Better Contrails Mitigation - BeCoM

D7.3

Initial Data Management Plan

Delivery Date:	30/11/2022(M6)
Date of submission:	20/12/2022

LEAD BENEFICIARY FOR THIS DELIVERABLE			
Name:	TU Delft		
Contact Person:	Feijia Yin		
E-mail:	f.yin@tudelft.nl		
AUTHORS			
Participants:	All partners		
Work Package:	7 - Project Management and Scientific Coordination		
Version:	V0.1	Number of Pages:	18

REVISION HISTORY			
VERSION	DATE	AUTHOR / REVIEWER	NOTES
0.1	20/12/2022	All partners	

REVIEWED AND SIGNED OFF BY			
ROLE	DATE	NAME	SIGNATURE

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

Table of contents

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

1 Will you re-use any existing data and what will you re-use it for?

BeCoM will be mainly reusing data delivered by the different partners within the collaboration. Hence this data is commercially sensitive, shared among collaborators for the purposes of this project only. The data correspond to humidity measurements which will be used to train the AI algorithms to detect and predict contrails based on weather forecast.

The project will also reuse the software Modular Earth Sub model System (MESSy; which allows the implementation of the EMAC Earth-System model to simulate air traffic and contrail formation) and Radiative Transfer for TOVS (RRTOV; which allows the simulation satellite radiances). MESSy is a software available to MESSy Consortium partners under a Memorandum Of Understanding and the corresponding MESSy Software License Agreement described in https://www.messy-interface.org/ RTTOV is an open-source code available upon request (https://nwp-saf.eumetsat.int/site/software/rttov/)

1.1 What types and formats of data will the project generate or re-use?

BeCoM will make use of pre-existing data (provided by partner institutions) and will generate data and code which will be suitable for reuse. The datasets that will be collected/generated will be:
*** Observational data e.g., measurements of temperature, cloud vertical structure and humidity level in the atmosphere. This data will be obtained from satellite and ground-based instrumentation. The aim of collecting/generating this data is to obtain a larger and higher resolution database of relative humidity at cruise levels for revisiting local contrail formation conditions and for testing their assimilation into Numerical Weather Prediction (NWP) models. Part of the measurements will also serve as the training dataset for the AI algorithms that will be developed for the contrail prediction. Format: NetCDF.**
*** Simulation code** e.g., NWP models with improved computation of ice cloud microphysics. Format: open/standard programming languages such as FORTRAN and Python. Currently also the C language is used for programming box model experiments, with the intention to switch to the ICON language, once the experiments within ICON will take place.
Perhaps we could say, the numerical procedures will be published in the form of manuals, papers, and theses.
*** Analysis code** e.g., AI algorithms trained for contrail detection and contrails risk prediction using proper data quality checks as part of the data assimilation process. Format: open/standard programming languages such as FORTRAN and Python.
*** Documents** e.g. recommendations for regulatory measures for effective contrail avoidance. Format: text (e.g. ASCII, .pdf).

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

1.2 What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The overall aim of BeCoM is to predict the vast weather-induced variability of the radiative effect of individual contrails, and formulate adequate mitigation measures and develop policy-driven implementation schemes to enable Air Traffic Management to reduce aviation's climate impact. In order to do this **observational data** will be used to obtain a higher resolution database of humidity at cruise levels for assimilation into Numerical Weather Prediction (NWP) models, and serve as the training dataset for the **AI algorithms** that will be developed for the contrail prediction.
The **simulation data/code** will be used to derive an improved ice cloud microphysics model and estimate optimized flight trajectories regarding contrail formation and consequent climate effects.

1.3 What is the expected size of the data that you intend to generate or re-use?

BeCoM will collect pre-existing and new observational data from aircraft, satellites, and ground-based sources. BeCoM will also deliver algorithms, analysis scripts, and documentation files related to methodologies and reports. The total amount of data is unknown, but given its nature (specially the observational data), it is expected to be in the order of 10s of TBs.

1.4 What is the origin/provenance of the data, either generated or re-used?

Each collected dataset will be properly documented with both embedded (e.g. file headers) and supporting documentation (e.g. README files). A top-level documentation file (README) in an open format (such as .txt or .md) will contain information about:

- purpose of data collection/generation
- how it was collected/generated (including experimental settings, facilities, protocols, and methodologies)
- software used to collect/generate, process and visualize the data (including versions, dependencies, etc.)
- structure of the project (file naming conventions, directory tree structure)
- headers and units (if not present in the files already)
- references, caveats and assumptions

In the case of code scripts developed within the project, researchers will be encouraged to work with a version control software, pay attention to the readability of the code as well as its structure. A top-level README (.txt or .md) file will also include information about:

- what the code does
- how to compile/run the scripts
- requirements (versions, libraries, dependencies, environment setup)
- Instructions or digital notebook (e.g. Jupyter notebooks) exemplifying how the code works with a small benchmark dataset

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

1.5 To whom might your data be useful ('data utility'), outside your project?

BeCoM will work towards a persistent contrail forecasting, which will ultimately enable the aviation technology sector to make informed design selections towards sustainable aviation. The results of BeCoM will improve the forecast of Ice Super Saturated Regions (ISSRs), which is critical for predicting persistent contrail formation, and thus allowing further research and development of climate-friendly technologies, operations, and regulatory measures.

In order to maximize the reuse of the data/code underlying the results generated in this project, data/code will be shared as open as possible and as closed as necessary. Data/code that can be published (i.e. not commercially constrained by IP) will be released via public archives like 4TU.ResearchData or Zenodo. Data/code that will not be published will be kept in-house at the corresponding owner institution, to be re-used by its employees and collaborators.

Regarding the utility of the data, the methodology behind the harmonization of the observational database will be useful for researchers working on interoperability of ground and space based observations. The observational database itself will be useful for other studies focused on climate effects (e.g. from different pollutant agents).

The climate-cost optimized flight paths for reduced contrails will be useful to researchers for benchmarking purposes, as well as projects working towards more sustainable air traffic management.

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

2 FAIR data

2.1 Making data findable, including provisions for metadata: Will data be identified by a persistent identifier?

Data/code can be publicly released will be published via the 4TU.ResearchData or Zenodo. Both are trusted data archives who ensure the published datasets will be findable and accessible in the long-term (>10 years from the publication date). These archives assign a persistent identifier (Digital Object Identifier DOI), which makes the datasets citable and persistently available. The 4TU.ResearchData archive in particular also uses OPeNDAP for NetCDF data which facilitates the findability and inspection of NetCDF records.

Data/code that will not be publicly released (i.e. commercially sensitive data) will be stored within the respective owner's institutional archive, accessible to the employees of the respective institution.

2.2 Making data findable, including provisions for metadata: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

During the project, metadata will be maintained and updated in the NetCDF files following CF metadata convention (https://cfconventions.org/). The simulation code also provides output following this convention.

For the publishing of the data/code, both publishing archives -4TU.ResearchData and Zenodo- use standard citation metadata compatible with Dublin Core.

2.3 Making data findable, including provisions for metadata: Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Relevant keywords following the Library of Congress subject Headings (LCSH) will be added in the citation metadata of the 4TU.ResearchData/Zenodo repositories.

Making data findable, including provisions for metadata: Will metadata be offered in such a way that it can be harvested and indexed?

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

Datasets that can be publicly released will be published via the 4TU.ResearchData or Zenodo. Both are non-discipline specific repositories and use a general citation metadata standard embedded in schema.org. This means the datasets will be indexed in Google Dataset Search.

2.4 Making data accessible - Repository: Will the data be deposited in a trusted repository?

Datasets that can be publicly released will be published via the 4TU.ResearchData or Zenodo. Both are trusted archives which ensure the archived datasets are findable and accessible in the long-term (>10 years from publication date). In addition to that, both allow for private and public repositories. For the purposes of this project, the datasets that can be publicly released will be archived in public repositories (without any embargo periods).

2.5 Making data accessible - Repository: Have you explored appropriate arrangements with the identified repository where your data will be deposited?

As mentioned before, two archives will be used: the 4TU.ResearchData and Zenodo.
The 4TU.ResearchData will be used for large datasets and/or publication of processed NetCDF files, since TU Delft researchers can upload up to 1 TB of data per year free-of-charge and the 4TU.ResearchData archive has extra features for the discoverability and reuse of NetCDF data (i.e. OPeNDAP).

Zenodo on the other hand will be used for the publication of smaller datasets ~50 GB max (free-of-charge) and datasets where no TU Delft researcher is involved in its generation/collection.
Regarding special arrangements with the archives themselves, we do not foresee any. This Data Management Plan has been drafted with the support of the Data Steward of the Faculty of Aerospace Engineering of the TU Delft.

The Data Steward works closely with the 4TU.ResearchData archive and has informed us about the policies and procedures required for data publication. If needed, the Faculty Data Steward will refer us to professionals at 4TU.ResearchData to discuss possible additional arrangements.

2.6 Making data accessible - Repository: Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Both the 4TU.ResearchData and Zenodo ensure the archived datasets are assigned an identifier (DOI).

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

2.7 Making data accessible - Data:

2.7.1 Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Not all data will be made publicly available. Data/code that will remain close is the commercially sensitive data (e.g., constrained by IP). Thus such data will be shared among relevant collaborators for the purposes of this project only. This data include the raw observational data, simulation code and part of the simulation data.

Data that will be publicly released via the 4TU.ResearchData or Zenodo are the processed observational database and part of the simulation data including some of the optimized trajectories which can be used to exemplify the results of the project.

Regarding the AI algorithms, digital (e.g., Jupyter) notebooks will be publicly released via 4TU.ResearchData/Zenodo-Github integration, exemplifying the prediction accuracy of the derived models.

In addition, not all data that we use need to be stored by consortium partners. Satellite data are usually stored by EUMETSAT or NOAA or similar institutions, NWP data are stored by the weather centres (not necessarily for 10 years).
Only data produced by the project partners will be addressed in this Data Management Plan.

2.7.2 If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

We do not foresee any embargo regarding the data/code collected/generated in this project.

2.7.3 Will the data be accessible through a free and standardized access protocol?

4TU.ResearchData uses the Hypertext Transfer Protocol Secure (HTTPS) protocol which is based on TCP/IP, which is an open protocol for most of the internet.

Zenodo uses HTTP and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) mechanism for repository interoperability.

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

2.7.4 If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

There will not be any restrictions on use for the datasets that will be publicly released via 4TU.ResearchData or Zenodo.

Regarding the datasets that will remain at the respective owner institution, contact points will be provided in this Data Management Plan in case access is required (e.g. regulatory purposes).

2.7.5 How will the identity of the person accessing the data be ascertained?

Identification of the person accessing the data is not necessary for the public repositories of the 4TU.ResearchData nor for Zenodo.

Regarding the datasets that are commercially sensitive, the respective owner institution has the right to decide under which circumstances access will be given to others if requested.

2.7.6 Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Non-applicable.

2.8 Making data accessible - Metadata:

2.8.1 Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

For the datasets that will be publicly released, the metadata is openly available and can be accessed directly from the dataset repository's landing page.

For the datasets that will not be publicly released, information about the data will be presented in the related articles (which will be published following the HE Open Science policy).

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

2.8.2 How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

The published datasets and the respectivee (citation) metadata will be preserved for > 10 years from the end of the publication data.

Datasets that will not be published will also be archived for the long-term following the owner institution archiving policies.

2.8.3 Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

The simulation code (e.g., EMAC/AirTraf) is available upon request, but the related documentation will be shared publicly for others to replicate the procedure/methodology.

Analysis scripts (used to e.g., to process or visualize the data) and digital notebooks showcasing the studied AI algorithms will be released publicly via 4TU.ResearchData/Zenodo-Github integration.

2.9 Making data interoperable:

2.9.1 What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

One of the research outputs of this project is the observational database which will collect and make interoperable measurement data coming from different instruments. Depending on the data format, the database will be preferably in NetCDF format following CF conventions as mentioned in a previous answer.

2.9.2 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

We will use common terminology used in the field to describe the data. No new ontologies nor vocabularies will be created.

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

2.9.3 Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

[1]A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/

For each published dataset, any other related datasets (e.g. code repository) will be tagged (via the DOIs) in the References citation metadata item of the repository. This also includes the DOI of the related article(s). Likewise, the DOI of the dataset will be added to the related article(s) in a Data Availability Statement or section alike.

2.10 Increase data re-use:

2.10.1 How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

All documentation needed to validate data analysis and facilitate data re-use will accompany the data via a README file as described in a previous answer.

2.10.2 Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

In principle we do not foresee the need to add specific restrictions to the published datasets. Thus data that will publicly release will be published under an open-content license such as a CC-BY. Code that will publicly release will be published under an open-source license such as MIT or Apache 2.0.

2.10.3 Will the data produced in the project be useable by third parties, in particular after the end of the project?

All data/code produced in this project will be useable by third parties, either publicly (datasets published via 4TU.ResearchData or Zenodo) or privately (commercially sensitive data).

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

2.10.4 Will the provenance of the data be thoroughly documented using the appropriate standards?

Each collected dataset will be properly documented with both embedded (e.g. file headers) and supporting documentation (e.g. README files) as explained in a previous answer. This will be done regardless if the datasets are published or not, as commercially sensitive data will also be reusable within the respective owner institution.

2.10.5 Describe all relevant data quality assurance processes.

The following checklist will be used as a reference for data quality:

- datasets should be deposited in a proper archive if published, or in a secure institutional storage if not published.
- Datasets should be in proper directory (tree) structure with consistent file/directory naming convention. Basic rules:
  - do not use only numbers to name files
  - avoid use of whites paces in file/directory names
  - separate data files from code files (e.g. in different sub directories)
  - distinguish datasets per processing level (raw data, processed data, visualizations, finalized data)
- Datasets must have at least a README file with basic structure: title, authors (and ORCIDs), description of data (units, instrumentation, software, etc.), description of the directory (tree structure and naming convention), caveats, references, copyright and license.

In addition to that, for published datasets:

- citation metadata items (such as authors, description, title, etc.) should describe the dataset accordingly.
- Citation metadata keywords should follow a standard vocabulary (e.g. LCSH).
- Uploaded files should not be corrupted.
- All files mentioned in the README should be present in the repository.
- There should be a license file.
- The DOI of the related article(s) and dataset(s) should be added in the citation metadata. Likewise, the DOI of the dataset should be added in the related article(s) and dataset(s).

2.10.6 Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

See the answers to the following questions.

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

3 Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

All digital outputs will be documented and stored accordingly. This includes data and code (incl. models). The management of these outputs is described in the answers of this DMP.
There will not be physical research outputs.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

Non-applicable.

4 Allocation of resources

4.1 What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

All partner institutions already provide the necessary infrastructure for the data/code management related to this project. This includes:

- software and processing power required to carry out the simulations;
- version control software;
- data storage during the project (institutionally managed servers);
- data archiving in public archives (TU Delft researchers can upload up to 1 TB of data to the 4TU.ResearchData per year free of charge, and Zenodo allows uploads of up to 50 GB each free-of-charge);
- data archiving for the data that will remain under closed-access due to commercial reasons (institutionally managed servers).

Thus no costs have been considered regarding research data/code management services/infrastructure.

Regarding responsibilities, researchers will manage the data related to their respective Work Packages (no need for a Data Manager or position alike).

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

4.2 How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Non-applicable

4.3 Who will be responsible for data management in your project?

Each partner institution will be responsible for the proper management of the data and code they generate/collect. The following table lists the team members who will be overseeing the data/code management at their respective institutions:

TU Delft	Heather Andrews Mancilla Feijia Yin
Deutscher Wetterdienst	Alexander Cress
Deutsches Zentrum für Luft - und Raumfahrt EV	Klaus Gierens
Thales	Teodora Petrisor
Centre National de la Recherché Scientifique CNRS	Philippe Keckhut
ENVISA SAS	Antoine Berthier
ECATS International Association AISBL	Simon Blakey
College of Medical and Dental Sciences, University of Birmingham	

All researchers will be encouraged to request advice to their respective support staff in case necessary.

4.4 How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

> 10 years as mentioned in previous answers.

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

5 Data security

5.1 What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

During the project, researchers will be actively working on the data/code in work laptops/stations provided by the respective institution and remotely maintained by the respective ICT Department. Researchers will be encouraged to make master copies of sensitive data in institutional drives, accessed remotely via secure protocols (e.g., SFTP). This will ensure sensitive data is backed up on a daily basis following the respective institution's ICT security protocols.

Data will be exchanged between partners via institutionally recommended file sharing services. Available services at partner institutions are: SurfDrive, SURFfilesender, XYZ.
For the development of code, researchers will be encouraged to use version control and remote repositories in institutional instances such as the TU Delft Gitlab, the DLR Gitlab, ABC.
After the end of the project, public datasets will be published via the 4TU.ResearchData or Zenodo where the data will be openly accessible to all.

6 Ethics

6.1 Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

No ethics nor legal issues that can have an impact on data sharing are foreseen for this project. No experiments will be done on humans or animals, and no personal data from human participants will be collected nor processed. Results of this project are intended exclusively for civil application.
The AI algorithms developed within BeCoM will be used for two purposes:

- Data assimilation in the numerical weather forecast system
- Persistent contrails detection in satellite image and classification

BeCoM has a specific task in Work Package 3 related to AI algorithms validation, where a thorough validation process will be performed to address the Trustworthiness aspects of using AI. The AI-based assimilation will help enhance the numerical weather prediction models and the AI-based contrail detection is an effective means to distinguish aviation contrails from natural clouds. These improvements eventually will improve our confidence in contrail prediction, hence beneficial to reduce environmental impact of aviation.

6.2 Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

Non-applicable

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885

7 Other issues

7.1 Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

BeCoM is committed to comply with the funder's research data management requirements of FAIR data, following the aim to publish as open as possible and as closed as necessary. Using that as a framework, the following policies regarding data/code management will also apply at the respective partner institutions:

TU Delft	TU Delft Research Data Framework policy TU Delft Research Software policy TU Delft Faculty of Aerospace Engineering Research Data Management Policy
Deutscher Wetterdienst	
Deutsches Zentrum fur Luft - und Raumfahrt EV	Research data management at DLR Guideline to Software-Engineering
Thales	
Centre National de la Recherché Scientifique CNRS	
ENVISA SAS	
ECATS International Association AISBL	
College of Medical and Dental Sciences, University of Birmingham	

The project will be conducted in line with the Netherlands Code of Conduct for Research Integrity which contains a framework for good research practice.

This project has received funding from the European Union's Horizon Europe Research and innovation program under Grant Agreement No 101056885