Better Contrails Mitigation - BeCoM

European Union's Horizon Europe Research and innovation program Under Grant Agreement No 101056885

D3.1

Description of training and testing datasets

Delivery Date:	30/11/2023 (M18)
Date of submission:	18/01/2024

LEAD BENEFICIARY FOR THIS DELIVERABLE	
Name:	Thales
Contact Person:	Teodora Petrisor
E-mail:	teodora.petrisor@thalesgroup.com

AUTHORS			
Participants:	Teodora Petrisor, Celine Reverdy (Thales), Matthias Mages (DWD), Klaus Gierens (DLR)		
Work Package:	WP3		
Version:	1	Number of Pages:	17

REVISION HISTORY			
VERSION	DATE	AUTHOR / REVIEWER	NOTES
0.1	10-11-2023	Teodora Petrisor	Initial draft
0.2	14-11-2023	DWD	Added DWD contributions on ground and satellite data
0.3	15-12-2023	Thales/DLR	Added DLR satellite data description
0.4	14-01-2024	Teodora Petrisor	Version ready for final review
1.0	18-01-2024	Teodora Petrisor	Integrated internal reviewers' comments

REVIEWED AND SIGNED OFF BY			
ROLE	DATE	NAME	SIGNATURE
Project Coordinator	18-01-2024	Oana Trifan	
WP1 Lead	18-01-2024	Philippe Keckhut	
WP2 Lead	18-01-2024	Alexander Cress	

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

Executive Summary

Observational data and in particular different kind of imagers can be used to verify and validate contrail predictions.

This report, highlighting the progress in WP3 during the first project period, presents the relevant data sources for the development of advanced AI algorithms for data assimilation and contrail classification and detection. The overarching goal is to reduce uncertainties in contrail prediction models by making use of available observations.

Data sources identification was the object of the first milestone in Work package 3, achieved in month 12 and joined in Appendix A for completeness.

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

List of figures

List of tables

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

Table of content

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

1 Introduction

1.1 Context and overall objectives

Condensation trails (contrails) left behind by aircrafts in the upper troposphere at cruising altitudes, between approximately 7km and 12 km, are a subject of growing scientific interest in recent years due to their non-negligible effect on climate change. To study their link to radiative forcing, partners in the project and other researchers developed several models based on the microphysics of the atmosphere and cloud formation. Contrail formation can thus be predicted with simulator models such as CoCIP [13]. However, contrail prediction is usually subject to very high uncertainty levels due to the lack of precision in sensors at the corresponding altitudes, as well as to the very high complexity of the physical models (turbulence flows, ice formation according to fuel/engine type, wind shear, stratification, optical depth and many more…).

BeCoM aims to develop and assess measures to largely reduce the overall impact of aviation on climate through achievable actions on a short-time horizon. Among the non-CO2 effects of aviation, the main focus is on the contrails factor, which we aim to better predict, verify through observations and assess in terms of their impact on climate.

We aim to enhance the reliability of persistent contrails forecast so as to reduce individual contrail radiative effects through successful avoidance planning. In particular strongly warming contrails should be avoided via global-climate-aware trajectory optimization techniques.

The scientific and technical work in BeCoM is structured to answer the following objectives, among which the current report focuses on **Objective #3**:

Objective #1: enhance the routine measurements of atmospheric humidity at the cruise altitude.

Objective #2: improve the treatment of ice supersaturated regions in numerical weather prediction models and the use of new and unique data sources and methods in the data assimilation process.

Objective #3: **develop appropriate AI algorithms** for data assimilation, contrail detection, contrails classification, and uncertainties of contrail prediction, by using the high diversity of available observations (ground images, LIDAR, satellite etc.).

- Identify adequate data sources for AI training
- Develop appropriate AI algorithms trained using a variety of datasets and prior knowledge from physical models.
- Evaluate the contrails prediction models through comparison with empirical observations concerning contrails localization, dynamics, and persistency from measurements.

Objective #4: minimize cost impact when implementing climate optimized trajectories.

Objective #5: Develop and evaluate non-CO2-based measures to be applied for ATM strategies for climate impact mitigation.

This deliverable focuses on the identification and description of data sources that can be used in the development of AI algorithms, with a focus on openly available data to increase future usage.

1.2 WP3 Advanced AI algorithms and evaluation of the data and model predictions

Predicted contrails can be verified in past collected data through different sources of imagery, called observations. Combining relevant information such as contrails localization, dynamics and/or

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

persistency is key prior knowledge to integrate in hybrid AI algorithms either for contrail verification or as data assimilation techniques.

This report focuses on evaluating the available observations and highlight relevant pre-processing steps needed for their use in the developed algorithms.

1.3 Sources of available or useful observations in data assimilation or contrail detection

Models can be complemented with actual contrail observations, available from many different sources:

- Satellite imagery: contrails are visible in channels used for distinguishing thin cirrus clouds, in particular by using brightness temperature differences between channels[1],
- Ground based cameras which can be:
  - regular planar or wide angle cameras,
  - fisheye spherical or hemispherical cameras,
- Lidar signals which can be useful for estimating more precisely their altitude and their optical depth through the signal attenuation through vertical levels of the atmosphere.

Through discussions inside the BeCoM consortium we have first chosen to build a sample collocated dataset by selecting a two-month period in May-June 2023 for which both ground-based images and Eumetsat/SEVIRI images should be readily available. The purpose of this action is to serve in calibrating the developed algorithms for genericity. For instance, a spherical neural network could be used for contrail classification/identification directly on the ground based fisheye images and on satellite images projected to the sphere. Conversely, we investigate projecting the fisheye images on the plane and using them to train a similar neural network as with patches of satellite images (in the planar domain). Further discussions need to be conducted internally in the project in order to also add Lidar signals when possible.

However, one of the limitations that we encounter with the collocated approach is the lack of unified annotation strategy. The annotation is highly dependent on the task to solve: e.g. whole image label for classification, labelled-polygon for segmentation (identification). Both are non-trivial and with increasing degree of complexity, i.e. dependent on image quality as well as on specialized domain knowledge.

Contrail overlap is smaller in ground-images due to their geographical locality, each image containing only a couple of occurrences, whereas in satellite images, depending on the satellite image patch size the density of contrails could be much higher. The co-localized identification may be challenging to perform simultaneously in this context. A hierarchical approach may be more suitable: e.g. first classify a given contrail in a ground-based image, then identify it in a spatial-temporally collocated satellite patch, provided that is not too young so as to be seen via satellite, e.g. usually more than 10-minutes old contrails.

Therefore, in order to simplify the problem in this period we considered contrail analysis in different data sources separately. We describe here the different datasets and/or data sources which are

[1] Contrails - when do we see them from satellites? | EUMETSAT

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

useful for each situation. Note that Milestone MS3.1 (see Appendix A) catalogues all the openly available sources that we have identified during this task. In this deliverable we only describe the sources we are considering for use in the developed AI algorithms.

To our knowledge, there are no annotated datasets for fisheye images. We have started building an annotated sample training dataset for binary classification. This constitutes an initial contribution of this workpackage as well as a basis for developing the robust AI algorithms. These can then be adapted to more complex ones such as object detection or semantic segmentation (e.g. by reusing the generalized convolution operators developed in Tasks T3.2, T3.3), provided that the annotation strategy is revised – or that new annotated data become available during the project[2]. Will further investigate annotation strategies for contrail identification in the following months.

2 Ground imagers

Ground-camera imaging offers several advantages in the study of contrails e.g. when compared to satellite imagery. They benefit from a higher spatial and temporal resolution, meaning that more detailed and precise images of contrails can be captured. This allows for better discrimination between contrails and other atmospheric features or objects and facilitates accurate measurements and analysis of contrail characteristics. Ground cameras can be placed strategically in areas where contrail study is of particular interest, allowing for a focus on specific geographic regions, for instance in adequate proximity of airports, where other instruments (such as Lidars) are deployed or regions of the Globe where air traffic activity is particularly dense in conjunction with contrail-favourable meteorological conditions, and where contrails can potentially be found in high numbers.

Despite the advantages of ground-camera imaging, several challenges must be addressed when utilizing this approach in contrail studies. Weather conditions and environmental factors, such as cloud cover, precipitation, as well as camera flares caused by direct sunlight, can hinder the visibility of contrails from the ground, whereas the use of cameras in the visible spectrum renders impossible the study of contrails during night-time. Moreover, the precise measurement of contrail properties such as altitude and width can be challenging from ground-based images, and needs at least two collocated sensors. Additionally, the localized nature of ground-based cameras is challenging for monitoring persisting contrails over the entire course of their lifetime in a wide geographical area, again needing several sensors placed strategically close to one another. Finally, the limited availability of labelled contrail datasets from ground observations is a challenge in the development and training of automated contrail detection algorithms using ground-based imagery.

2.1 Identified datasets and data sources

At the start of the project we focused on identifying available camera/imagers data, keeping in mind that this data should be used for training AI algorithms. This implies the need for ground truth about the content of these images and in particular we were looking for annotated contrails.

We started a survey of the potential sensors and data sources publicly available with a focus on data availability but also on sensor type, one of the goals of the project being to explore using whole-sky imagers in conjunction with more widely used modalities such as satellite. The purpose of this is to

[2] EUROCONTROL has recently announced leading a new European initiative called ContrailNET (https://www.eurocontrol.int/news/eurocontrol-launches-contrailnet-new-network-create-common-repository-contrail-observation) aiming to collect and store unified data for contrail mitigation, along with bringing together researchers and scientist from the various associated disciplines to work on a common data management methodology. Thales is a partner in this initiative.

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

enhance the probability of detecting contrails early on from their appearance and also calibrate multi-modal algorithms which will contribute to reducing prediction uncertainties through verification.

Existing datasets from the literature

CCSN: Among the existing works using ground-based images for contrail detection, one source of data is the CCSN (Cirrus Cumulus Stratus Nimbus) database. This database was constructed by (Zhang et al., 2018) [1] in order to develop the CloudNet model for cloud classification. Availability of sufficient training samples is a known fundamental issue in cloud classification research (Xiao et al., 2016; Zhuo et al., 2014)[2],[3]. A previously existing dataset for ground-based cloud classification is the SWIMCAT dataset (Dev et al., 2015) [4], but it does not include precise cloud type labelling and does not include labelled contrail data. The CCSN dataset was built to be three times larger than the SWIMCAT dataset and consists of 2543 unique ground-based cloud images. Each image has a resolution of 256×256 pixels in JPEG format [1]. The images were labelled by meteorological experts and divided by cloud type into 11 categories: cirrus (Ci), cirrostratus (Cs), cirrocumulus (Cc), altocumulus (Ac), altostratus (As), cumulus (Cu), cumulonimbus (Cb), nimbostratus (Ns), stratocumulus (Sc), stratus (St), and contrail (Ct). In this dataset the contrail class contains 200 labelled images of contrails. In Figure 2-1 we show a few examples of labelled contrail images from the CCSN dataset.

Figure 2-1 : Examples of contrail labeled data from the CCSN dataset (available at https://github.com/upuil/CCSN-Database)

In (Sharma et al., 2023) [5], the CCSN dataset is used in order to compare models for contrail detection in ground-based imagery. Since this study focuses on contrails detection, and not classification of other cloud types, the CCSN dataset was redistributed into two classes: “contrail” and “non-contrail”. The “contrail” category consists of all the contrail images from CCSN (200 images), while the “non-contrail” category consists of all the other categories in the CCSN dataset (2343 images), which corresponds to the other types of clouds.

TSI dataset: Another dataset used for contrail detection in ground images by (Siddiqui, 2020)[6] was built using images captured by a total-sky imager (TSI) belonging to the U.S. Department of Energy’s Atmospheric Radiation Management user facility. The TSI unit is comprised of a camera placed vertically above a convex mirror, allowing for the visualization of the sky from zenith to horizon. Images in JPEG format are captured by this TSI unit every 30 seconds. The dataset building process comprised a visual annotation process in order to label images and divide them into 2 categories: images with no visually perceived contrails (label of 0), and images with visually perceived contrails (label of 1). Each category equally contains 800 images for a total dataset of 1600 images taken during the month of March 2017 by the TSI. Some examples of images containing contrails from this dataset are shown in Figure 2-2.

This project has received funding from the European Union’s Horizon Europe Research and innovation program under grant agreement No 101056885

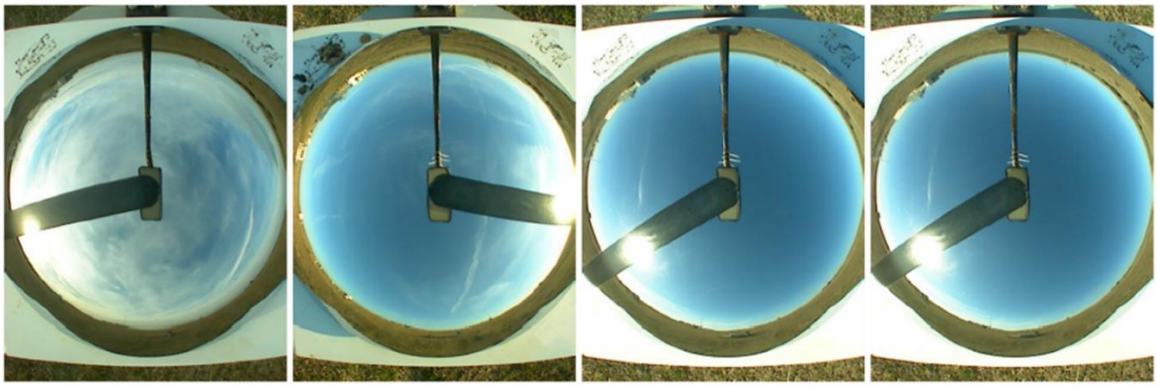


Figure 2-2 : Examples of images captured by total-sky imager with visually perceived contrails. Taken from [6]

The authors in [6] note that the contrail annotation process is a difficult task prone to human error and which also needs human expertise – especially for distinguishing between older contrails and regular cirrus in images. This is due in particular to the low visibility in the images, when other clouds besides contrails are present, contrails mixed with other cirrus-like clouds, visual distortion of the sensor etc.

However, to the best of our knowledge this dataset is not available to use. This remark is indeed very important for the BeCoM project, because the lack of annotated data is a limiting factor in developing advanced AI algorithms.

Potential data sources open for research usage

Faced with the scarcity of ground based annotated images we have searched other potential sources from which to build our own datasets for training and test of the developed AI algorithms.

The FRIPON data source: We have found an entire network of fisheye cameras in the visible spectrum, active throughout Europe and part of North America, Africa and Australia, called FRIPON [7](https://fireball.fripon.org/), dedicated to meteoroid tracking, see Figure 2-3 for the major repartition and the status of the cameras (green cameras are currently active). The raw images

Figure 2-3: Excerpt of the FRIPON camera network coverage. The biggest coverage is in Europe.

acquired from these cameras can be used for research purposes provided the appropriate source acknowledgements.

These cameras acquire images every ten minutes, and are stored in .jpeg or .fit[3] format, at 1280x960 pixels on 3 channels. The median distance between the cameras is of 80km.

They capture many phenomena, the main purpose of this network being meteoroid detection and tracking, but since they use visual sensors other objects of interest can occur. In particular, they can

[3] Flexible Image Transport System standard - https://en.wikipedia.org/wiki/FITS

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

capture contrails early on and may allow to track contrails which persist above 10 minutes by considering a time interval around a given contrail from the images right before its appearance to the last image in which it is seen on a particular location.

Moreover, when using sufficiently close cameras on a given flight trajectory, contrails can be tracked geographically as well as temporally, as shown in Figure 2-4.

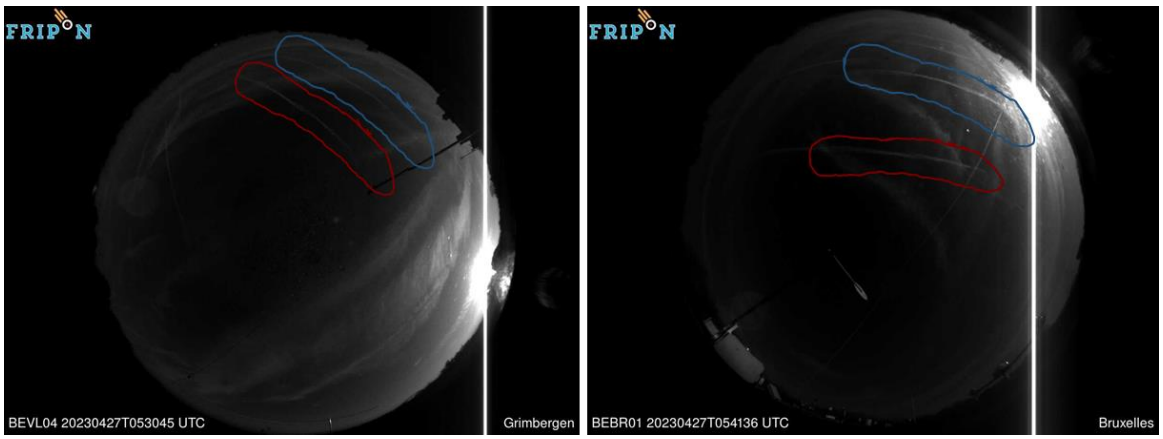


Figure 2-4 : Two contrails highlighted in red and blue which can be observed in images from two neighbouring cameras.

However, as with other whole-sky imagers which are not particularly calibrated for contrail observation, as well as due to the different acquisition conditions, the quality and visibility of contrails is highly variable, as illustrated in Figure 2-5, showing visible contrails from different cameras.

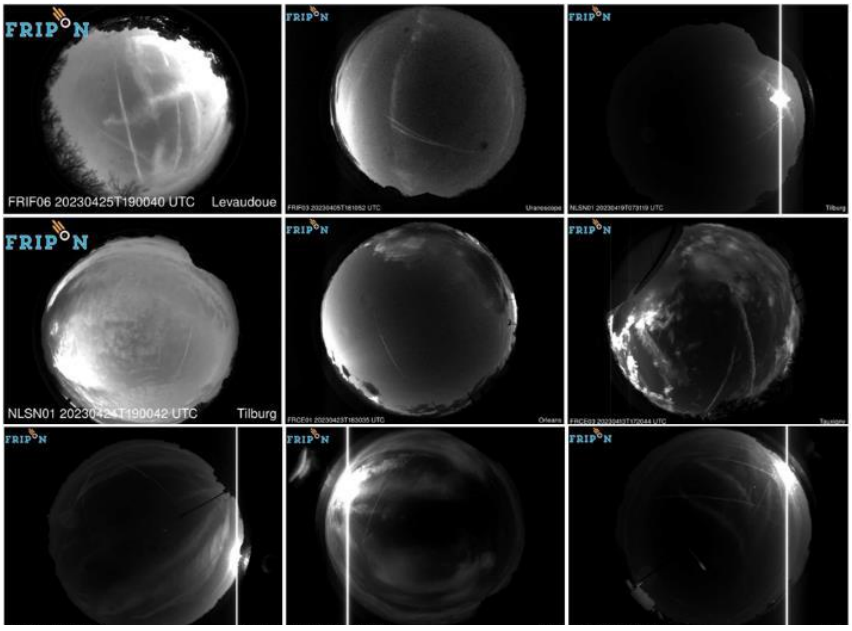


Figure 2-5 : Examples of FRIPON images from various stations

Nevertheless, the important quantity of available images in this network is promising for finding enough good-quality contrail examples to include in our relevant data. We have therefore started building a dataset for training by manually selecting images from different stations in which contrails are clearly observable. These images are in .jpeg format.

We have collected about 3000 images from 19 FRIPON stations covering a period between April and May 2023 which amount to approximately 300MB of data. To collect this data we have used

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

several flight trajectories[4] mostly in France which instrumented the time interval to collect images as well as the locations of the FRIPON cameras, and we included temporal context to maximize the probabilities of seeing actual contrails along these trajectories. The temporal context consists in considering collecting images from 10 minutes before a flight passes in the radius of a FRIPON camera up to 30 minutes after, meaning that for each point a contrail may appear in several images, this making its identification easier, along with a reference image without contrail.

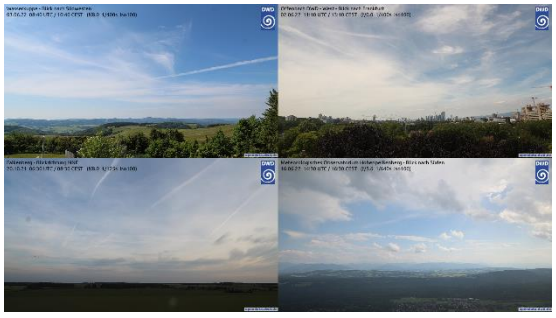
We have currently identified around 100 images which we labelled as contrails (we chose to label an image as `contrail = TRUE` if at least one contrail is present in the image), along with a couple of tens of images without contrails for building a sample dataset. A key point in machine learning algorithms is to have balanced datasets so as to avoid algorithmic biases.

This first annotation may be used with classification algorithms as those present in the literature, but also allows to start working on different annotation strategies, e.g. extend in later stages to contrail identification in an image, provided an agreed-upon methodology and enough domain knowledge to distinguish between regular clouds and older contrails is defined.

We view this as an iterative process to be continued throughout the duration of the workpackage in order to refine algorithm performance, this initial sample serving in the algorithm design stage. Algorithm design should, in addition, serve in refining the criteria for data collection since the two are intertwined: the size and the variety of the training data guides the size of the models to implement and vice versa. Moreover, one of the goals in designing our robust AI algorithms is data frugality. This initial sample should be the minimally required data for the first prototypes.

The goal is to extend this dataset to some ~200-300 observed contrails over time, possibly collocated with SEVIRI observations.

In-project data sources: Through exchanges with partners in WP1, a complementary source of ground-based fisheye images could be used in the BeCoM, alongside with the FRIPON images to increase the diversity of the training data (as illustrated by the sample image). The CNRS partner provided a dataset of about ten years of historical data in the visible range. We have yet to analyse and label (following the simple strategy above, i.e. “contrail – non-contrail” examples) this dataset. Additional calibration and preprocessing work will be needed in order to combine the two fish-eye data sources.



The DWD partner disposes of another source of ground imagery, in the planar domain. These are wide angle images (see left side) in the visible spectrum, collected over several locations in Germany, over different periods in 2021-2022. The images are available in .jpeg compression with respect to different resolutions: 5184 x 3456, 1200 x 675, 816 x 419 and have a temporal resolution of 10 minutes. The spatial coverage of these cameras is about 50x50km.

[4] Thales Flights Footprint Estimator - Flights Footprint (flights-footprint.com)

This project has received funding from the European Union’s Horizon Europe Research and innovation program under grant agreement No 101056885

This allows to build a dataset containing 390 images for training and 58 images for test, with the aim of verifying Ice SuperSaturated Regions (ISSR) from the ICON model in WP2. An annotation methodology has been defined on these images which could serve as a basis for further investigations in the other ground imagers, should we project the fisheye images on the plane.

3 Satellite data

The most explored, as well as the richest source of contrail and particularly persistent contrail data, is of course satellite data. The very wide area coverage allows for tracking various phenomena (clouds, temperatures, dust, etc.) at much larger scale than localised ground sensors.

Among the various satellite technologies, weather geostationary satellites such as:

- the US GOES series[5] covering the Americas,
- the European Meteosat Second Generation (MSG)[6] [7], covering Europe, Africa and a part of the Indian Ocean,
- or the Japan Himawari[8] covering Eastern Asia and Oceania

are a very important complement to ground images since they provide images of the full disk, with different temporal-spatial resolution trade-offs.

During an internal consortium workshop aiming to identify the relevant data source for BeCoM, in view of achieving the first milestone for this workpackage, we have established that data availability for AI algorithms should be used in conjunction with the geographic region above Europe. Indeed the numerical weather prediction models developed in WP2, in particular through assimilation of cloud observations focus on using data from Eumetsat.

Therefore, the main satellite source data for the project should come from the SEVIRI channels for the Eumetsat MSG (or MTG) satellites.

However, as with ground images, the lack of annotated data needs to be handled before implementing robust machine learning algorithms. We have thus followed the same two directions as in the previous section, in parallel:

- identify openly available datasets and analyse their potential for use, and
- start collecting SEVIRI data from EUMESAT along with initiating a labelling methodology.

Labelled datasets available for research

In [8] a first labelled contrail dataset is available on low-orbit satellite LANDSAT-8, which has sufficient spatial resolution for contrail identification. Due to its orbit type and mission, however, the availability of the relevant data (e.g. during nighttime or above water) is an issue.

Moreover, as we intend to use Eumetsat data, there is a problem of homogeneity; thus, we shall not consider this dataset further.

The OpenContrails Dataset

As with ground images a key feature for using satellite data in AI algorithms is the availability of a sufficiently large corpus of annotated data. The most complete to date one is the OpenContrails

[5] GOES Imagery Viewer - NOAA / NESDIS / STAR
[6] https://www.eumetsat.int/meteosat-second-generation
[7] The newly launched MTG – third generation European satellite system at the end of 2022 is expected to greatly increase the precision of the collected observations with higher temporal and spatial resolution and the addition of a sounding satellite.
[8] Meteorological Satellite Center (MSC) | Himawari-8/9 Imager (AHI) (jma.go.jp)

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

dataset provided by Google [9], [10] of GOES-16 Advanced Baseline Imager (ABI). This is a 450GB human-labelled dataset split into training and test. Data covers a period of one year starting in April 2019 through 2020 and contains 244400 images each with the 16 spectral bands of the GOES-16 satellite, labelled with contrail masks. This dataset has been released for the ”Google Research - Identify Contrails to Reduce Global Warming” kaggle competition[9], where it is stated that the original full-disk images were re-projected using bilinear resampling to generate a local scene image.

A sample of the images in this dataset is given in Figure 3-1 with the original image in false colour on the left and its annotated version on the right. A key point of this dataset is the availability of the used annotation methodology [12], which the consortium has started to adopt for annotating SEVIRI images in a similar manner.

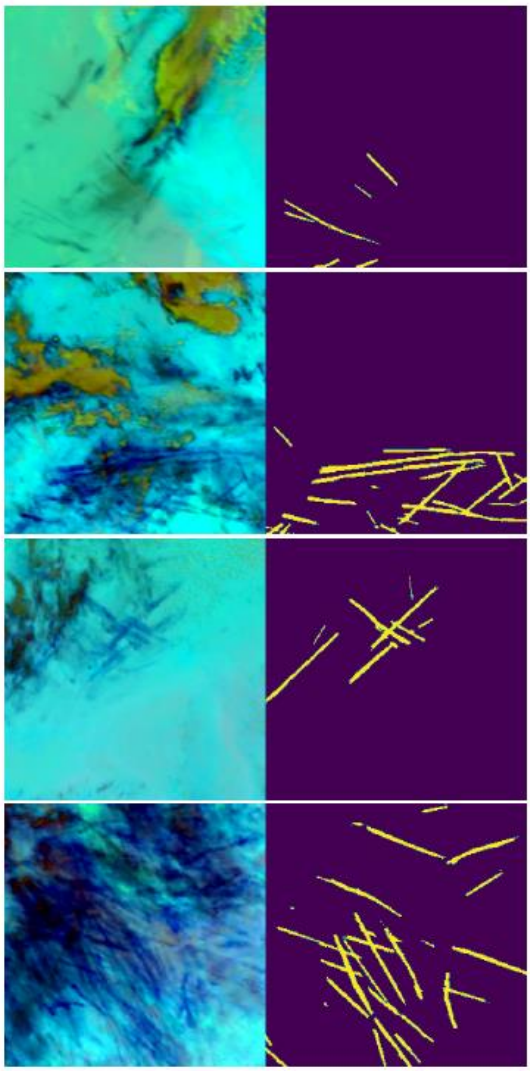


Figure 3-1 : Sample OpenContrails images (taken from [9])

As we did for the FRIPON images, temporal context was provided for the GOES16-ABI images in order to identify contrails more easily. Thus, for each contrail, a sequence of images at 10-minute intervals is provided.

[9] https://www.kaggle.com/competitions/google-research-identify-contrails-reduce-global-warming/data

This project has received funding from the European Union’s Horizon Europe Research and innovation program under grant agreement No 101056885

In addition, several neural networks architectures have been trained for contrail detection based on this dataset, which constitutes a highly valuable baseline for the robust algorithms under development in WP3.

In-project satellite data

The DWD partner has constituted a dataset of 339 images for train and 24 images for test using the Eumetsat SEVIRI satellite images, using the 11th and the 12th SEVIRI spectral channels in HRIT compression. In Figure 3-2 we present an example of the considered images in this training/test dataset. The temporal resolution is of 15 minutes and the spatial resolution is of 3 km at sub-satellite point for the 11th spectral channel and of 1 km at sub-satellite point for the High Resolution Visible (HRV) channel. Contrails are visible when using Brightness Temperature in the following InfraRed (IR) SEVIRI channels: IR 8.7, IR 10.8, IR 12.8. DWD aims at continuing the annotation of the SEVIRI images using guidelines from the existing datasets in the literature.

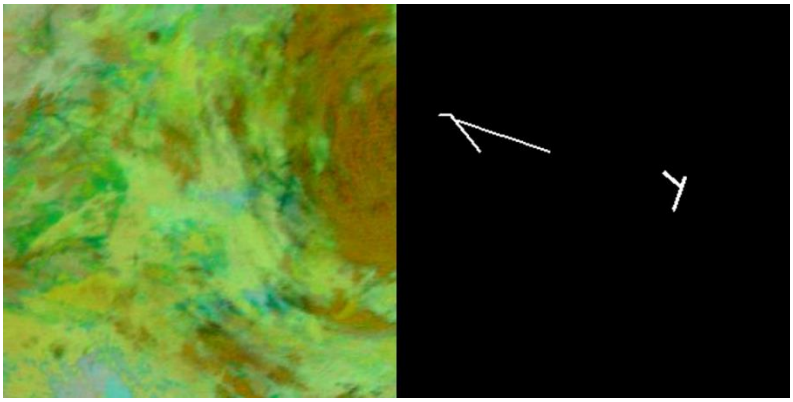


Figure 3-2 : Sample SEVIRI brightness temperature image (left) with corresponding annotated contrail mask (right)

The DLR partner provided another sample of SEVIRI images taken for the 18th of June 2022 (around 125MB of data). The sample, from which an example is given in Figure 3-3 contains:

- images of the brightness temperature difference (between 8.7 and 12µm & 10.8 and 12µm),
- images of detected contrails with the contrail detection algorithm (CDA)[11] which can be used as a preliminary annotation, and
- false-color RGB images from 0 to 12 UTC, respectively.

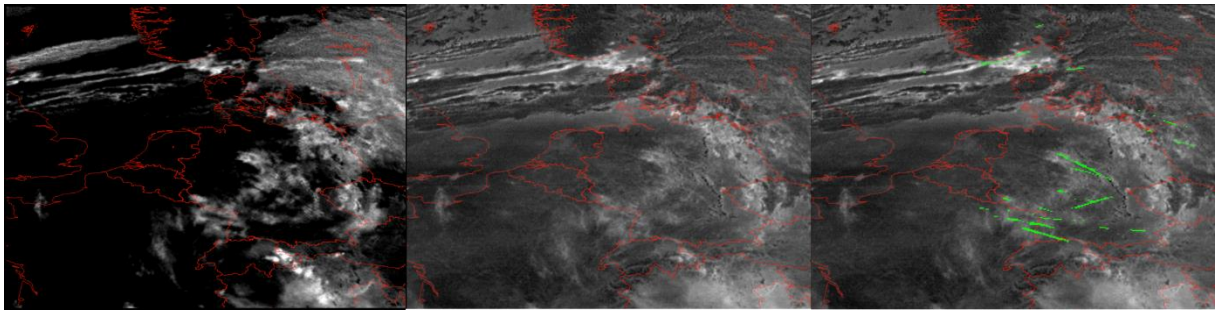


Figure 3-3: Sample images from June 18th SEVIRI Brightness Temperature Difference. Left: channels 8.7 and 12µm, Middle: channels 10.8 and 12µm and Right: detected contrails with the CDA algorithm in green

We are planning to use the SEVIRI data in conjunction with the GOES-16 data for training robust neural networks.

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

4 Conclusion

In summary, this deliverable presented the various data sources which will be used for the development of our AI algorithms in workpackages WP3 for contrail analysis and WP2 for enhancing numerical weather prediction models through data assimilation.

After analysing the available data in the literature, we have identified, collected sample datasets and initiated relevant pre-processing stages for two main complementary sources of observations:

1. Ground-based images which offer good detection possibility for very young low-altitude contrails in local key points (e.g. on particular flight trajectories), such as images coming from the French network of cameras called FRIPON and,
2. Satellite geostationary data from GOES-16 and SEVIRI which cover a wide geographical area albeit with lower spatial and temporal resolution. Lidar data may be included in later iterations in the project.

For the ground images, we have built a first **FRIPON sample annotated dataset** of fish-eye images covering several points in France relevant to usual flight trajectories and with images related to several flights time-schedules in April and May 2023, which we plan to complement with images from 2019 and 2020 to collocate with some of the annotated GOES-16 data. In-house images provided by the CNRS/LATMOS partner and covering 10 years of historical data could help in extending this sample dataset in the following months. This could be useful when using bigger models in the AI algorithms.

For the satellite modality the existence of the human-labelled, curated, **OpenContrails dataset** is a valuable addition to use in complement with SEVIRI data, reducing the need for extended labelling. This should be especially useful in combination with transfer/few-shot learning algorithms that we aim to investigate in T3.2/T3.3.

This initial collection of datasets, in terms of data availability and temporal range is sufficient for prototyping our algorithms both on ground images and on satellite.

Since other data collection initiatives have emerged at European level since the beginning of the project, such as the ContrailNet initiative from Eurocontrol we aim to keep this task alive throughout the following year so as to benefit from future iterations which could enhance our algorithmic capabilities.

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

References

[1] Zhang, J. L., Liu, P., Zhang, F., and Song, Q. Q.: CloudNet: Ground-based Cloud Classification with Deep Convolutional Neural Network. Geophysical Research Letters, 45. https://doi.org/10.1029/2018GL077787, 2018.

[2] Xiao, Y., Cao, Z., Zhuo, W., Ye, L., & Zhu, L.: mCLOUD: A multiview visual feature extraction mechanism for ground-based cloud image categorization. Journal of Atmospheric and Oceanic Technology, 33(4), 789–801. https://doi.org/10.1175/JTECH-D-15-0015.1, 2016.

[3] Zhuo, W., Cao, Z., & Xiao, Y.: Cloud classification of ground-based images using texture-structure features. Journal of Atmospheric and Oceanic Technology, 31(1), 79–92, 2014.

[4] Dev, S., Lee, Y. H., & Winkler, S.: Categorization of cloud image patches using an improved texton-based approach. In IEEE International Conference on Image Processing (pp. 422–426) Quebec City, QC, Canada, 2015.

[5] Sharma, K., Jain, S., Wu, E., Fattah, Z. M., Sarin, C., Maeshiro, D., Kumar, S., & Rajaram, A.: Utilizing Computer Vision Algorithms to Detect Contrails. In *AIAA AVIATION 2023 Forum* (p. 3769), 2023.

[6] Siddiqui, N.: Atmospheric Contrail Detection with a Deep Learning Algorithm. Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal, 7(1), 5, 2020.

[7] F. Colas et al - FRIPON: a worldwide network to track incoming meteoroids, A&A 644, A53 (2020)

[8] McCloskey, K. J. F., Geraedts, S. D., Jackman, B. H., Meijer, V. R., Brand, E. W., Fork, D. K., ... & Van Arsdale, C. H. (2021). A human-labeled Landsat contrails dataset.

[9] Ng, J. Y. H., McCloskey, K., Cui, J., Brand, E., Sarna, A., Goyal, N., ... & Geraedts, S. (2023). OpenContrails: Benchmarking Contrail Detection on GOES-16 ABI. *arXiv preprint arXiv:2304.02122*.

[10] GitHub - patmejia/contrails_vision: Climate Analytics | contrail detection in GOES-16 imagery, aiding climate change mitigation. | Preprint: https://arxiv.org/abs/2304.02122

[11] Mannstein, H., Meyer, R., & Wendling, P. (1999). Operational detection of contrails from NOAA-AVHRR-data. *International Journal of Remote Sensing*, *20*(8), 1641-1660.

[12] OpenContrails Labeling Instruction (submission) (storage.googleapis.com)

[13] Schumann, U., 2012: A contrail cirrus prediction model. Geosci. Model Dev., 5, 543–580, doi: 10.5194/gmd-5-543-2012.

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

5 Appendix A – MS 3.1 Relevant sources of data for training have been identified

During the first year of the BeCoM project we have achieved the first milestone, MS3.1, by establishing a catalogue of potential data sources out of which we were able to select the most relevant ones. This was reported internally during the Year One general assembly. We include the report in this deliverable for completeness purposes.

To build this catalogue we have also defined the necessary metadata where for each datasource we listed:

- Name
- Source-type/author
- Last updated
- Download link
- Data format
- Licence type
- Acquisition type
- Size (or sample size: for continuously updated data we have initially only tested the download capability from the data source).
- Files characteristics

In Table 5-1 we give an overview of the initially identified data sources for this milestone, along with their most important metadata. This catalogue allowed the selection of only a pertinent achievable subset for the project. The final selected data sources in this deliverable reflect the internal consortium discussions and are considered to be the most relevant for the following work.

Table 5-1 : Milestone MS3.1 Relevant source of data for training

Source Nb	Name	Download Link	Data Format	Acquisition type	Size/Sample size	License
1	95-Cloud: Cloud Segmentation On Satellite Images	Kaggle Link	.txt; .csv; .tif; .xlsx	Satellite: Landsat 8	35.6 GB	CC0: Public Domain
2	A human-labeled Landsat-8 Contrails dataset	Google Cloud Link	.json	Satellite: Landsat 8	5.2 GB	CC BY 4.0
3	Simulations of Contrails Under COVID-19 Effects	Zenodo Link	.nc: .F90	Flight track data	32.7 GB	CC BY 4.0 International
4	Aviation contrail cirrus	Zenodo Link	.nc	Satellite : SEVIRI MSG	115.4 MB	CC BY 4.0

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

	and Radiative forcing over Europe for six months in 2020 during COVID-19 compared With 2019: Observations and model results					International
5	SUCCESS Utah Polarization Diversity LIDAR data set	Earthdata Link	.ps; .dat	Satellite: CALIPSO – LIDAR	10.7 MB	NASA Web Privacy Policy
6	EUMETSAT records dataset	EUMETView Link	.mp4; .xml	Geostationary satellites: Meteosat 0 Degree; Meteosat IODC; GOES-16; GOES-17; Himawari-8	1.1 GB (only a sample for download testing)	Attribution-ShareAlike 3.0 IGO
7	FRIPON (Fireball Recovery and InterPlanetary Observation Network) fish eye Contrails Dataset	Link	.jpg; .csv	Camera : Fisheye	Variable	Free of use for education, Non commercial purpose only
8	CALIPSO LIDAR Dataset	Earthdata Link	.hdf	Satellite : CALIPSO Lidar	38.5 GB	NASA Web Privacy Policy
9	OpenContrails:	Kaggle link	npy, json, csv	Satellite: GOES-16	450.91 GB	CC BY 4.0
10	Cirrus Cumulus Stratus Nimbus (CCSN) Database	CCSN_database_link	.jpg	Camera: planar	104MB	CC0: Public Domain

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