CAMS Service Evolution

D7.4 Data Management Plan

Due date of deliverable	June 2023
Submission date	04/07/2023
File Name	CAMEO-D7-4-V1.0
Work Package /Task	WP7/ T7.4
Organisation Responsible of Deliverable	ECMWF
Author name(s)	Rhona Phipps, Tanya Warnaars and CAMEO Consortium
Revision number	V1.0
Status	Issued
Dissemination Level	Public

Funded by the European Union

The CAMEO project (grant agreement No 101082125) 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 CAMEO 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 CAMEO 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 CAMEO 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 CAMEO. 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 CAMEO.

This CAMEO 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 CAMEO 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 European Union's flagship Space programme Copernicus, with the Copernicus Atmosphere Monitoring Service (CAMS) providing free and continuous data and information on atmospheric composition.

The CAMS Service Evolution (CAMEO) project will enhance the quality and efficiency of the CAMS service and help CAMS to better respond to policy needs such as air pollution and greenhouse gases monitoring, the fulfilment of sustainable development goals, and sustainable and clean energy.

CAMEO will help prepare CAMS for the uptake of forthcoming satellite data, including Sentinel-4, -5 and 3MI, and advance the aerosol and trace gas data assimilation methods and inversion capacity of the global and regional CAMS production systems.

CAMEO will develop methods to provide uncertainty information about CAMS products, in particular for emissions, policy, solar radiation and deposition products in response to prominent requests from current CAMS users.

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

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

The CAMEO consortium, led by ECMWF, the entity entrusted to operate CAMS, includes several CAMS partners thus allowing CAMEO developments to be carried out directly within

the CAMS production systems and facilitating the transition of CAMEO results to future upgrades of the CAMS service.

This will maximise the impact and outcomes of CAMEO 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 D7.4 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, WP7 T7.4) was performed.

2.2.3 Deviations and counter measures

No deviations have been encountered.

2.2.4 CAMEO Project Partners:

ECMWF	EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS
Met Norway	METEOROLOGISK INSTITUTT
BSC	BARCELONA SUPERCOMPUTING CENTER-CENTRO NACIONAL DE SUPERCOMPUTACION
KNMI	KONINKLIJK NEDERLANDS METEOROLOGISCH INSTITUUT-KNMi
SMHI	SVERIGES METEOROLOGISKA OCH HYDROLOGISKA INSTITUT
BIRA-IASB	INSTITUT ROYAL D'AERONOMIE SPATIALEDE BELGIQUE
HYGEOS	HYGEOS SARL
FMI	ILMATIETEEN LAITOS
DLR	DEUTSCHES ZENTRUM FUR LUFT - UND RAUMFAHRT EV
ARMINES	ASSOCIATION POUR LA RECHERCHE ET LE DEVELOPPEMENT DES METHODES ET PROCESSUS INDUSTRIELS

CNRS	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS
GRASP-SAS	GENERALIZED RETRIEVAL OF ATMOSPHERE AND SURFACE PROPERTIES EN ABREGE GRASP
CU	UNIVERZITA KARLOVA
CEA	COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES
MF	METEO-FRANCE
TNO	NEDERLANDSE ORGANISATIE VOOR TOEGEPAST NATUURWETENSCHAPPELIJK ONDERZOEK TNO
INERIS	INSTITUT NATIONAL DE L ENVIRONNEMENT INDUSTRIEL ET DES RISQUES - INERIS
IOS-PIB	INSTYTUT OCHRONY SRODOWISKA - PANSTWOWY INSTYTUT BADAWCZY
FZJ	FORSCHUNGSZENTRUM JULICH GMBH
AU	AARHUS UNIVERSITET
ENEA	AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE

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 7.4 (WP7) is responsible for this activity.

The products of CAMEO 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 CAMEO project has finished.

Datasets arising from the project:

- Delivery of 2 1-year 3MI proxy data
- Uncertainties in CAMS emission temporal profiles
- Uncertainties in Isoprene CAMS-GLOB-BIO emissions at the grid cell level
- Uncertainties in primary PM emissions from CAMS-REG at the grid cell level
- Uncertainties in CAMS-GLOB-ANT emissions at the country and sector-level
- Intercomparison of satellite-derived CO2, CH4 and NO2 emissions

CAMEO 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). CAMEO's results and uncertainty information once implemented in CAMS, will be directly available together with the corresponding products.

The provision of uncertainty information has been a recurring user request in CAMS. These user requests have been collected by CAMS in a systematic way and reflect the opinions from users and stakeholders in key positions. Uncertainty information is a key input to any data-based decision-making and planning process. Provision of uncertainty information will also increase the credibility of the respective CAMS products as it is a requirement for quality assured data sets. The methods applied in CAMEO to systematically derive uncertainty information will instigate the further development of the products, will provide a framework to monitor progress and help to set future research priorities. By providing uncertainty information it can be expected that more users, especially from application areas such as air quality management, solar energy companies and ecosystem monitoring, are inclined to use CAMS products for their mandated roles to report on the environment.

Extending the knowledge of uncertainty of solar radiation products will enable better planning processes in the solar energy application sector. This will result in larger security of energy supply with reduced investment costs, more reliable feasibility and viability assessments, larger bankability of solar investments and increased financial yields for investors. European

citizens will benefit from reduced costs of their individual energy supply invoice if the energy supply system planning and design relies on data with better known uncertainty structures.

Dust deposition is a major impact for solar energy production. Taking dust deposition into account properly in the solar energy system design process and its operation will ensure economic viability and reduce overall system costs of solar energy production. The application of deposition data without a realistic uncertainty specification strongly limits the application of the data, and the application of these data might even reduce the economic yield and the accuracy of predictions. The availability of accuracy information will therefore greatly improve the applicability of the data.

Quantifying the uncertainty in CAMS emission products will allow not only to increase the number of applications in which these datasets can be used, but also to improve the associated results. This includes, for instance, the Screening for High Emission Reduction Potential on Air (SHERPA) tool, developed by the Joint Research Centre of the EU (JRC) to support national to local authorities in the design and assessment of their air quality plans, and which currently uses as input CAMS regional emissions with simplified uncertainty assumptions due to lack of more detailed information. Moreover, the methods developed in CAMEO to quantify uncertainty in emissions can be later extrapolated to other CAMS emission products not considered in the present proposal such as shipping or soil emissions (CAMS-GLOB-SHIP and CAMS-GLOB-SOIL). Emission uncertainty information derived from CAMEO will also allow us to prioritise future efforts for improving the accuracy of current CAMS emission inventories and guide decisions on methodological choice. The final version of the CAMEO uncertainty information will be provided together with the releases of the corresponding CAMS emission products. The CAMEO methods to provide regular updates of the emission uncertainty could be implemented in CAMS.

Better knowledge of the uncertainties in the sector/source attribution in the CAMS policy products will help policy makers to make more robust decisions when prioritizing measures for specific activity sectors. Furthermore, there have been repeated user requests in the CAMS Policy User Workshops that CAMS should deliver air quality information at finer spatial resolution, making the results more relevant for the urban scale. In CAMEO, the linkages between urban and regional scales will be made, introducing a variability/correction estimate that will cover the resolution issue. This new information will make the source receptor results more applicable for urban areas and give better information on the local (urban) versus non-local contributions of emissions. In turn, this will better guide policy makers on the scale of actions that should be targeted when designing air quality policies. Furthermore, linking regional and urban scales in the CAMS policy products will make these products more relevant for policy analysis such as the impact assessments being done at present for the Ambient Air Quality Directive, where exceedances in hot spot areas possibly will be driving policy.

The largest single source of uncertainty in air pollution forecasting (and source apportionment) is probably related to emissions. Although there have been several studies where emission uncertainties have been propagated into air quality forecasts/source apportionment, to our knowledge there has been no attempt to generate such information in NRT. Clearly, such information will be very valuable for interpreting the drivers behind episode situations, during the episode itself, supporting communication both to the public as well as to policy makers. At present, only ozone and particulate matter (PM) are covered in the Source Contribution to EU Cities service.

The regional model resolution does currently not allow for reasonable results for NO_2 due to the short lifetime and large variability close to sources. However, the methodology developed in CAMEO could pave the way for NO_2 to be included in the city SR service.

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 CAMEO 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 CAMEO 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[2] observations, prior emissions, and results from inversion studies.

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

[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

[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

4 FAIR Data

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

CAMEO's results and uncertainty information 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, CAMEO 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 CAMEO partnership will ensure 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 CAMEO 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 CAMEO 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,

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

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, CAMEO 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 CAMEO 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.

CAMEO 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, some of CAMEO'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 CAMEO consortium includes ECMWF, all contractors operating the regional CAMS production system, and additional partners with leading roles in CAMS service contracts. Other CAMEO developments can be transferred to CAMS through future Invitations to Tender, translating them into future tier-1 and tier-2 developments in CAMS.

CAMEO 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. CAMEO'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, CAMEO 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 CAMEO “FAIR” have been included in the budget of the project. In general, the CAMEO 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 CAMEO budget.

7 Data security

The CAMEO 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.

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 CAMEO. Additionally, the Grant Agreement and the CAMEO 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 CAMEO Data Management Plan has been initiated.

Whilst this provides a good starting point for the FAIR data activities of the CAMEO project, it nevertheless needs careful further reflection and updating when appropriate to ensure that new developments (technical as well as strategy) within the CAMEO project and beyond are well reflected by the DMP. The CAMEO 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 CAMEO. 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 CAMEO. 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

Completed by: Angela Benedetti with input from WP partners

Data set	**Dataset of the retrieved aerosol properties from CAMS-based simulations**
Data set description	Dataset will present microphysical and optical properties of aerosol derived from the retrieval of the top-of-atmosphere simulations based on CAMS aerosol model and atmosphere parameters. There are no limitations or usage constraints for the final version of the dataset.
Standards and metadata	The data will be provided in the form of NetCDF files, containing the aerosol optical properties. The dataset will be provided with metadata. The metadata and information related to the dataset will be properly documented in a deliverable report in the project.
Data Sharing	The final dataset is expected to be distributed through the GRASP projects web page (I.e. https://www.grasp-open.com/products/) and will have open access. For direct use in the project, the data will be shared via FTP site with the 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)	GRASP will ensure a local archiving facility for the datasets generated within the CAMEO project.

Workpackage 2

Completed by: Antje Inness with input from WP partners

Dataset	**Global IFS analysis data**
Data set description	Global IFS analysis data based on assimilation tests with a range of new datasets, model and data assimilation configuration developments These are test experiments to document the impact of the assimilation of new satellite data, 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. CAMEO 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 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.

Workpackage 3

Completed by: Vincent Guidard with input from WP partners

Dataset	Regional analysis data from various partners
Data set description	Regional model analysis data, based on assimilation tests with a range of new datasets, model and assimilation configuration developments. These are test experiments to document the impact of the assimilation of new satellite data, 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 regional 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 regional models which are well documented. CAMEO will use their standard chemistry schemes. The models' 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 peer-reviewed publication.
Data Sharing	Data shared within the consortium on request.
Archiving and preservation (including storage and backup)	The model simulations will be archived by each partner at their own archiving facility and will be classified as "publication datasets". The datasets preservation will be reviewed on a regular basis and may differ from one partner to another.

Workpackage 4

Completed by: Marion Schroedter-Homscheidt with input from WP partners

Dataset	**Collection of 1-min measurements from different scientific solar radiation networks (is available and will be further collected)**
Data set description	For the analysis, bias-correction and uncertainty estimation of CAMS Radiation Service, a global dataset of radiation measurements from different scientific radiation networks has been collected. The current dataset includes 183 stations from the following public networks: • ABOM (22) • BSRN(77) • enerMENA (13) • NREL (12) • SAURAN (21) • SOLRAD (9) • SURFRAD (7) • SKYNET (6) • ESMAP (16) The integration of further networks (e.g. SONDA in Brazil, recent dataset published by the World Bank) is also planned. All stations provide 1-min measurements of the global, diffuse and direct radiation over a period ranging from a few years to several decades depending on the station. The data are automatically updated every 15 days. The data is originally aimed to be used for the evaluation of CRS but it is currently also used for developing algorithm for the CAMS UV service. A potential use for activities on aerosol can also ne envisaged. The data have been found in the public domain and can be used for our research activities. The data source should be mentioned in communication and publication using the data. The redistribution of the data is not allowed: the access to the collected data will therefore be limited to the CAMEO team.
Standards and metadata	The metadata will be prepared according to the CF convention and to fulfil the requirements of the ISO 1915, ISO 1939 and INSPIRE standards. Additional information will be added to address common needs of solar practitioner as well as to indicate information on the original data provider. Link to the WMO activities (WIGOS OSCAR) will also ensure by indicating the WMO and WIGOS identifiers when possible. The data format (netCDF structure as well as variable description) has been prepared to be compliant with the CF conventions. The data have successfully passed tools to verify the compliance with the CF standards. A careful plausibility check of the data has been conducted and is currently being improved. This quality control is made of a set of automatic tests and a visual inspection tool. This work has been started in CAMS2-73 and is carried out in cooperation with the expert team of the task 16 of the IEA PVPS program.

	Further details on our activities on standards and metadata on in-situ measurements can be found on https://hal-mines-paristech.archives-ouvertes.fr/OIE/hal-03811628v1
Data Sharing	Each of the station's dataset will be then transformed into a standard and interoperable NetCDF-CF file and according to the use-cases specifications, prepared for data sharing over the network using an open-source application server. In this project, the Thredds Data Server (TDS) will be used; it is a web application server providing metadata and data access supporting the following open and standard protocols including: OPeNDAP (Open-source Project for a Network Data Access Protocol), OGC Web Map Service (WMS), Web Coverage Service (WCS) and HTTP (enabling download). These protocols would allow to perform several key operations including: • Download. • Sub-setting in space and time and across all available data dimensions using NetCDF subset • Local and remote access based on the OpenDAP protocol. Thanks to the functionalities of the Thredds Data Server the data can easily be accessed (with sub-setting functionalities) using a simple URL access. The access to the data is protected by a password. Credentials have been shared to the CAMEO team working with the data. As previously mentioned, the data have been found in the public domain, but data redistribution is generally not allowed. We have therefore limited the access to the CAMEO team. The license is inherited from the original data-providers or network operator. Information about the license but also original provider (institution, operator, contact person) is systematically provided in the metadata. This step will tick the "A" (Accessible) from the FAIR principles. The "I" (Interoperable) will be enabled via the request to TDS from local or remote applications including Jupyter Notebook, API framework (Python) or web-based applications. The data and metadata format will be prepared to be findable by the GEO Discovery and Access Broker (DAB) so that their visibility will be maximised by a reference on the GEO knowledge hub.
Archiving and preservation (including storage and backup)	The in-situ data collected from the local sites and stored in a local database and according to the various use-cases of the project will follow the "FAIR" approach; MINES Paris PSL, member of the Open Geospatial Consortium (OGC), will implement this step. OGC is "an international voluntary consensus standards organization for geospatial content and location-based services, sensor web and Internet of Things, GIS data processing and data sharing." "It supports commercial, governmental, nonprofit and research organizations in a consensus process encouraging development and implementation of open standards with a mission to make location information FAIR – Findable, Accessible, Interoperable, and Reusable". This will first include a data transformation to create datasets per station in a standard NetCDF format improving standardisation and interoperability, to leverage development of new tools. "NetCDF is a set of software libraries and self-describing, machine-independent data

	formats that support the creation, access, and sharing of array-oriented scientific data." MINES Paris - PSL has developed an open-source library, "libinsitu", to transform solar in situ data into a standard NetCDF format. This set of Command Line Interface (CLI) utilities and Python functions allows to: • Transform raw input files into NetCDF format. • Explore & query NetCDF files, and transform them to various formats (CSV, JSON, text, pandas dataframes). • Flag data with quality checks and produce graphs for visual quality control. • Align variable naming following Climate and Forecast (CF) conventions. The libinsitu library is available on https://git.sophia.mines-paristech.fr/oie/libinsitu . This first step will tick the "R" (Reusable) from the FAIR principles. The storage need for the total processing and archiving chain of the data is estimated to 0.5 - 1 To. As previously mentioned, the data are stored on a local database on Mines Paris premises and the TDS server used to share the data among the consortium is also hosted by Mines Paris. This infrastructure has been financed through our participation in CAMS2-73 as well as our own investments.

Dataset	**Archive of 1-min global solar radiation measurements from Météo-France** **(collected)**
Data set description	To conduct a detailed analysis of the error and uncertainty of the CAMS Radiation Service, archive of pyranometric measurements operated by Meteo-France will be used. This dataset includes 1-min measurements of the global solar radiation over several years from more than 150 stations installed in France. The data is originally only destinated for internal use within the CAMS and CAMEO projects and is not intended for dissemination. Any use and publication made with the data is submitted to an explicit agreement of Météo-France. A specific agreement of Météo-France will be needed for not-planned use of the data or prior to any communication based on the data.
Standards and metadata	The metadata will be prepared according to the CF convention and to fulfil the requirements of the ISO 1915, ISO 1939 and INSPIRE standards. Additional information will be added to address common needs of solar practitioner as well as to indicate information on the original data provider. Link to the WMO activities (WIGOS OSCAR) will also ensure by indicating the WMO and WIGOS identifiers when possible. The data format (netCDF structure as well as variable description) has been prepared to be compliant with the CF conventions. The data have successfully passed tools to verify the compliance with the CF standards.

	A careful plausibility check of the data has been conducted and is currently being improved. This quality control is made of a set of automatic tests and a visual inspection tool. This work has been started in CAMS2-73 and is carried out in cooperation with the expert team of the task 16 of the IEA PVPS program. Further details on our activities on standards and metadata on in-situ measurements can be found on https://hal-mines-paristech.archives-ouvertes.fr/OIE/hal-03811628v1
Data Sharing	Each of the station's dataset will be then transformed into a standard and interoperable NetCDF-CF file and according to the use-cases specifications, prepared for data sharing over the network using an open-source application server. In this project, the Thredds Data Server (TDS) will be used; it is a web application server providing metadata and data access supporting the following open and standard protocols including: OPeNDAP (Open-source Project for a Network Data Access Protocol), OGC Web Map Service (WMS), Web Coverage Service (WCS) and HTTP (enabling download). These protocols would allow to perform several key operations including: • Download. • Sub-setting in space and time and across all available data dimensions using NetCDF subset • Local and remote access based on the OpenDAP protocol. Thanks to the functionalities of the Thredds Data Server the data can easily be accessed (with sub-setting functionalities) using a simple URL access. The access to the data is protected by a password. Credentials have been shared to the CAMEO team working with the data. The restriction on the use of the data previously mentioned will be indicated in the license section of the metadata. This step will tick the "A" (Accessible) from the FAIR principles. The "I" (Interoperable) will be enabled via the request to TDS from local or remote applications including Jupyter Notebook, API framework (Python) or web-based applications. The metadata format will be prepared to be findable by the GEO Discovery and Access Broker (DAB) so that their visibility will be maximised by a reference on the GEO knowledge hub.
Archiving and preservation (including storage and backup)	The in-situ data collected from the local sites and stored in a local database and according to the various use-cases of the project will follow the "FAIR" approach; MINES Paris PSL, member of the Open Geospatial Consortium (OGC), will implement this step. OGC is "an international voluntary consensus standards organization for geospatial content and location-based services, sensor web and Internet of Things, GIS data processing and data sharing." "It supports commercial, governmental, nonprofit and research organizations in a consensus process encouraging development and implementation of open standards with a mission to make location information FAIR – Findable, Accessible, Interoperable, and Reusable". This will first include a data transformation to create datasets per station in a standard NetCDF format improving standardisation and interoperability, to leverage development of new tools. "NetCDF is a set

	of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data." MINES Paris - PSL has developed an open-source library, "libinsitu", to transform solar in situ data into a standard NetCDF format. This set of Command Line Interface (CLI) utilities and Python functions allows to: • Transform raw input files into NetCDF format. • Explore & query NetCDF files, and transform them to various formats (CSV, JSON, text, pandas dataframes). • Flag data with quality checks and produce graphs for visual quality control. • Align variable naming following Climate and Forecast (CF) conventions. The libinsitu library is available on https://git.sophia.mines-paristech.fr/oie/libinsitu . This first step will tick the "R" (Reusable) from the FAIR principles. The storage need for the total processing and archiving chain of the data is estimated to 0.5 - 1 To. As previously mentioned, the data are stored on a local database on Mines Paris premises and the TDS server used to share the data among the consortium is also hosted by Mines Paris. This infrastructure has been financed through our participation in CAMS2-73 as well as our own investments.

Dataset	**CAMS Radiation Service time series**
Data set description	The CAMS solar radiation services provide historical values (2004 to present) of global (GHI), direct (BHI) and diffuse (DHI) solar irradiation, as well as direct normal irradiation (BNI). Additionally, an ASCII "expert mode" format can be selected which contains in addition to the irradiation, all the input data used in their calculation (aerosol optical properties, water vapour concentration, etc). Details as in Qu, Z. et al., 2017. Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method, Meteorol. Z., 26, 33-57, doi: 10.1127/metz/2016/0781. Lefèvre, M. et al, 2013. McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. AMT, 6, 2403-2418, doi: 10.5194/amt-6-2403-2013. Gschwind, B., et al., 2019. Improving the McClear model estimating the downwelling solar radiation at ground level in cloud free conditions – McClear-V3., Meteorol. Z., doi:10.1127/metz/2019/0946.
Standards and metadata	The CAMS Radiation Service CSV formatted products are presented as a metadata header section, followed by lines of values (columns). The metadata header section helps the user to understand the data, and describing the various features of the products. These metadata are

	written as text in the delivered file, and conform to the ISO standard where available. Csv files Data is regularly controlled (https://atmosphere.copernicus.eu/supplementary-services#fa6856b7-a306-4cc4-9137-f3e0cb703093)
Data Sharing	Data can be downloaded under the Copernicus data policy via https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-solar-radiation-timeseries?tab=overview. Documentation: https://confluence.ecmwf.int/display/CKB/CAMS+solar+radiation+time-series%3A+data+documentation
Archiving and preservation (including storage and backup)	Data is stored long-term in the Copernicus ADS. Data amount: <100 GB

Dataset	**Cloud products from CAMS processing chain**
Data set description	As described in Schroedter-Homscheidt et al., Surface solar irradiation retrieval from MSG/SEVIRI based on APOLLO Next Generation and HELIOSAT-4 methods, Meteorol. Z./Contrib. Atm. Sci., doi: 10.1127/metz/2022/1131
Standards and metadata	HDF 4 files with metadata internally in the header. Data is regularly used in CAMS Radiation service and quality controlled routinely.
Data Sharing	Consortium internal data share. Data can be made available on request, but is seen as intermediate in the operational CAMS processing chain.
Archiving and preservation (including storage and backup)	Long-term storage on DLR-internal hardware in owned data centers. Data amount: <100 GB for extracted time series

Dataset	**Ultraviolet (UV) radiation (UV index) measurements at surface level**
Data set description	Dataset consists of surface UV index observations. There are a total of 39 stations in the dataset that provide the UV index measurement. The observations have been collected directly from multiple data providers using personal communication. We have agreements with all data providers that allow us to use the data for the project with no cost.
Standards and metadata	Standards and metadata vary station by station.
Data Sharing	The data will not be shared. The agreements with the data providers only allow use of the data for the project but not sharing of the data. Possible inquiries about the data can be sent directly to the station managers responsible for the original data.

Archiving and preservation (including storage and backup)	The data will be archived and preserved at FMI computers at least until the end of the project.

Dataset	**Grimm EDM-164 Particulate Matter**
Data set description	Description: PM1, PM2.5 and PM10 observations collected with the optical Grimm EDM-164 particle counter - Origin: CIEMAT'S Plataforma Solar de Almería (PSA, since 2013), Zagora (since 2020) and Missour (2015-2017) - Nature: 1min resolution observations of particulate matter (units of $\mu g/m^3$) collected with an optical particle counter - Possible users: CAMS community - Scientific publication: Hanrieder (2016), (Wolfertstetter (2016) - Publicly available for research purposes. Commercial uses can be discussed, charges may apply
Standards and metadata	Specifications described in Mol, W. (2013) and EU (2001) ASCII file, Excel, .mat file Quality control based on Spangl, W. (2019) and Aslan, M.E. (2022)
Data Sharing	Data stored in an internal DLR-ISF SQL repository. Available upon demand via personal communication
Archiving and preservation (including storage and backup)	Data is long-term stored in a DLR-ISF SQL database and backed-up in several internal servers Total volume of data: ~50MB/year Volume of distributed data will depend on demand (years, channels, etc.)

Dataset	**Soiling ratio and soiling loss rate observations**
Data set description	- Description: Soiling ratio and soiling loss rate observations - Origin: CIEMAT's PSA from 2017 - Nature: Daily observations with reference cells and DUST IQ sensor - Possible users: CAMS community and solar energy industry - Scientific publication: Norde-Santos (2022) Publicly available for research purposes. Commercial uses can be discussed, charges may apply
Standards and metadata	IEC-61724 ASCII file, Excel, .mat file Data collected following the guidelines described by the IEC-61724 standard and manually quality-checked by operator
Data Sharing	Data stored in an internal DLR-ISF SQL repository. Available upon demand via personal communication

Archiving and preservation (including storage and backup)	Data is long-term stored in a DLR-ISF SQL database and backed-up in several internal servers Total volume of data: ~10 KB/year Volume of distributed data will depend on demand

Dataset	**WP4-WP6 IFS-COMPO ensemble**
Data set description	Medium resolution IFS-COMPO 50-member ensemble forecast simulations (40km and 137 vertical levels) will be performed for periods of 2019. These IFS simulations will include the latest developments in the aerosol and chemistry schemes, as described in the IFS cycle 48R1 documentation. Supplementary simulations perturbing some aerosol and chemistry emissions using the emission uncertainty estimated provided by WP5 will be carried out, also perturbing online primary aerosol emissions. Supplementary simulations also using different deposition schemes and key reaction rates will be carried out. These simulations will provide an estimate of uncertainty in the simulated atmospheric concentration, burden and deposition rates of aerosol and chemical species of interest.
Standards and metadata	The simulations are based on the state-of-the-art NWP model which is very well documented, also using the IFS(CB05) and IFS(AER) modules, which are documented in the IFS cycle 48R1 documentation. The model parameters are based on the WMO standard meteorological parameters and described in grib parameter database (https://apps.ecmwf.int/codes/grib/param-db). The metadata will also be documented in a publicly available deliverable report, including the evaluation simulations with independent data and observations. The data will be available in grib or netCDF formats.
Data Sharing	The data will be publicly available and it will be accessible through the ECMWF API (https://www.ecmwf.int/en/forecasts/accessforecasts/ecmwf-web-api).
Archiving and preservation (including storage and backup)	The IFS simulations will be archived in the mars tapes at ECMWF and they will be classified as “publication datasets” which means they will be preserved for at least 5 years, and after which the dataset preservation will be reviewed.

Workpackage 5

Completed by: Marc Guevara with input from WP partners

Dataset	**Gridded maps of uncertainties for annual global NOx, CO, SO2 and OC emissions per sector**
Data set description	The final dataset will consist of a collection of gridded global and annual emission maps with associated uncertainty. The emission maps will be produced for selected species (NOx, CO, SO2 and OC) and sector (road transport, energy industries, manufacturing industries, and residential/commercial combustion activities). The compiled emission factors for each country, fuel type, and sector used to produce the final emission uncertainty products will be gridded for easy map visualization and analysis before being shared with partners. Furthermore, all this information will be entered into the online tool under development. The web-based tool will also include all the activity data required to estimate emissions. A publication on the development of the online system is foreseen.
Standards and metadata	The data will be provided in the form of NetCDF files, containing gridded emissions and associated uncertainties per country for a few sectors over a few years, as well as CSV files. The dataset will be provided with metadata. The metadata and information related to the dataset will be properly documented in a deliverable report in the project. Comparisons will be made with existing datasets in order to evaluate the quality of the data produced by the online tool.
Data Sharing	The final dataset is expected to be distributed through a dedicated Copernicus CAMS website repository. For direct use in the project, UT3 will share the dataset through its FTP site. The access to the online tool will be shared with the partners of the project when it will be operational. The tool will be made usable by a general audience once the project is completed
Archiving and preservation (including storage and backup)	All the datasets generated by the tool 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.

Dataset	**Gridded maps of uncertainties for annual European primary PM emissions per sector**
Data set description	The dataset will (mostly likely, pending discussions with the main users) consist of a family of 10 gridded emission datasets. It will be generated by TNO based on the CAMS-REG methodology, combined with information on the uncertainty of emissions. It will be useful to regional air quality modellers who want to assess the impact of emission uncertainty on the modelling results. Similar data were generated before for other species (CO2, CO) and described in a scientific paper (Super et al., 2020) Not all correlations and dependencies are currently taken into account, hence the dataset should be regarded as a first approximation of the uncertainty in PM emissions. Because of the uncertainties, the dataset should be used with caution. It will be made available to all interested in the different work packages of the CAMEO project.
Standards and metadata	Reference to CAMS-REG emission inventory for Europe, which is documented in Kuenen et al. (2022) The metadata and information related to the dataset will be properly documented in a deliverable report in the project. Additionally, a readme file will be provided along with the dataset to provide practical information on how to interpret and use the dataset. Data will be provided in CSV and netcdf format like the regular CAMS-REG emission datasets Like for the regular CAMS-REG emissions, several QA/QC steps will be performed.
Data Sharing	The dataset is expected to be distributed through a dedicated Copernicus CAMS website repository. For direct use in the project, TNO will share the dataset through its FTP site.
Archiving and preservation (including storage and backup)	The dataset is expected to be archived in a dedicated Copernicus CAMS website repository, as mentioned above, providing also a DOI that will be assigned by ECMWF.

Dataset	Ensemble of estimates of monthly, weekly, daily and hourly temporal profiles
Data set description	The dataset will consist of an ensemble of monthly (month-of-the-year), weekly (day-of-the-week), daily (day-of-the-year) and hourly (hour-of-the.day) emission temporal profiles for a selection of sectors, including energy industry, residential and commercial combustion activities and road transport. The profiles will be provided per sector and country. The temporal profiles will be constructed making use of several public databases, including national electricity generation statistics, TomTom congestion statistics and ECMWF Reanalysis (ERA5) meteorological data, among others. The methodology to construct the profiles will follow the approach proposed for the construction of the CAMS-TEMPO database (Guevara et al., 2021[4]). The final dataset will also include statistics and information that defines the spread of the data and probability distributions associated to each temporal weight factor, so it can be used for uncertainty propagation. Users are potentially atmospheric chemistry and inversion modellers, as well as emission inventory developers. The sources of information considered to construct the ensemble of profiles present limitations in terms of spatial coverage (e.g., TomTom congestion statistics does not report information for China) The quality of the final product is determined by the quality of the data sources and estimation methodologies considered.
Standards and metadata	Reference to the CAMS-TEMPO database and documentation of the original database. A proper report and metadata file will be produced. The metadata file will consist of a TXT README file provided together with the final datasets, which will include a detailed description of the different files and fields of information included in them. The ensemble of profiles and associated statistics will be provided in CSV format 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 previous CAMS-related daily emission datasets (e.g., https://atmosphere.copernicus.eu/node/731). In the short term, it will be shared among CAMEO 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. The dataset is expected to be licensed under CC BY 4.0 license. The CAMS website repository will have an URL access

[4] Guevara, M., Jorba, O., Tena, C., Denier van der Gon, H., Kuenen, J., Elguindi, N., Darras, S., Granier, C., and Pérez García-Pando, C.: Copernicus Atmosphere Monitoring Service TEMPOral profiles (CAMS-TEMPO): global and European emission temporal profile maps for atmospheric chemistry modelling, Earth Syst. Sci. Data, 13, 367–404, https://doi.org/10.5194/essd-13-367-2021, 2021.

Archiving and preservation (including storage and backup)	The dataset will be archived in a dedicated Copernicus CAMS website repository, as mentioned above, providing also a DOI that will be assigned by ECMWF. The volume of the data will depend on the amount of temporal profiles constructed, each of them taking approximately between 10-50kb.

Dataset	**Gridded maps of monthly global isoprene emissions with range of uncertainty**
Data set description	The dataset will consist of global gridded maps containing information of an uncertainty range of biogenic isoprene emissions. The maps will be provided with monthly temporal resolution for at least 2 years with distinct climatology. Each grid cell will include the mean isoprene emission from the CAMS-GLOB-BIO dataset (Sindelarova et al., 2022[5]) together with high and low emission values at a given confidence interval. The gridded maps will be provided with 0.25°x0.25° horizontal spatial resolution. The high and low emission values will be obtained from simulations of the MEGANv2.1 emission model based on uncertainty of different model input parameters. The potential users are atmospheric chemistry and inversion modellers, as well as emission dataset developers. The horizontal spatial resolution of the dataset is limited with spatial resolution of the input parameters (e.g. ERA5 meteorology). The temporal resolution is limited with computational and storage costs. No usage constraints are known at the moment.
Standards and metadata	Reference to the CAMS-GLOB-BIO inventory development (Sindelarova et al., 2022[2]) and report describing estimation of isoprene emission uncertainty range. A report with detailed description of methodology of isoprene emission uncertainty estimation will be released together with the gridded maps. Short description will be included in the attributes of the resulting netCDF files. NetCDF files including global monthly gridded fields with isoprene mean emissions, high and low emissions at given confidence level. There will be a review process to quality-check the data.

[5] Sindelarova, K., Markova, J., Simpson, D., Huszar, P., Karlicky, J., Darras, S., and Granier, C.: High-resolution biogenic global emission inventory for the time period 2000–2019 for air quality modelling, Earth Syst. Sci. Data, 14, 251–270, https://doi.org/10.5194/essd-14-251-2022, 2022.

Data Sharing	The data is expected to be shared through the publicly accessible Atmosphere Data Store (ADS, https://ads.atmosphere.copernicus.eu/cdsapp#!/home) managed by ECMWF, similarly to the CAMS-GLOB-BIO inventory. The dataset will be publicly available, and no software will be necessary to enable re-use. Before the upload to ADS, the data can be shared to partners in the project through a Cloud repository. The dataset will be assigned a unique DOI. The dataset is expected to be licensed under CC BY 4.0 license. The ADS website will provide a URL of the dataset.
Archiving and preservation (including storage and backup)	As mentioned above, the dataset is expected to be stored on the Atmosphere Data Store (ADS) and to be assigned with a unique DOI, similarly as the DOI was assigned to the CAMS-GLOB-BIO dataset within the CAMS project. One file with monthly mean isoprene emissions and uncertainty range is approximately 2.5 MB in size. The two years of data will be around 60 MB in size.

Dataset	**Observation-based NOx emission estimates**
Data set description	Gridded NOx emissions calculated with the DECSO algorithm for the various regions in the world. The DECSO algorithm derives NOx emissions using TROPOMI NO2 observations and the chemical transport model CHIMERE. Resolution will be 0.2 x 0.2 degree. Every major upgrade of the algorithm is accompanied by a scientific publication
Standards and metadata	The data format will be netcdf. Our data will be accompanied by metadata following the CF metadata convention. Maps of the data will be made and compared to earlier versions of the algorithm for quality check.
Data Sharing	The produced data will be provided as open data. The data will be made available via the GlobEmission web-site www.globemission.eu
Archiving and preservation (including storage and backup)	The data will be archived both on the GlobEmission web-site and the IT facilities of SURF (www.surf.nl). Note that the data of the GlobEmission web site is hosted on AWS (S3).

Dataset	Observation-based CO_2 and CH_4 emission estimates
Data set description	The CO2 point source datasets include longitude, latitude, day of emission, magnitude of emission inverted from OCO2 and OCO3 satellites. Currently about 300 sources, updated regularly with new measurements Limitations. Coverage limited by clouds. Snapshot estimates, only large point sources The CH4 point source datasets include longitude, latitude, day of emissions include longitude, latitude, day of emission, magnitude of emission from TROPOMI analysis. About 3000 sources, updated regularly with new measurements Limitations. Coverage limited by clouds. Snapshot estimates, only large point sources > 20 tCH4 h-1 The CH4 large area source datasets include shapefile, month of emissions magnitude of emission from TROPOMI analysis. About 9 basins, updated regularly with new measurements Limitations. Coverage limited by clouds
Standards and metadata	Meta data will include units, formats and reference to methodology used (if versions change, this will be updated) Metadata will be in the header of the files Data format will be text files csv QAQC from the inversion algorithm will be used to quality-check the data
Data Sharing	The unpublished data will be available for WP5 and to other members of the consortium upon request. License N/A Access URL from LSCE sharebox https://sharebox.lsce.ipsl.fr/index.php/login
Archiving and preservation (including storage and backup)	Data will be long term archived on the LSCE server (shared file system) Current standards of CEA for data archive

Workpackage 6

Completed by: Hilde Fagerli with input from WP partners

Dataset	**IFS-COMPO ensemble simulations**
Data set description	Medium resolution IFS-COMPO 50-member ensemble forecast simulations (40km and 137 vertical levels) will be performed for periods of 2019. These IFS simulations will include the latest developments in the aerosol and chemistry schemes, as described in the IFS cycle 48R1 documentation. Supplementary simulations perturbing some aerosol and chemistry emissions using the emission uncertainty estimated provided by WP5 will be carried out, also perturbing online primary aerosol emissions. Supplementary simulations also using different deposition schemes and key reaction rates will be carried out. These simulations will provide an estimate of uncertainty in the simulated atmospheric concentration, burden and deposition rates of aerosol and chemical species of interest.
Standards and metadata	The simulations are based on the state-of-the-art NWP model which is very well documented, also using the IFS(CB05) and IFS(AER) modules, which are documented in the IFS cycle 48R1 documentation. The model parameters are based on the WMO standard meteorological parameters and described in grib parameter database (https://apps.ecmwf.int/codes/grib/param-db). The metadata will also be documented in a publicly available deliverable report, including the evaluation simulations with independent data and observations. The data will be available in grib or netCDF formats.
Data Sharing	The data will be publicly available and it will be accessible through the ECMWF API (https://www.ecmwf.int/en/forecasts/accessforecasts/ecmwf-web-api).
Archiving and preservation (including storage and backup)	The IFS simulations will be archived in the mars data archive 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.

Dataset	**Modelled source receptor data (for cities and sector/sources)**
Data set description	This data set comprises modelled source receptor data for cities and sources generated with - EMEP MSC-W model - LOTOS EUROS model - CHIMERE CAMS-ACT model Data will cover Europe and cover the period 2015-2019 (for some models less year). Data will be useful for scientists involved in CAMEO WP6
Standards and metadata	Data will be provided in json format for the country-to-cities and country-to-country contributions. The format of the source contribution data has not been decided yet, but it will most probably be netcdf.

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)	MET Norway will ensure local archiving of EMEP MSC-W model data, TNO and INERIS will do the same for LOTOS EUROS and CHIMERE - ACT model data, respectively.

Dataset	**Source apportionment data from observations**
Data set description	A collection of observational based source attribution data from several locations and time periods has been collected from different sources. ● PM10 PMF data has been obtained through private communication with data providers. Currently, this covers datasets from Spain, Italy, France, Netherlands, Switzerland, Greece, Germany, Belgium. The datasets are for different time periods. The retrieved source factors differ per location. Most of the datasets have been presented in scientific publications. The data are allowed to be used within the context of the CAMEO project but are not allowed to be further distributed. In case of publication the data providers/owners need to be asked for approval and co-authorship should be offered. In all cases, an acknowledgement must be made to the data providers or owners. ● eBC PMF data has been obtained from the EMEP/ACTRIS/COLOSSAL field campaign in winter 2017/2018 from the NILU team. This dataset contains eBC differentiated between biomass burning and fossil fuel sources from 57 locations over Europe. The dataset also contains co-located levoglucosan data. The dataset is described in Platt et al., 2023 (in preparation). The data are available free of charge for non-commercial and scientific use. An offer of co-authorship needs to be made to the data providers/owners whenever substantial use is made of their data. In all cases, an acknowledgement must be made to the data providers or owners and to the project (“The EMEP intensive measurement campaign, winter 2017-2018”) when these data are used in a publication. When used, the dataset should be cited using the Digital Object Identifier (DOI). https://doi.org/10.21336/gen.h8ds-8596 ● Data for specific tracers and PM composition has been retrieved from the EBAS system. This includes Levoglucosan (tracer for biomass burning), Nickel and Vanadium (tracer for Heavy fuel oil), Natrium (seasalt), EC and OC.

	For a more detailed overview of the datasets we refer to the excel sheet describing the data. (A first version will be made available by TNO in July 2023). This excel sheet may be updated in the course of the CAMEO project when new datasets become available which can be included in the collection.
Standards and metadata	We will not create metadata for the above mentioned data collection, because of the large diversity in dataset contents and because the dataset is not made for sharing because of limitations on its use by data providers. The datasets will be converted to a standard format (format to be decided, but likely to be xlsx. or csv. format) for facilitating comparison with model results. The data will be visually analysed as a basic review process to check quality, no extensive review process is foreseen.
Data Sharing	The dataset cannot be shared because of its origin from a large variety of data providers. The data providers need to be contacted individually for each use outside of the CAMEO project
Archiving and preservation (including storage and backup)	The dataset collection will be stored on TNO's high performance cluster system.

Document History

Version	Author(s)	Date	Changes
0.1	Rhona Phipps, Tanya Warnaars, Johannes Flemming, CAMEO project partners	June 2023	Initial version
0.2	Rhona Phipps, Tanya Warnaars, Johannes Flemming, CAMEO project partners	June 2023	Removed advisory text and included annex 2 for internal project reviewers
1.0	Rhona Phipps, Tanya Warnaars, Johannes Flemming, CAMEO project partners	04 July 2023	Issued

Internal Review History

Internal Reviewers	Date	Comments
Anne Caroline Lange (FZJ)	June 2023	As per mark-ups in document.