number order)

HYGEOS SARL (HYGEOS,

NEDERLANDSE ORGANISATIE VOOR TOEGEPAST NATUURWETENSCHAPPELIJK ONDERZOEK TNO (TNO),

RESEARCHCONCEPTS IO GMBH (RC.io),

METEOROLOGISK INSTITUTT (METNorway),

KONINKLIJK NEDERLANDS METEOROLOGISCH INSTITUUT-KNMI (KNMI),

ILMATIETEEN LAITOS (FMI),

BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION (BSC CNS),

EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS (ECMWF),

METEO-FRANCE (METEO-FRANCE),

INSTYTUT OCHRONY SRODOWISKA - PANSTWOWY INSTYTUT BADAWCZY (IOS-PIB),
FORSCHUNGSZENTRUM JULICH GMBH (FZJ),
AARHUS UNIVERSITET (AU),
SVERIGES METEOROLOGISKA OCH HYDROLOGISKA INSTITUT (SMHI),
AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE (ENEA),
INSTITUT NATIONAL DE L ENVIRONNEMENT INDUSTRIEL ET DES RISQUES - INERIS (INERIS),

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 11.4 (WP11) is responsible for this activity.

The products of CAMAERA 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. Improved methods, algorithms and code are meant to form the basis for follow-on development after the CAMAERA project has finished.

Datasets arising from the project:

- Best estimate of global dust emissions
- Whitecap fraction estimated offline by deep neural network/machine learning techniques
- Intercomparison of dry deposition simulated by various regional models and the global system
- Multi model data applying different biogenic emissions and secondary organic aerosol schemes
- Intercomparison of global and regional models using similar resolution and emissions for the year 2018

CAMAERA can make use of existing CAMS infrastructure for data sharing and data delivery. CAMS information products are freely available and efficiently disseminated by the Copernicus Atmosphere Data Store (ADS). CAMAERA's results once implemented in CAMS, will be directly available together with the corresponding products.

Dust emissions are one of the most hardest process to model, as it cumulates uncertainties from the inputs (meteorology, soil characteristics) and modelling uncertainties (the processes are not entirely known and occur at scales much smaller than the global model). The provision of best estimates of dust emissions, computed offline through an ensemblist approach, will serve many purposes. First, internally to CAMAERA, this dataset will be used to train the deep neural network in work package 6. Second, internally to CAMS, this dataset can be used to modulate the emissions of the global system in order to correct for systematic errors at specific locations. Finally, for CAMS users and the general public, this dataset will provide a reference dataset that can be used for scientific studies, to compare to other models or for other purposes.

Whitecap fraction is used in the global CAMS system as an observable proxy to sea-salt aerosol emissions. However, outside of CAMS, whitecap fraction is used mainly to compute the albedo of ocean surfaces, in particular for remote sensing applications. The provision of a reference dataset for whitecap fraction can thus be of use to a variety of users, in and outside of CAMS.

Dry deposition is a key driver of simulated PM2.5 and PM10 in the CAMS regional and global systems. It is also a process that is relatively overlooked in terms of model intercomparison, unlike gaseous dry deposition. The publication of a dataset of model intercomparison of dry deposition as well as that of the observational dataset used to evaluate the simulated dry deposition will thus be of interest to the whole atmospheric composition modelling community. It can also be used in the future to evaluate dry deposition schemes to be implemented in the regional or global CAMS models.

The CAMS global and regional use secondary organic aerosol production schemes of various complexity, together with specific online biogenic emissions modules. The publication of a dataset resulting from the intercomparison of regional and global model focusing on secondary organic aerosols will benefit future CAMS work, as well as the scientific community.

An intercomparison of regional models focusing on deposition has been carried out in 2023 in the framework of the CAMS2_40 project. Within CAMAERA, this intercomparison will be extended with more regional models, as well as the global CAMS system using a similar resolution and emissions as the regional models. Selected fields from the multi model dataset will be made public and will provide insights into how the different processes impacting the aerosol life cycle are represented in the regional and global systems.

3.1 Definitions related to the approach to Open Science

The Horizon Europe programme guide states[1]: “*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 CAMAERA 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 CAMAERA 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

[1] 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

preserved. The types of data that will be used or produced in the project are satellite and in-situ[2] observations, prior emissions, and results from inversion studies.

CAMAERA has a strong link to the CAMS Service. The close collaboration will ensure that the CAMAERA activities are complementary to what is done elsewhere in other projects/ initiatives e.g. the EU projects CAMEO, CATRINE, CORSO and CoCO2.

4 FAIR data principles

The data of the project will comply with the FAIR data principles, as much as possible.

CAMAERA's results and code once implemented in CAMS, will be available via the Copernicus ADS together with the corresponding products.

The ADS has 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[3]. 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, CAMAERA 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).

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 CAMAERA partnership will ensure

[2] 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

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

the integrity of personal data and sensitive information prior to the dissemination of the datasets.

The project will maintain a list of data sets it accesses for the purposes of CAMAERA 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 CAMAERA 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, CAMAERA 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. The public 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 a third party which has denied permission for sharing the data;
- 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 CAMAERA DMP Annex 2 provides the specific information indicating the versions or parts of the data sets that can(not) be freely shared. The repositories for data set publication and preservation may be further defined 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.

CAMAERA will carefully consider the CAMS performance measures and integration procedures during the project and will structure its work in such a way as to minimise future technical implementation efforts for CAMS. To enable such a cost-efficient implementation, most of CAMAERA’s research and development work will be carried out directly with the regional and global CAMS systems and their computing, archiving, dissemination and software environment. Such close alignment with the operational CAMS service is possible because the CAMAERA consortium includes HYGEOS, currently leading the development work of the global system, ECMWF, all contractors operating the regional CAMS production system, and additional partners with leading roles in CAMS service contracts.

CAMAERA can make use of existing CAMS infrastructure for data sharing and data delivery. CAMS information products are freely available and efficiently disseminated by the Copernicus ADS. CAMAERA's results and uncertainty information once implemented in CAMS, will be directly available together with the corresponding products.

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, CAMAERA 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 Allocation of resources

The resources required for making the data generated by CAMAERA "FAIR" have been included in the budget of the project. In general, the CAMAERA consortium as a whole will decide and contribute to relevant aspects of the data management cycle during and after the completion of this project. 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 CAMAERA budget.

6 Data security

The CAMAERA consortium places 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 or storage environments, we will investigate in depth options such as Zenodo.

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.

7 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 CAMAERA. Additionally, the Grant Agreement and the CAMAERA 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

8 Conclusion

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

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

Annex 1:

Annex I includes the template used to collect the information from WP leaders regarding data to be used or produced. The completed tables are in Annex 2

WP leaders to complete the 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 work package of CAMAERA. The table below shows each data set that:

- *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?* *Usage constraints?*
Standards and metadata	*Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.* *Will you generate proper metadata for your data?* *If yes: how do they look like?* *If no: why?* *Data format?* *Will there be a review process to quality-check the data?*
Data Sharing	*Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).* *In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).*

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

Annex 2:

Annex 2 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 work package of CAMAERA. 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).

Workpackage 1/2

Completed by: Melanie Ades with input from WP partners

Data set	**Five-year dataset (2017-2022) of dust emissions**
Data set description	First guess and analysis total dust emissions from 2017-2022. The emissions are calculated using an ensemble of IFS dust forecasts input into an offline ensemble Kalman smoother. The method provides a correction factor to the original dust emissions based on the ensemble uncertainty combined with observational information to improve the estimate of emissions.
Standards and metadata	The data will be provided in the form of Netcdf files with metadata internally in the header. The metadata and information related to the dataset will also be documented in a publicly available deliverable report in the project.
Data Sharing	The data will be publicly available and accessible through Zenodo (https://zenodo.org/). For direct use in the project, the data will be shared as necessary between partners of the project. The dataset will be made available to a general audience once the project is completed.
Archiving and preservation (including storage and backup)	The final dataset will be archived in Zenodo.

Data set	Global IFS analysis data
Data set description	Global IFS analysis data based on assimilation tests with new datasets, model and data assimilation configuration developments. These are test experiments to document the impact of the assimilation of new lidar/ceilometer profiles, model developments or data assimilation method developments and while publicly available are not really intended for public use. If successful, the improvements will be implemented in the operational CAMS system and those data are publicly available from the CAMS ADS: https://atmosphere.copernicus.eu/data
Standards and metadata	The simulations are based on the state-of-the-art global IFS model which is well documented. CAMAERA will use the standard IFS-COMPO chemistry scheme. The model parameters are based on the WMO standard meteorological parameters and described in grib parameter database (https://apps.ecmwf.int/codes/grib/param-db). The data will be available in GRIB and netcdf format. The data (and any metadata) will be documented in a deliverable report.
Data Sharing	The data will be publicly available and accessible through the ECMWF API (https://www.ecmwf.int/en/forecasts/access-forecasts/ecmwf-web-api).
Archiving and preservation (including storage and backup)	MARS archive. The IFS simulations will be archived in the MARS tapes at ECMWF and will be classified as "publication datasets" which means they will be preserved for at least 5 years after which the dataset preservation will be reviewed.

Workpackage 3/4

Completed by: Vincent Huijnen with input from WP partners

Dataset	**Global IFS hindcast data**
Data set description	Global IFS hindcast data based on forecast tests with a range of new model configuration developments. These are test experiments to document the quality of an alternative aerosol scheme, and alternative parameterizations for biogenic emissions which drive secondary organic aerosol formation. The simulation data of test experiments will be publicly available, but not intended for public use. Any successful model developments will be implemented in the operational CAMS system and those data are publicly available from the CAMS ADS: https://atmosphere.copernicus.eu/data.
Standards and metadata	The simulations are based on the state-of-the-art global IFS model which is well documented. CAMAERA will use the standard IFS-COMPO chemistry scheme. The model parameters are based on the WMO standard meteorological parameters and described in grib parameter database (https://apps.ecmwf.int/codes/grib/param-db). The data will be available in GRIB format. The data (and any metadata) will be documented in a peer-reviewed publication.
Data Sharing	The data will be publicly available and accessible through the ECMWF API (https://www.ecmwf.int/en/forecasts/access-forecasts/ecmwf-web-api).
Archiving and preservation (including storage and backup)	MARS archive. The IFS simulations will be archived in the MARS tapes at ECMWF and will be classified as “publication datasets” which means they will be preserved for at least 5 years after which the dataset preservation will be reviewed.

Work package 5/6

Completed by: Samuel Remy with input from WP partners

Data set	**Two years dataset (2014 and 2017) of whitecap fraction**
Data set description	Estimated whitecap fraction using deep neural network/machine learning techniques on meteorological and wave predictors.
Standards and metadata	The data will be provided in the form of Netcdf files with metadata internally in the header. The metadata and information related to the dataset will also be documented in a publicly available deliverable report in the project.
Data Sharing	The data will be publicly available and accessible through Zenodo (https://zenodo.org/). For direct use in the project, the data will be shared as necessary between partners of the project. The dataset will be made available to a general audience once the project is completed.
Archiving and preservation (including storage and backup)	The final dataset will be archived in Zenodo.

Data set	**0D dry deposition intercomparison dataset**
Data set description	This dataset provides the simulated dry deposition velocity as function of particle size for several land types. The participants models are the IFS, SILAM, MATCH, MINNI, LOTOS-EUROS and GEM-AQ. The observational dataset that will be used to evaluate the model output (already made public in Pleim et al. 2022) will be included.
Standards and metadata	The data will be provided in the form of Ascii files with metadata internally in the header. The metadata and information related to the dataset will also be documented in a publicly available deliverable report in the project.
Data Sharing	The data will be publicly available and accessible through Zenodo (https://zenodo.org/). For direct use in the project, the data will be shared as necessary between partners of the project. The dataset will be made available to a general audience once the project is completed.
Archiving and preservation (including storage and backup)	The final dataset will be archived in Zenodo.

Work package 7/8

Completed by: Hilde Fagerli with input from WP partners

Dataset	**Multi-model data applying different BVOC emissions and different SOA schemes**
Data set description	Global and Regional scale simulations (for 2019) from IFS, EMEP, CHIMERE, MATCH, DEHM; MONARCH model with different SOA schemes and different BVOC emissions. The simulation data from these comparisons will be publicly available.
Standards and metadata	The data will be available in netcdf format. The data (and any metadata) will be documented in openly available reports. We will aim at peer-reviewed publications where possible.
Data Sharing	Data will be shared among partners for analysis via external ftp or another platform to be agreed upon. Dissemination of data through existing repositories is currently under investigation.
Archiving and preservation (including storage and backup)	The IFS simulations will be archived in the MARS tapes at ECMWF and they will be classified as “publication datasets” which means they will be preserved for at least 5 years, and after which the dataset preservation will be reviewed. MET Norway will ensure local archiving of EMEP MSC-W model data, SMHI, AU-ENVS, BSC and INERIS will do the same for MATCH, DEHM, MONARCH and CHIMERE model data, respectively

Work package 9/10

Completed by: Samuel Remy with input from WP partners

Data set	**2018 multi-model datasets**
Data set description	This dataset includes selected simulated data by the participants of the CAMS2_40 regional model intercomparison (EMEP, SILAM, MOCAGE, LOTOS_EUROS, DEHM, MONARCH and CHIMERE), with additional simulated data from GEM-AQ, MINNI and from the global system IFS using similar emissions and resolution as the regional models.
Standards and metadata	The data will be provided in the form of Netcdf files with metadata internally in the header. The metadata and information related to the dataset will also be documented in a publicly available deliverable report in the project.
Data Sharing	The data will be publicly available and accessible through Zenodo (https://zenodo.org/). For direct use in the project, the data will be shared as necessary between partners of the project. The dataset will be made available to a general audience once the project is completed.
Archiving and preservation (including storage and backup)	The final dataset will be archived in Zenodo.

Document History

Version	Author(s)	Date	Changes
1.0	Silvia Jacob, Samuel Rémy,	08/07/2024	Issued version

Internal Review History

Internal Reviewers	Date	Comments

This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.