Physically Consistent Contrail-to-Flight Attribution

Andrei - Ionut Paraschiv

Delft University of Technology

TU Delft

This page is intentionally left blank.

Physically Consistent Contrail-to-Flight Attribution

by

Andrei - Ionut Paraschiv

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday February 2, 2026 at 12:45.
Faculty of Aerospace Engineering TU Delft, Hall E (formerly K)

Student number:	5955556	
Project duration:	December, 2024 – January, 2026	
Thesis committee:	Dr. F. Yin,	TU Delft, Chair
	Dr. V.R. Meijer,	TU Delft, Supervisor
	Dr. C. Varriale,	TU Delft, Independent examiner

Cover:	Photo of Airplane Across the Clouds during Night Time by Seven-Storm JUHASZIMRUS (Modified)
Style:	TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

Preface

Over the past months, I had the opportunity to dive deep into aviation's climate impact, with a particular focus on contrail-to-flight matching. This period has been both intense and rewarding, and completing this thesis marks the final step of my time at TU Delft. Looking back, my time here has helped me grow both professionally and personally, shaped by challenges and important moments of reflection.

I want to express my sincere gratitude to Vincent Meijer and Feija Yin for their guidance and valuable feedback throughout the thesis. Their input consistently helped me improve the quality of my work and approach problems with greater thoroughness. A special thank you goes to Vincent Meijer, with whom I spent many hours discussing ideas and results. His patience, support, and clear explanations were crucial throughout the entire process.

Finally, I would like to thank my family, friends, and everyone who supported me during this journey. Their encouragement, understanding, and trust meant a great deal to me.

Andrei - Ionut Paraschiv
Delft, January 2026

Summary

Aviation's climate impact is driven not only by CO_2 emissions but also by non-CO_2 effects, of which contrail cirrus is the dominant contributor. In particular, non-CO_2 effects were estimated to account for roughly two-thirds of aviation's net effective radiative forcing in 2018. Because persistent contrails form only under specific atmospheric conditions and produce a net warming effect, operational mitigation concepts such as contrail region avoidance require reliable identification of which flights produced which observed contrails.

This thesis investigates contrail-to-flight attribution for contrails detected in satellite imagery, where they are typically displaced from the aircraft trajectory due to advection, and where uncertainty in the wind field can make it difficult to identify the true source flight. Most recent approaches rely on geometric matching. A key limitation of geometric-only matching is that it neglects wind uncertainty: because advection integrates wind errors over time, small forecast biases can grow into large spatial offsets, increasing ambiguity and leading to false positive matches. To address this, the core idea in this work is to incorporate physical consistency across groups of nearby contrails, motivated by the expectation that wind errors are locally correlated in space and time.

To enable controlled evaluation with known ground truth, the thesis constructs synthetic datasets by advecting real ADS-B flight trajectories using ERA5 ensemble meteorology. Contrails are generated with a single calibrated ensemble member, while candidates are generated with the ensemble mean of the remaining members, introducing realistic but controlled displacement consistent with meteorological uncertainty. Flights are filtered prior to advection to retain only cases that are physically plausible for persistent contrail formation, using ice-supersaturated regions (ISSRs) and the Schmidt-Appleman criterion. Two dataset variants are produced: one using whole trajectories and one using segmented trajectories, enabling a fair comparison with existing baselines.

The proposed attribution method is formulated as a mixed-integer linear programming optimization that selects contrail–flight matches while encouraging global physical consistency across clusters of contrails. Its objective combines three components: an advection-error penalty, an altitude consistency rule, and a geometric match score derived from the baseline approach. Performance is evaluated using per-contrail and per-flight precision and recall, following established metrics used in recent contrail attribution literature.

Across the evaluated synthetic datasets, the physically consistent algorithm applied to whole trajectories delivered the strongest overall results. Compared to the baseline whole-trajectory approach, it achieved clearly higher precision and recall at both the contrail level and the flight level, indicating more reliable attribution with fewer false matches. Segmenting trajectories helped the baseline recover more true matches, but this came with a noticeable drop in precision, highlighting the recall–precision trade-off inherent to purely geometric matching. In terms of computational cost, the physically consistent whole-trajectory approach remained practical, with runtime scaling favourably as dataset size increased.

Finally, an ablation study confirmed that all three objective components contribute positively to attribution accuracy, and a statistical analysis indicated that the observed performance gaps between methods are unlikely to be explained by sampling variability alone. At the same time, the evaluation is based on a relatively small synthetic dataset (304 contrails), and further validation on real contrails detected in satellite imagery is required. Future work should include applying the framework to real GOES-16 images, improving clustering stability, and addressing additional sources of uncertainty.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
ADS-B	Automatic Dependent Surveillance-Broadcast
ARCO	Analysis-Ready, Cloud Optimized
CNN	Convolutional Neural Network
CO	Contrail Object
CoAtSaC	Contrail Attribution Sample Consensus
CoCiP	Contrail Cirrus Prediction
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF Reanalysis v5
ERF	Effective Radiative Forcing
FAA	Federal Aviation Administration
FIR	Flight Information Region
FL	Flight Level
GOES	Geostationary Operational Environmental Satellite
ISSR	Ice Supersaturated Region
MILP	Mixed-Integer Linear Programming
PDF	Probability Density Function
RF	Radiative Forcing
SAC	Schmidt-Appleman Criterion
SAF	Sustainable Aviation Fuel
TFMS	Traffic Flow Management System
UTC	Coordinated Universal Time

Symbols

Symbol	Definition	Unit
$a\,(i)$	Area index associated with contrail i based on the clustering rule	[-]
$\mathcal{C}$	Set of contrails used in the physically consistent algorithm	[-]
C_{fit}	Constant used in the cost function described in Geraedts et al. (2024)	[km^{-2}]
C_{shift}	Constant used in the cost function described in Geraedts et al. (2024)	[km^{-2}]
C_{angle}	Constant used in the cost function described in Geraedts et al. (2024)	[-]
C_{age}	Constant used in the cost function described in Geraedts et al. (2024)	[hour]
c_p	Specific heat capacity of air at constant pressure	[$\text{J kg}^{-1}\,\text{K}^{-1}$]
$D_{\max}$	Maximum distance between two contrails in a dataset (Haversine)	[km]
$\text{EI}_{\text{H}_2\text{O}}$	Emission index of water vapour	[kg H_2O (kg fuel)$^{-1}$]
$\mathbf{e}_{ij}$	Implied advection error vector between contrail i and candidate flight j	[degree] or [degree s^{-1}]or [m] or [m s^{-1}]
F	False-alarm rate	[-]
$\mathcal{F}$	Set of candidate flights used in the physically consistent algorithm	[-]
FN	False negative count	[-]
FP	False positive count	[-]
G	Mixing line slope in the Schmidt-Appleman criterion	[Pa K^{-1}]
G_{ij}	The match score between contrail i and candidate j computed by the algorithm in Geraedts et al. (2024)	[-]
H	Hit rate	[-]
h	Scaling factor chosen as the average height of an ISSR	[m]
K	Number of contrail clusters used in the physically consistent algorithm	[-]
N	Number of contrails	[-]
$\text{Precision}_{\text{contrail}}$	Per-contrail precision	[%]
$\text{Precision}_{\text{flight}}$	Per-flight precision	[%]
p	Ambient Pressure	[Pa]
$\text{Recall}_{\text{contrail}}$	Per-contrail recall	[%]
$\text{Recall}_{\text{flight}}$	Per-flight recall	[%]
Q	Heat released per unit mass of fuel burned	[J (kg fuel)$^{-1}$]
S_k	Set of contrail-candidate pairs currently assigned to cluster k	[-]
S	Base rate	[-]
s	Chosen maximum distance from one side of the contrail cluster to the other	[km]
TN	True negative count	[-]
TP	True positive count	[-]
V	Shift parameter in the v-direction	[km]

Symbol	Definition	Unit
v	Distance along axis in a rotated local coordinate system aligned with contrail direction (contrail axis)	[km]
vertRate	Vertical rate of descent of an aircraft	[ft min^{-1}]
W	Shift parameter in the w-direction	[km]
w	Distance along axis in a rotated local coordinate system perpendicular to contrail direction (cross-contrail axis)	[km]
x_{ij}	Binary decision variable for MILP	[-]
z_j	Altitude of candidate flight j	[m]
Δz	Altitude difference between two consecutive waypoints on a flight path	[m]
ε	Ratio of the molar mass of water vapour to dry air	[-]
λ_1, λ_2, λ_3	Weights of the terms in the objective function	[-]
μ_k	Average implied advection error for cluster k	[degree] or [degree s^{-1}] or [m] or [m s^{-1}]
η	Aircraft engine efficiency	[-]
θ	Rotation parameter for the baseline algorithm	[rad]
$\zeta_{a(i)}$	Mean of the altitudes of all candidates that were previously assigned to contrails in the cluster that contains contrail i	[m]

List of Figures

List of Tables

Chapter 1
Introduction

In this chapter, I review the background and recent literature relevant to contrail-to-flight attribution and outline the objectives of this thesis. Section 1.1 introduces the physical processes behind contrail formation, their climate impact, and current mitigation strategies. Section 1.2 reviews existing flight attribution methodologies, including data inputs, detection methods, algorithmic structures, outputs, and evaluation strategies, as well as their key limitations. Finally, Section 1.3 formulates the research objectives, which aim to address these limitations by developing a physically consistent attribution algorithm and a new evaluation dataset.

1.1. Contrail Formation and Climate Impact

In this section, I introduce the scientific background of contrail formation, its climatic implications, and how these motivate mitigation strategies and the need for better observations. Subsection 1.1.1 discusses the physical mechanisms of contrail formation and outlines their classification into non-persistent and persistent contrails depending on environmental conditions. Subsection 1.1.2 examines the contribution of contrails to aviation's overall climate impact. Subsection 1.1.3 then reviews technological and operational options for mitigating contrail climate effects and highlights the main limiting factors for operational contrail avoidance. Finally, Subsection 1.1.4 and Subsection 1.1.5 introduce satellite-based remote sensing of contrails and explain why contrail-to-flight attribution is needed, why it is challenging, and how this motivates the objectives of this thesis.

1.1.1. Formation and Classification

Condensation trails (Brewer, 1946), commonly referred to as contrails (Appleman, 1953), are line-shaped clouds that consist of condensed and frozen water vapour. They form in the wake of aircraft engines when atmospheric and exhaust conditions are favourable, typically at cruise altitudes in the upper troposphere. According to U. Schumann (1996), contrail formation is primarily driven by the emission of water vapour, but it is also influenced by additional factors such as ambient temperature and humidity, engine heat release, and particle formation processes in the exhaust plume.

Contrail formation generally requires ambient temperatures below approximately -40ºC. However, the exact threshold is determined by the Schmidt–Appleman criterion (SAC), which defines this critical temperature based on the slope G (U. Schumann, 1996), which is given by

$$G = \frac{\mathrm{EI}_{\mathrm{H_2O}}\, c_p\, p}{\varepsilon Q\ (1-\eta)} \tag{1.1}$$

where $\mathrm{EI}_{\mathrm{H_2O}}$ is the emission index of water vapour, c_p is the specific heat capacity of air at constant pressure, ε is the ratio of the molecular masses of water vapour to dry air, Q is the heat released per unit mass of fuel burned, and η is the aircraft engine efficiency. This criterion accounts for the combined effects of ambient pressure and thermodynamic properties of the exhaust plume.

Contrails are generally classified according to their lifetimes. Non-persistent contrails dissipate within minutes, whereas persistent contrails can last for tens of minutes to several hours, sometimes evolving into contrail cirrus. Contrails persist exclusively in ice-supersaturated regions (ISSRs), where the relative humidity with respect to ice exceeds 100 % (Klaus Gierens et al., 2020). In the upper troposphere,

ISSRs typically appear as localized, intermittent regions whose horizontal and vertical extents vary over relatively short spatial (tens to hundreds of kilometres) and temporal scales (Spichtinger et al., 2016). Over time, persistent contrails typically widen due to wind shear, potentially transforming into extensive contrail cirrus clouds (Kärcher et al., 2018).

1.1.2. Climate Impact

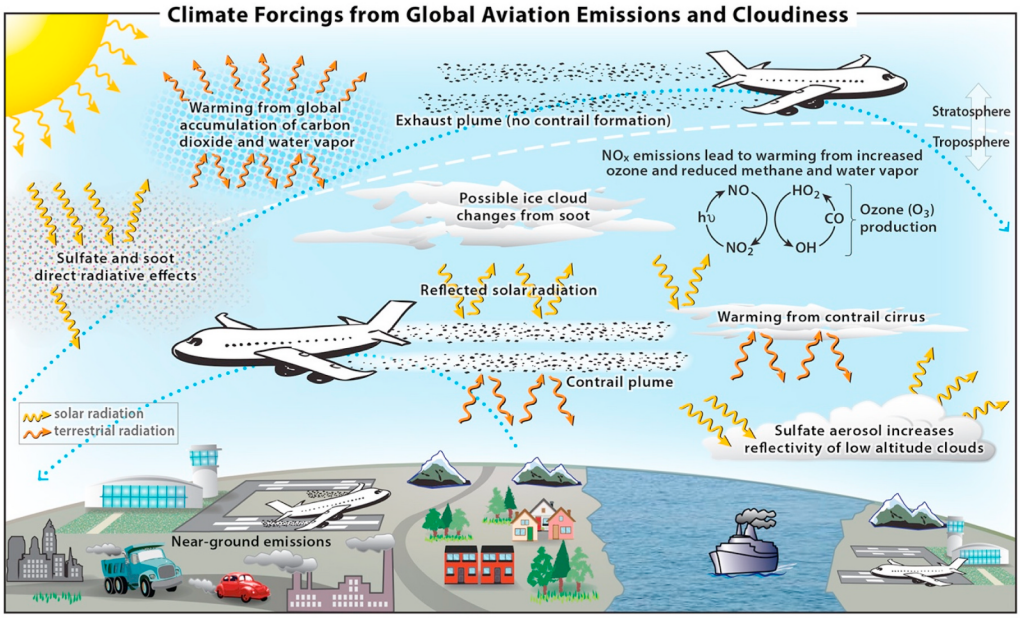


Figure 1.1: Overview of aviation climate impact (Lee et al., 2021)

Global aviation climate impact is usually split into CO_2 and non-CO_2 effects. Non-CO_2 effects include the impact of contrail cirrus, which is the largest contributor to aviation-induced climate forcing. Lee et al. (2021) state that, in 2018, non-CO_2 effects were responsible for approximately two-thirds of the total net effective radiative forcing (ERF) from aviation.

Effective radiative forcing (ERF) is a widely used metric across multiple sectors, including aviation, to quantify the climate impact of different forcing agents. It serves as a proxy for the change in global mean surface temperature attributable to a specific forcing mechanism (Intergovernmental Panel on Climate Change (IPCC), 2014). Contrails alter the radiative energy balance of the Earth through two opposing processes: they can contribute to cooling by reflecting incoming shortwave solar radiation to space, and to warming by reducing the outgoing longwave radiation emitted by Earth. Overall, contrails are estimated to produce a net positive ERF, corresponding to a warming effect (Lee et al., 2021). These processes are illustrated in Figure 1.1.

1.1.3. Contrail Mitigation

Because non-CO_2 effects contribute significantly to aviation's climate impact, several mitigation strategies have been proposed. These can be grouped into two main categories:

- Technological improvements
- Operational measures

One way to address contrails through technological improvements is to look at the formation of persistent contrails. The soot particles present in an aircraft's exhaust plume are an important driving factor behind ice particle creation (Ulrich Schumann, 2005). Therefore, one avenue in contrail mitigation would be to decrease or eliminate the soot concentration of aircraft exhaust. For example, Märkl et al. (2024) show that the use of sustainable aviation fuels (SAFs) can decrease soot emissions and, ultimately, ice particle formation, resulting in a lower ERF caused by contrails. However, these measures

require changes in fuel supply and/or engine technology, so they are typically developed over long time scales.

Operational measures seek to reduce the climate impact of contrails without modifying the existing aircraft. A central idea is contrail avoidance: modifying flight trajectories to reduce the amount of time spent in ISSRs where persistent contrails are likely to form. If regions with a high probability of persistent contrail formation can be identified in advance, flights can be rerouted to avoid these regions. This concept is illustrated in Figure 1.2, where the top diagram presents a plan view of the avoidance region and the bottom diagram shows a vertical cross-section. As observed by Spichtinger et al. (2016), ISSRs often cover large horizontal areas while remaining relatively small in the vertical dimension, suggesting that altitude changes might suffice.

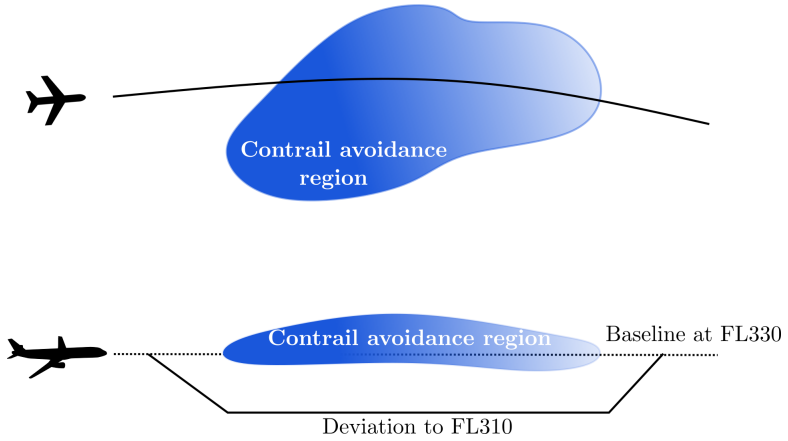


Figure 1.2: Contrail avoidance concept (Meijer, 2024)

The effectiveness of operational contrail avoidance depends on how well these regions can be predicted. A major challenge in assessing the climate impact of contrail cirrus lies in quantifying it accurately. Lee et al. (2021) note that non-CO_2 effects contribute substantially to the overall uncertainty in aviation's net ERF. Meijer (2024) groups the sources of uncertainty into two main categories:

- Uncertainty in contrail formation regions and lifetime of contrails
- Uncertainty in radiative forcing measurements

For contrail avoidance, the first category is particularly important. Avoidance methods require reliable forecasts of where ISSRs will occur and how long contrails will persist. In practice, current prediction systems rely on weather data, which can show substantial errors in relative humidity with respect to ice, the most important variable when it comes to ISSR predictions (Hofer et al., 2024). Hofer et al. (2024) suggest that weather models could be improved by equipping more aircraft with hygrometers and using the measured humidity to improve the predictions of flight humidity.

1.1.4. Remote Sensing of Contrails

Remote sensing means collecting information about an object without directly touching it. It plays a central role in contrail research because contrails are observable as linear cloud structures in satellite imagery. This is made possible by the fact that contrails interact with both incoming solar radiation and outgoing terrestrial radiation, as discussed in Subsection 1.1.2 as well.

Modern geostationary satellites provide continuous, high-frequency observations over large geographical regions, making them valuable for detecting the formation, evolution, and dissipation of contrails. These sensors allow researchers to track contrails even in situations where ground-based camera observations are unavailable due to cloud coverage (Ng et al., 2023). Remote sensing plays an important role in contrail research, as it provides large-scale observational evidence of when and where contrails form and how they evolve. As a result, satellite imagery can be used to build consistent datasets of contrail occurrence and evolution.

However, to determine the time and location of ISSRs, it is not enough to use remote sensing because the contrails observed on satellite imagery are already blown away from their initial position. This phenomenon (referred to as appearance lag) represents the delay between contrail formation and its detection in satellite imagery. Therefore, without any additional data, remote sensing would not be able to determine the exact location and time of contrail formation to be then used for contrail avoidance. Consequently, remote sensing is an excellent tool for gathering data on contrails, but by itself it cannot provide the formation information required for contrail avoidance.

1.1.5. Contrail-to-Flight Attribution

As mentioned before, to determine contrail avoidance regions, one should find a way to discover where contrails are initially formed. Remote sensing would offer insufficient data to do this constantly. However, given a contrail detected on satellite imagery, it can be attributed to the flight that initially created it by comparing their trajectories. This process is called contrail-to-flight attribution.

As mentioned before, winds transport contrails away from their point of formation, meaning that, together with appearance lag, a contrail may first appear displaced from the original flight trajectory of the aircraft that produced it. This displacement can be corrected by advecting the flight path using wind data from reanalysis datasets such as European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5). Afterwards, the advected trajectory can be compared to the contrail to determine if the contrail can be attributed to the flight. By these attributions, contrail formation regions could be successfully determined, and operational contrail avoidance could be used reliably.

The GOES-16 is widely used in flight attribution studies (Chevallier et al., 2023; Geraedts et al., 2024; Sarna et al., 2025) due to its high temporal resolution (5 minutes), i.e., each satellite image is captured 5 minutes after the previous one (this thesis will refer to each image's time as *frame time*). However, its comparatively coarse spatial resolution (approximately 2 km) limits the detection of newly formed, optically thin contrails. Such contrails become visible only after reaching the minimum optical thickness and/or width required for satellite detection (Meijer, 2024). On the other hand, as the contrail grows older and thicker, it might look like it has disappeared in the background from the satellite's perspective. Therefore, it can be concluded that satellite imagery can effectively detect contrails only over a limited period of their lifetime (this thesis will refer to this period as the *observation window*).

One major challenge in flight attribution is the uncertainty in weather data, particularly wind measurements. Suo et al. (2024) evaluated ERA5 wind speed accuracy only up to approximately 5 km altitude, reporting errors of up to $10\,\mathrm{m\,s^{-1}}$ near the surface and at their uppermost evaluated level, with better performance in the middle troposphere. The accuracy at upper-tropospheric levels relevant to contrail formation may differ due to variations in measurement coverage and atmospheric dynamics at these altitudes. Wind errors accumulate over longer advection periods, reducing the accuracy of the flight attribution algorithm.

Altitude estimation of contrails presents another challenge for flight attribution. These estimates are typically uncertain, which can result in misattributions when two or more flights overlap in satellite imagery. A further limitation is the scarcity of validation data. Until recently, there was no comprehensive reference dataset or benchmark to evaluate the performance of attribution algorithms. This gap has been addressed by Sarna et al. (2025), who introduced a benchmark dataset specifically designed for contrail-to-flight matching.

These challenges motivate the development of attribution algorithms that are more robust to meteorological uncertainties and that enforce physical consistency across multiple contrail–flight pairings in the same region. The rest of this thesis looks at existing methods in more detail and builds on this idea by designing and evaluating a physically consistent contrail-to-flight attribution algorithm and by constructing a synthetic evaluation dataset created for this problem.

1.2. Literature Review

In this section, I review the state of the art in contrail-to-flight attribution methodologies. Subsection 1.2.1 outlines the types of input data employed across studies, including flight trajectory datasets, meteorological reanalyses, and satellite-based contrail detection models. Subsection 1.2.2 compares algorithmic structures, distinguishing between single-frame and multi-frame approaches and discussing

key techniques used in these methods. Subsection 1.2.3 summarizes the outputs generated by different attribution methods, and reviews their evaluation strategies. Finally, Subsection 1.2.4 identifies key limitations in the existing literature that motivate the objectives of this thesis.

In the past five years, several studies have addressed the problem of flight matching, each proposing its own attribution algorithm. While all share the same underlying objective, their approaches differ significantly. Table 1.1 summarizes five representative studies, highlighting their key characteristics:

- Algorithm inputs in terms of weather and flight data (GOES-16 imagery is omitted, as it is used in all cases)
- Contrail detection method
- Algorithm structure
- Use of multi-frame or single-frame contrail tracking
- Algorithm output
- Evaluation methods

Table 1.1: Summary of flight-matching methodologies used in recent studies

Study	Chevallier et al. (2023)	Geraedts et al. (2024)	Gryspeerdt et al. (2024)	Barbosa (2024)	Sarna et al. (2025)
Weather Data	ERA5 reanalysis	ERA5 reanalysis	ERA5 reanalysis	ERA5 reanalysis	ARCO-ERA5
Flight Data	ADS-B (OpenSky)	ADS-B (FlightAware)	FAA TFMS (New York FIR)	ADS-B (FlightAware)	ADS-B (FlightAware)
Contrail Detection	In-house contrail detection algorithm	CNN-based contrail detection (Ng et al., 2023)	Modified Res-UNet CNN-based contrail detection (Zhang et al., 2018)	CNN-based contrail detection (Meijer, Kulik, et al., 2022)	CNN-based contrail detection (Ng et al., 2023)
Algorithm Structure	MILP optimization	Cost function minimization	Post-processing of contrail detections using linearity constraints	Probabilistic flight attribution model using Gaussian-based probability distributions	Cost function minimization
Multi-frame Tracking	Yes	No	Yes	Yes	Yes
Output	List of matched contrail-flight pairs, ranked by confidence score	Binary classification of flight segments (”matching” or ”not matching” a contrail)	Filtered contrail-flight pairs	Probabilistic outputs, ranking possible matches by likelihood	Score-based match between multiple contrails and flights, across frames
Evaluation	Manual validation on several case studies	Human-labelled validation dataset (1000 flight segments manually labelled)	No synthetic dataset or manual ground-truth dataset	Ground-truth dataset of 180 unique contrails and 1980 labels attributed manually	SynthOpenContrails benchmark dataset (in-house)

1.2.1. Data and Contrail Detection

Firstly, all five studies rely on flight trajectory data to match contrails to the flights that created them. For this, three sources are primarily used: Automatic Dependent Surveillance–Broadcast (ADS-B) provided by the OpenSky Network or FlightAware, and aircraft position reports from the Federal Aviation Administration Traffic Flow Management System (FAA TFMS) for aircraft intersecting New York’s Flight Information Region (FIR).

Sarna et al. (2025), Geraedts et al. (2024), Barbosa (2024) use ADS-B data from FlightAware. Sarna et al. (2025) mentions that, for contrail attribution algorithms, flight data from FlightAware might be incomplete, since some operators request that their data be obfuscated or excluded from the dataset.

Chevallier et al. (2023) choose to use ADS-B data from the OpenSky Network, an open-source dataset that is not affected by exclusion requests from operators (Strohmeier et al., 2022). However, it can still miss some flights that are available on FlightAware because OpenSky’s coverage depends on the location and density of volunteer ground stations, which vary regionally and can result in gaps in reception. Alternatively, Gryspeerdt et al. (2024) use aircraft position reports from the FAA TFMS. Their study aims to discover how the operational differences of aircraft affect contrail lifetime. To do this, Gryspeerdt et al. (2024) need information about the aircraft type to correlate it with the engine’s parameters.

Secondly, weather data were used in all five studies to simulate the evolution of contrails. While four of the five studies relied on the ERA5 reanalysis dataset, Sarna et al. (2025) employed the ARCO-ERA5 dataset, which is derived from ERA5 but offers improved vertical resolution. Instead of 37 pressure levels, ARCO-ERA5 provides 137 model levels.

Satellite imagery is only the first step in a contrail-to-flight attribution algorithm. The raw images must be processed to identify contrails and distinguish them from other features such as clouds or rivers. This requires a dedicated contrail detection model.

Chevallier et al. (2023) trained an instance segmentation model on a synthetic dataset generated from GOES-16 imagery and simulations from the Contrail Cirrus Prediction (CoCiP) model. All other studies relied on existing convolutional neural networks (CNNs), which were adapted for contrail detection. Barbosa (2024) employed the CNN developed by Meijer, Kulik, et al. (2022) and post-processed the detected contrail regions into line segments. Both Sarna et al. (2025) and Geraedts et al. (2024) used the CNN developed by Ng et al. (2023). In contrast, Gryspeerdt et al. (2024) adopted a different approach, applying the Res-UNet CNN described by Zhang et al. (2018) to identify regions of contrails in satellite imagery, which were subsequently post-processed into “contrail objects” (COs) by preserving only contrails that were observed in two or more consecutive satellite images (separated by 5 minutes).

1.2.2. Algorithm Structure

The primary structural distinction between the algorithms lies in whether they perform contrail–flight matching across multiple temporal frames or only within a single frame. The former, referred to as multi-frame tracking, ensures that the match is consistent throughout the contrail’s observed lifetime, thereby reducing false positives from coincidental proximity on the two-dimensional frame. Among the reviewed studies, Chevallier et al. (2023), Gryspeerdt et al. (2024), and Sarna et al. (2025) implement multi-frame tracking, while Barbosa (2024) applies it in a subset of experiments. In contrast, Geraedts et al. (2024) restrict their approach to single-frame matching, which may result in ambiguous attributions for contrails observed near multiple flights.

The algorithm developed by Geraedts et al. (2024) is based on cost-function minimization: for each candidate flight, waypoints are advected forward in time using three-dimensional wind fields and then compared to the observed contrail’s position in a rotated coordinate system centred around the contrail. The cost function penalizes poor linear fit, large spatial shifts, and large rotation angles, with a time-dependent relaxation term for older contrails that would be affected by errors in the wind data. A match is declared if the total cost is below a set threshold.

Sarna et al. (2025) builds upon the method from Geraedts et al. (2024) by extending it into a multi-frame tracking algorithm. The process begins with the same cost function minimization method, after which similar matches are grouped according to their spatial shift. Implausible candidates are then rejected using temporal consistency checks across frames, removing cases where flights pass near an already formed contrail. This multi-frame tracking approach provides an additional consistency constraint by ensuring that a contrail is attributed to the same flight throughout its observed lifetime, as illustrated in Figure 1.3.

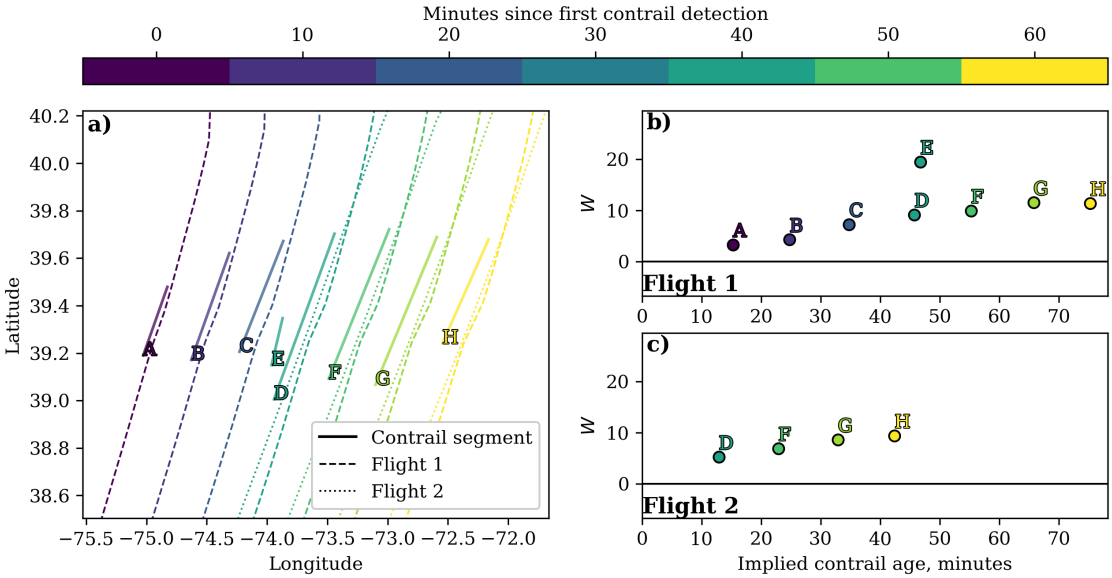


Figure 1.3: Multi-frame tracking process (Sarna et al., 2025)

Figure 1.3 illustrates the multi-frame tracking process. Panel **(a)** shows consecutive detections of a single contrail together with the advected trajectories of two candidate flights. Panels **(b)** and **(c)** present the perpendicular shift parameter as a function of the implied contrail age for each candidate. A smaller perpendicular shift usually corresponds to a better match score for each contrail detection at a given timestamp. Based on individual detections, Flight **2** would provide a better fit for detections **D**, **F**, **G**, and **H**, while detections **A**, **B**, and **C** would be attributed to Flight **1** under a single-frame tracking approach. In contrast, the multi-frame tracking method enforces temporal consistency across all detections, ultimately identifying Flight **1** as the candidate whose trajectory exhibits the highest overall agreement with the contrail over its entire observed lifetime, and therefore attributing the contrail to this flight.

The approach described in Chevallier et al. (2023) comprises two sequential algorithms. The first operates on a single timestamp, generating a list of potential flight candidates for each contrail polygon detected in that frame. Candidate selection is based on several filtering metrics: the mean Haversine geographical distance, the Hausdorff distance, and a custom metric that assigns additional weight to flights oriented parallel to the contrail polygon. The second algorithm advects each contrail polygon to the subsequent timestamp and checks for overlap with already-detected contrails. This chain-building process is repeated until no further viable successors are identified. A match score is then computed for each potential chain, and a mixed-integer linear programming (MILP) method is used to determine the optimal set of contrail–flight associations.

Gryspeerdt et al. (2024) employs a filtering strategy in which COs detected by their algorithm are matched to flights based that could have produced the contrail (i.e., the contrail did not appear before the respective flight) and on two geometric constraints: the spatial separation between the contrail and the flight track, and the alignment of their respective headings. Only contrails that are sufficiently linear and satisfy both criteria are considered matches. The method remains effective in regions of high flight density because overlapped or irregular contrails are removed during a preliminary filtering step.

Barbosa (2024) creates and applies eight probabilistic flight attribution algorithms. The best match is determined by the smallest difference in three *similarity measures*: position, heading, and altitude between the contrail and the candidate flight. The algorithms differ in their combinations of these measures and in whether they incorporate information from contrails detected at previous timestamps. The study compares the performance of all eight algorithms on a common dataset.

1.2.3. Algorithm Output and Evaluation

The approach defined in Chevallier et al. (2023) results in confidence-weighted match lists. Each contrail is matched to a flight based on the MILP optimization process, which assigns a confidence score to each match individually. The scores can then be used to rank match quality, allowing someone to filter results by their confidence threshold. Evaluation of the algorithm is also performed by the authors through the use of case studies. The performance observed in isolated areas was compared to that of a more congested region, resulting in more confidence in the former.

Geraedts et al. (2024) adopts a binary classification approach, where each flight segment is labelled as either *matching* or *not matching* the respective contrail. A match is assigned when the resulting cost is below a predefined threshold. Performance is evaluated manually by inspecting 1000 selected flight segments and verifying whether the attributed contrail was in fact produced by the flight.

Sarna et al. (2025) uses the same attribution score as in Geraedts et al. (2024), but redefines it within a multi-frame tracking framework. The algorithm produces score-based matches between multiple contrails and flights across consecutive frames. To evaluate the method, Sarna et al. (2025) constructs a synthetic dataset, *SynthOpenContrails*, in which contrails and their source flights are generated with known physical properties. Four performance metrics are defined, grouped into “per-contrail” and “per-flight” categories:

- **Per-contrail recall:** percentage of linear contrails correctly attributed to a flight
- **Per-contrail precision:** percentage of the algorithm’s contrail attributions that are correct
- **Per-flight recall:** percentage of flights that formed at least one contrail in *SynthOpenContrails* and were attributed at least one contrail (regardless of match validity)
- **Per-flight precision:** percentage of flights to which the algorithm attributed at least one linear contrail that also formed at least one linear contrail in *SynthOpenContrails*

As discussed earlier, Gryspeerdt et al. (2024) follows a much simpler approach. The method outputs a set of contrail–flight matches, which are accepted only if the predefined geometric criteria are satisfied. No formal evaluation or validation procedure is presented, as the study’s primary objective is to analyse the relationship between engine parameters and contrail lifetime rather than to benchmark attribution accuracy.

Barbosa (2024) presents a probabilistic output system in which every flight–contrail pair is assigned a likelihood score derived from Gaussian probability density functions (PDFs) corresponding to the previously defined similarity measures. The final output is a ranked list of matches based on these likelihood scores. Evaluation is conducted using a manually labelled ground truth dataset containing 180 distinct contrails and 1,980 manually attributed labels. Part of this dataset is used to perform parameter sweeps on four of the algorithms to determine the optimal Gaussian parameters for each similarity measure. Subsequently, all eight algorithms are tested with these optimal parameters, and their performance is evaluated using a *performance accuracy* metric, defined as the proportion of contrail–to-flight attributions that exactly match the ground truth labels.

1.2.4. Key Limitations

Some of the reviewed studies introduce novel approaches to the flight attribution problem, while others refine existing methods to address known limitations. Two key challenges are recurrent across the literature. First, wind uncertainty is inherent in all meteorological datasets and can lead to compounding errors when advecting flights over extended periods. This may cause misalignment between flight trajectories and contrail detections, ultimately resulting in incorrect attributions. Second, until recently, there has been a lack of ground-truth or evaluation datasets, which has hindered the ability to rigorously assess algorithm performance and compare results across studies. Sarna et al. (2025) partially addresses this limitation through the development of *SynthOpenContrail*, providing a benchmark dataset that enables more systematic evaluation and comparison of attribution methods.

1.3. Objectives

There are two hypotheses I would like to test to challenge the current limitations of contrail-to-flight attribution algorithms:

- **Wind error correlations:** Nearby contrails should exhibit similar displacements relative to their attributed advected flight trajectories because wind errors should be similar in direction and magnitude over small spatio-temporal scales.
- **Vertical extent of ISSRs:** Contrails that are close to each other should be assigned to flights that do not show large differences in altitude to account for the small vertical extent of ISSRs.

Therefore, the objective of this research is to design, implement, and assess a physically consistent contrail-to-flight attribution algorithm that improves accuracy and consistency between multiple contrail-flight pairings in the same region compared to current independent-pair approaches.

The specific objectives of the thesis are to:

- Define physical constraints based on atmospheric processes, including correlated wind prediction errors and the limited vertical extent of ISSRs
- Implement and compare the algorithm with a baseline method (Geraedts et al., 2024)
- Create a new synthetic ground-truth dataset from calibrated ERA5 data (Meijer, 2024)
- Evaluate performance in terms of attribution accuracy and computational efficiency
- Provide recommendations on the applicability of physically consistent methods and their potential integration into contrail mitigation strategies

Chapter 2
Methodology

In this chapter, I describe the methodology developed to implement and evaluate two variations of the same physically consistent contrail-to-flight attribution algorithm. Section 2.1 presents where flight trajectory and meteorology data are sourced from, as well as other post-processing that this data is subjected to. Section 2.2 explains how the data is used to construct two types of synthetic datasets used by the algorithms, together with known contrail–flight associations that are used for evaluation purposes. Section 2.3 outlines the baseline geometric attribution algorithm used as the primary form of comparison for evaluating the proposed algorithms. Section 2.4 introduces the two variations of the proposed physically consistent attribution algorithm, formulated as a Mixed-Integer Linear Programming (MILP) problem. Finally, Section 2.5 presents the choice of datasets used to test the algorithms and defines the metrics used to evaluate them both.

2.1. Data Acquisition

In this section, I describe the types of data required for the contrail-to-flight attribution algorithms and the post-processing applied. Subsection 2.1.1 explains how raw flight trajectories are filtered, resampled, and transformed to ensure consistency and sufficient resolution for advection. Subsection 2.1.2 explains how meteorological reanalysis datasets are selected and how the wind fields are calibrated to be used subsequently for the synthetic datasets.

2.1.1. Flight Data

Attribution algorithms use flight trajectories that can be advected forward in time. In this thesis, flight trajectories are derived from ADS-B telemetry obtained via the *Contrails API* (*Contrails API* 2025). ADS-B data consist of telemetry broadcast by aircraft, including position (latitude, longitude), altitude, timestamp, and heading, typically transmitted every second (Sun, 2021).

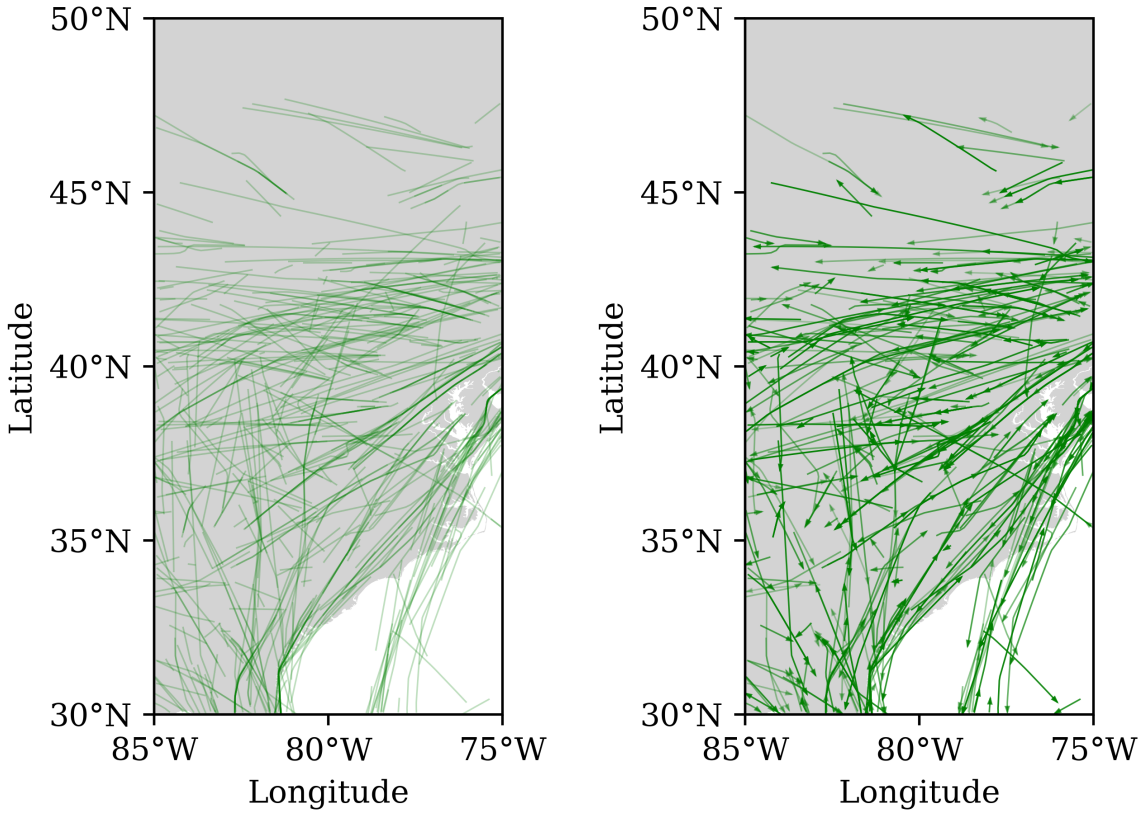
For each case study, ADS-B data is downloaded for a selected time window and geographical region. Unlike most approaches in the literature, where contrails are first detected in satellite imagery and corresponding flight data is then retrieved for the same period, this thesis reverses the process: flight data serves as the starting point, from which contrails are synthetically generated. This setup allows the analysis to be applied to any region or time of interest, independent of the real presence of visible contrails.

After collection, the ADS-B data is filtered to retain only flights at altitudes relevant for contrail formation. Persistent contrails rarely form below 8 km or above 14 km (U. Schumann, 1996; Meijer, Eastham, et al., 2024), so waypoints outside this range are excluded. The geographical regions for which data is downloaded can be chosen arbitrarily. However, for this thesis, Section 2.5 explains the selection in more detail.

The first post-processing step is to resample the flight data into one-minute-spaced waypoints. Later, in Section 2.3, it is explained how the flight trajectories are segmented to ensure a fair comparison with the algorithm presented by Geraedts et al. (2024). Afterwards, two preliminary filters are used to ensure that only meaningful trajectories are retained for subsequent steps. Trajectories with less than two minutes of flight time are removed as they do not provide enough temporal extent to form a representative track at contrail altitudes. What is more, trajectories containing fewer than two waypoints are discarded. At

least two waypoints are required to compute geometric distances such as the Hausdorff-based metrics later used in the thesis.

A final quality-control step removes implausible trajectories. Flights are removed based on vertical rates or, if not available, on the altitude difference between two consecutive waypoints. A flight is removed if any waypoint has $|\texttt{vertRate}| > 1920\ \mathrm{ft\,min^{-1}}$. This specific limit was chosen based on *Aircraft Performance Database* (2025), allowing for average rates during climb and descent phases of flight for commercial aircraft. If `vertRate` is missing for any waypoint, the altitude change to the previous point is checked; the flight is removed if $|\Delta z| > 600\ \mathrm{m}$ for that step (in accordance with the distance flown in a minute of flight if $|\texttt{vertRate}| > 1920\ \mathrm{ft\,min^{-1}}$).



(a) Raw ADS-B flight trajectories

(b) Resampled and filtered flight trajectories

Figure 2.1: Example of one hour flight trajectory data for a $20° \times 20°$ region

Figure 2.1 illustrates one hour of ADS-B trajectories. Figure 2.1a presents 542 flights (raw ASD-B data), while Figure 2.1b shows 429 trajectories that remained after the post-processing was completed. Arrows show the flight direction.

2.1.2. Meteorological Data

In addition to flight trajectories, attribution algorithms require access to meteorological data describing the atmosphere through which aircraft fly and contrails are advected. In this thesis, reanalysis data is obtained from the ERA5 ensemble product at multiple pressure levels (Hersbach et al., 2020). The variables retrieved are air temperature, specific humidity, and the three wind components (eastward, northward, and vertical velocity).

For each case study, meteorological data is downloaded for a period longer than the corresponding flight data window. For instance, when analysing flights between 08:00 UTC and 10:00 UTC, ERA5 data is collected from 07:00 UTC to 13:00 UTC. This extended window ensures that sufficient data is available for advection, accounts for contrails that may persist beyond the active flight period, and

provides a consistent basis for modelling contrail evolution over at least two hours, in line with the typical upper bound of the observation window reported by Chevallier et al. (2023) and Geraedts et al. (2024).

Meteorological ensemble forecasts such as ERA5 are known to be under-dispersed, meaning that the spread of the ensemble members underestimates the actual atmospheric uncertainty (Wilks, 2006). This poses a challenge for trajectory advection, where small wind errors can accumulate into significant spatial displacements that do not reflect the actual uncertainty in a trajectory's location. To address this limitation, a calibration procedure following *Appendix D* of Meijer (2024) is applied to the ERA5 wind ensemble. The calibration adjusts both the ensemble mean and the standard deviation of the eastward and northward wind components, producing wind fields that more accurately represent the uncertainty of the real atmosphere. In this thesis, only a single calibrated ensemble member is retained for contrail generation, while the ensemble mean of all remaining members is computed to generate the candidate trajectories. This setup ensures that contrail advection is based on a physically consistent wind field, while the advected trajectories incorporate the type of wind uncertainty expected in practice.

Finally, in addition to wind and temperature, specific humidity is also stored. This variable is later used to identify ice supersaturated regions that also satisfy the Schmidt-Appleman criterion, which are required to generate synthetic contrails in the following section of this thesis.

2.2. Synthetic Dataset

In this section, I explain how the two types of synthetic datasets are constructed and how they are used for evaluation. Subsection 2.2.1 describes the procedure for creating synthetic contrails and candidate flights using meteorological data, while Subsection 2.2.2 provides a representative case to illustrate the resulting displacement between contrails and their candidates.

2.2.1. Generation

To evaluate attribution algorithms under controlled conditions, two types of synthetic datasets are generated. Both datasets are constructed by advecting real flight trajectories using ERA5 ensemble meteorology, but they differ in the structure of the starting trajectories:

- **Whole initial trajectories:** this dataset is used together with one variation of the contrail-to-flight matching algorithm to simulate a scenario in which whole flights are attributed to contrails
- **Segmented initial trajectories:** this dataset is used together with the other variation of the contrail-to-flight matching algorithm to allow for a fairer comparison with the baseline algorithm

Motivation

Real contrails observed in satellite imagery are displaced from the aircraft's initial position due to advection caused by wind. However, the real, certain wind field is unknown. Using a synthetic dataset allows the user to introduce controlled, realistic wind errors and generate a ground-truth-based dataset with known contrail-flight pairs ready for evaluation.

Common Structure

Two types of trajectories are generated from the initial flight path using two different meteorological datasets:

- **Contrails**, generated using a single calibrated ensemble member of ERA5, which is assumed to recreate true atmospheric uncertainties and therefore approximates the real location of contrails as they would appear in satellite imagery
- **Candidates**, generated using the ensemble mean of the remaining nine ensemble members, which represents the displacement that existing algorithms show between real contrails and advected flight traces

As stated before, this introduces a controlled but physically plausible displacement between contrails and their candidate flights, reflecting the uncertainty inherent in meteorological data.

Several conditions are applied to both types of datasets to ensure only flights capable of producing observable contrails are retained:

- **Flight timing relative to the satellite frame:** Flights that occur entirely after the chosen frame time are excluded, since their contrails would not yet be visible in satellite imagery for the chosen frame time
- **Observation window:** Only contrails persisting between 30 minutes and 2 hours are retained; shorter lifetimes are excluded because contrails are not optically thick enough to be observed, while older ones cannot be distinguished from surrounding cloud structures (Chevallier et al., 2023; Geraedts et al., 2024)
- **SAC and ISSR:** Only flights passing through ISSRs satisfying SAC are considered, ensuring the physical feasibility of persistent contrail formation

The ISSR/SAC filtering step is essential because real, persistent contrails can only form if the surrounding atmosphere permits the water vapour present in the engine's exhaust to freeze into ice crystals. If the ambient air is too warm or subsaturated, the exhaust plume evaporates within seconds, leaving no visible trace. Satisfying SAC alone would not guarantee persistence, since additional conditions such as ice supersaturation are required for contrails to grow and become detectable (U. Schumann, 1996). In this dataset, this limitation is addressed by pairing ISSR/SAC filtering with the observation-window constraint (30 minutes – 2 hours). Together, these criteria ensure that only flights capable of forming persistent contrails observable by satellite imagery are retained, thus improving the physical realism of the synthetic dataset.

To add more details, all initial flight waypoints are tested to see if they satisfy the ISSR/SAC filter. In practice, it was found that these waypoints can occur in multiple disconnected fragments. To avoid creating unrealistic or discontinuous synthetic contrails, small gaps (at most one missing waypoint between two valid points) caused by noise or interpolation are filled. However, larger discontinuities that indicate genuinely unsuitable atmospheric conditions are preserved.

The synthetic dataset creation process can be seen by the flowchart in Figure 2.2.

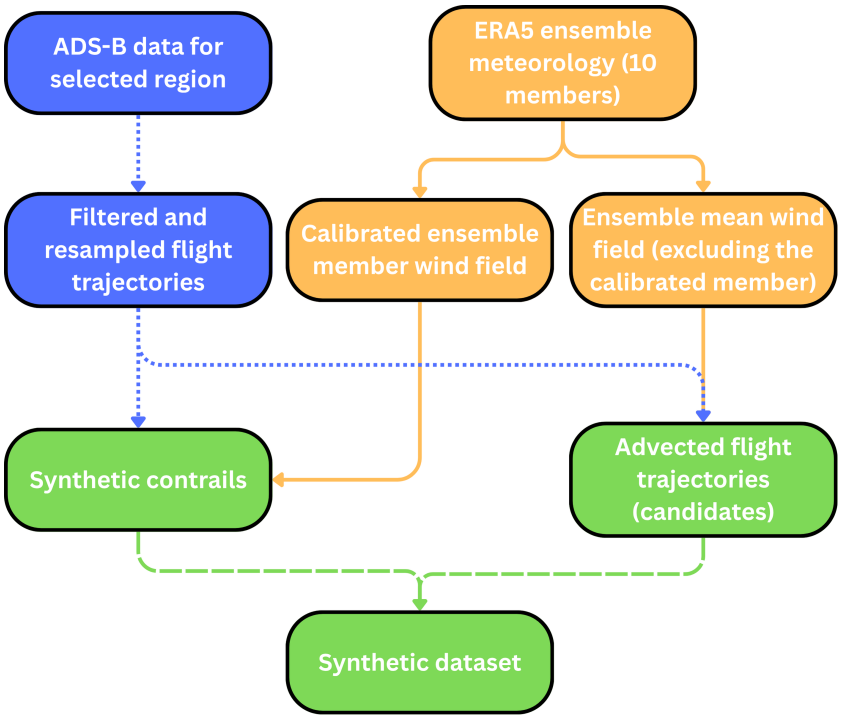


Figure 2.2: Synthetic dataset creation process

Why Use 2 Different Datasets?

As mentioned before, two synthetic datasets are generated in this work — one based on whole initial flight trajectories and one based on segmented trajectories. Each dataset serves a distinct methodological purpose and enables a different type of evaluation of the attribution algorithms.

Using whole, unsegmented flight trajectories allows the attribution algorithm to work on the complete trajectory of each aircraft, rather than on artificial segments. This could represent a future approach in which the algorithm uses complete flight tracks to determine which flight corresponds to the observed contrail. Therefore, evaluating whole trajectories tests the algorithm's ability to recover the correct, complete match, offering insight into its behaviour under more realistic conditions.

In contrast, prior work, such as Geraedts et al. (2024), relies on shorter, fixed-length flight segments presented to the attribution algorithm. To enable a fair comparison with these established baselines, a second synthetic dataset is constructed where each full trajectory is divided into such shorter segments. Each segment is then treated as a separate candidate flight. This dataset reproduces the assumptions, structure, and difficulty profile of earlier attribution frameworks, allowing the performance of the new algorithm to be directly benchmarked against existing methods.

Resulting Datasets

The resulting datasets consist of contrail and candidate trajectories, each represented as time-ordered waypoints defined by longitude and latitude. For candidate trajectories, additional variables such as altitude, timestamp, and age are retained, as these are available from flight data but cannot be inferred from satellite imagery for contrails.

An initial trajectories dataset is also maintained, which stores flight-level attributes (flight identification and aircraft type) and visibility flags indicating whether the corresponding contrail would be observable in satellite imagery according to the filtering rules or whether the initial trajectory was advected into a candidate flight. These flags are later used during evaluation but are not stored as attributes of the advected trajectories themselves. Each contrail is explicitly linked to its originating flight, ensuring that ground truth correspondence is preserved for evaluating attribution algorithms. Table 2.1 presents the structure of the initial trajectories dataset.

Variable	Type	Description
`callsign`	string	Unique flight callsign
`flight id`	string	Unique flight identifier
`aircraft type`	string	Aircraft type
`visible contrail`	boolean	Whether the flight formed a contrail
`visible candidate`	boolean	Whether flight appears in candidate set

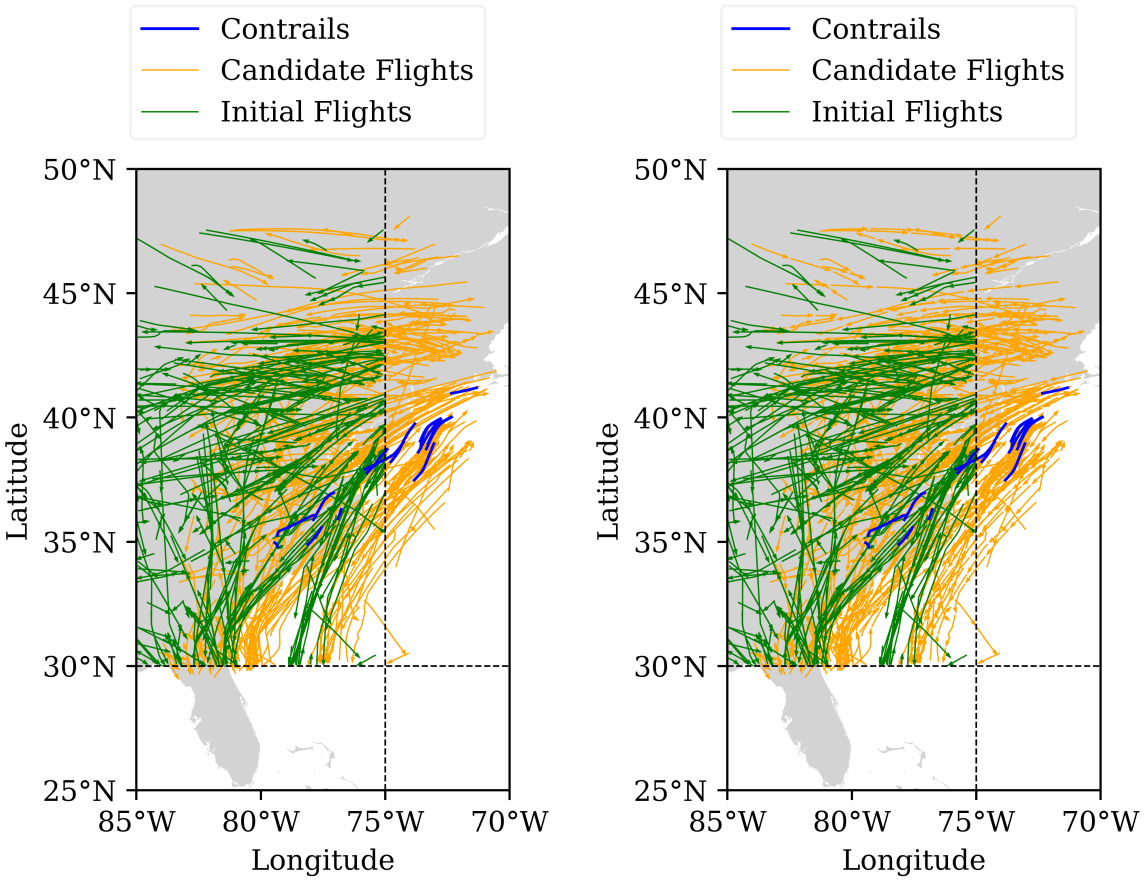
Table 2.1: Variables contained in the flight dataset

The key assumption underlying this framework is that differences between the calibrated ensemble member and the ensemble mean are large enough to introduce meaningful but manageable wind errors, without producing inconsistencies that would make attribution impossible. This calibration not only grounds the synthetic dataset in realistic meteorological variability but also provides a robust benchmark for testing attribution algorithms.

2.2.2. Example

To illustrate the construction of the synthetic dataset, I present one representative case from a region on the East coast of the United States of America. The flight data spans from 15 April 2025 10:00 UTC to 11:00 UTC, and the corresponding meteorological data covers 15 April 07:00 UTC to 15 April 15:00 UTC. Aircraft cruise altitudes range between 10 - 12 km (a shorter altitude range was chosen to reduce the dataset size). The satellite frame time is fixed at 12:00 UTC on the same day.

As mentioned briefly before, after applying the post-processing of the flight data, 429 initial flights are retained in the domain bounded by longitudes $[-85°, -75°]$ and latitudes $[30^o, 50^o]$. From these, 15 are identified as producing observable contrails, while 429 remain only as candidates. Figure 2.3 shows the resulting datasets: contrails (blue) appear spatially displaced from their corresponding candidates (orange) due to the use of different meteorological inputs for advection, while the initial flight trajectories are shown in green.



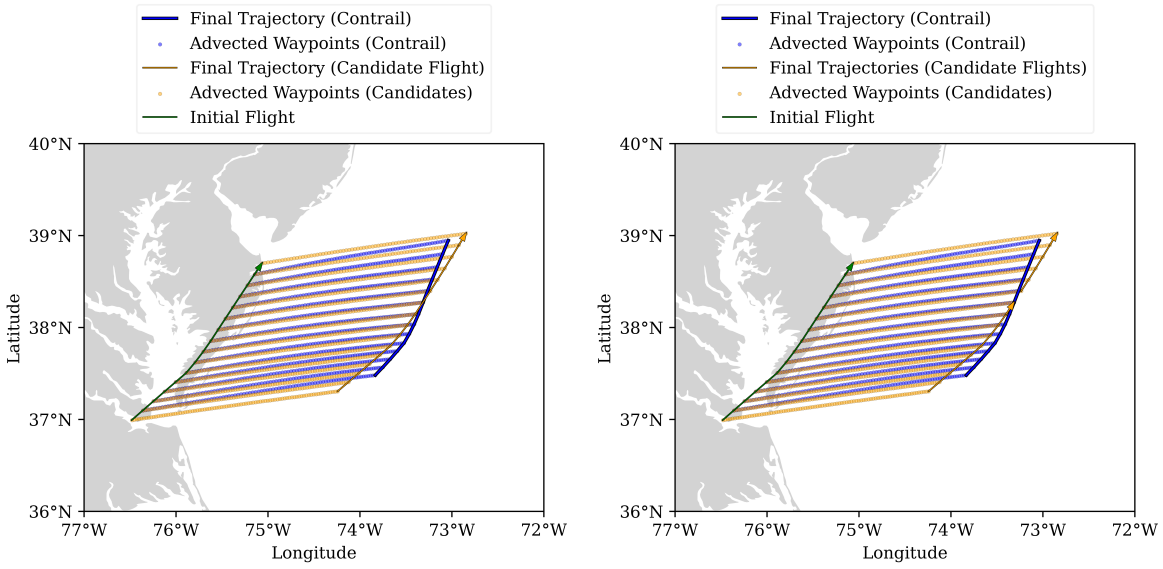
(a) Whole trajectories as candidates

(b) Segmented trajectories as candidates

Figure 2.3: All initial trajectories, contrails, and candidate flights in the North Atlantic region at 12:00 UTC on 15 April 2025

Figure 2.3a presents the synthetic dataset consisting of 429 candidates advected from each whole initial flight. Figure 2.3b shows 1 131 candidates from segmented initial flights.

A closer look at a single flight reveals how displacement develops during advection. Both types of datasets are presented for better understanding. Figure 2.4 shows the initial trajectory (green), the contrail (blue), and the candidate (orange). Intermediate advected waypoints for the contrail and candidates are also shown. While the contrail and candidate originate from the same initial trajectory, differences between the calibrated ensemble member and the ensemble mean cause a systematic separation in their positions.



(a) Whole trajectories

(b) Segmented trajectories

Figure 2.4: Advection process for a single flight

Figure 2.5 shows the meteorologically valid region where persistent contrails can form, defined by ISSR/SAC filter. The synthetic contrail (blue) is generated because the initial flight (green) is present inside such a valid region.

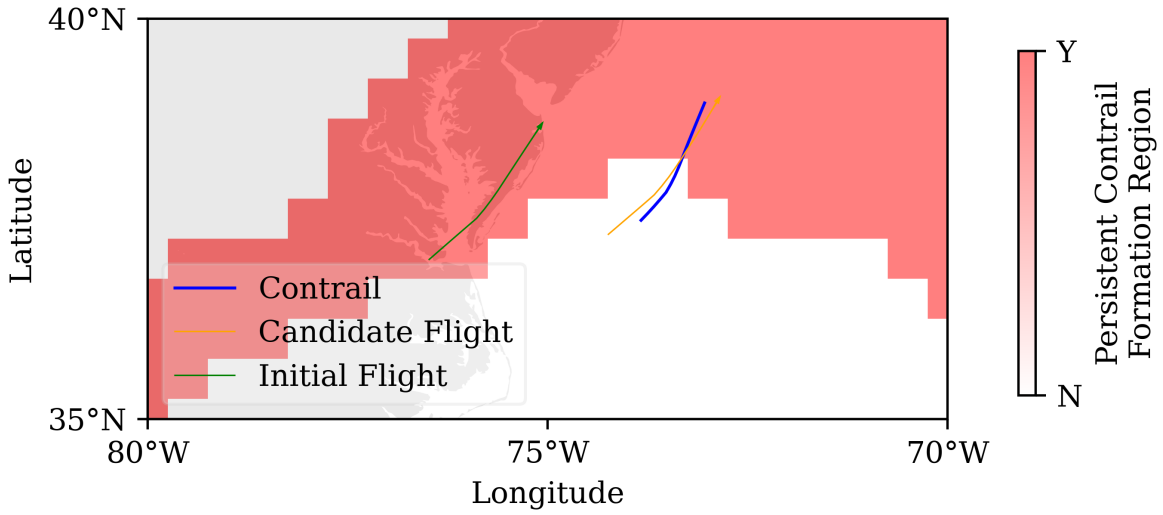


Figure 2.5: Example of persistent contrail formation regions

Figure 2.6 illustrates how synthetic contrails and candidate trajectories are generated using two different wind fields. Starting from the same initial flight track (dashed line), the trajectory is advected once using the calibrated ensemble member (left panel) to produce the synthetic contrail, and once again using the ensemble-mean wind field of the remaining members (right panel) to produce the candidate trajectory. The difference between the two resulting paths represents the controlled wind-field uncertainty that the attribution algorithm must resolve.

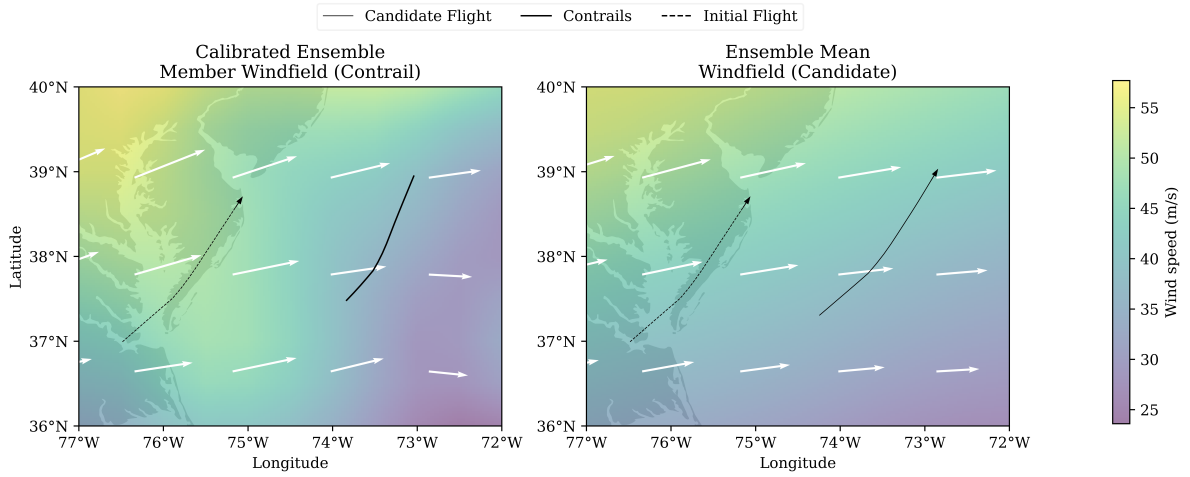


Figure 2.6: Wind data differences between the two meteorological datasets

2.3. Baseline Contrail-to-Flight Matching Algorithm

To provide a benchmark for evaluating the physically consistent attribution algorithm developed in this thesis, I implemented the geometric attribution method of Geraedts et al. (2024). This method matches contrails to flights using only geometric criteria, under the assumption that contrails appear as straight-line segments in satellite imagery and can be matched to flights by aligning candidate flight trajectories to these segments through geometric transformations.

Contrails are first detected automatically from GOES-16 imagery using the computer vision approach of Ng et al. (2023). Flight trajectories are obtained from ADS-B data (FlightAware) and preprocessed to one waypoint per minute. Each trajectory is then divided into 10-minute segments, corresponding to a typical distance of about 150 km, consistent with the expected scale of flight paths within ice-supersaturated regions (K. Gierens et al., 2000). Segments with fewer than six waypoints are discarded, and all waypoints below 7 km altitude are excluded, as persistent contrail formation is rare below this threshold. The retained segments are subsequently advected forward for two hours, covering the next eleven GOES-16 image frames (see leftmost plot of Figure 1.3).

For geometric alignment, each contrail is first rotated into a local (v, w) coordinate system such that the contrail lies horizontally along the v-axis with its midpoint at the origin. Candidate flight segments are transformed into the same coordinate system. A further transformation is then applied to shift and rotate candidates to minimize their distance from the contrail segment in this rotated frame:

$$\begin{aligned} w &\to (w + W)\cos\theta + (v + V)\sin\theta \\ v &\to (v + V)\cos\theta - (w + W)\sin\theta \end{aligned} \tag{2.1}$$

where V, W, and θ are the shift and rotation parameters. These parameters are optimized by minimizing the cost function defined by Geraedts et al. (2024):

$$\text{match error} = \underbrace{C_{\text{fit}} \frac{1}{N} \sum_{i=1}^{N} w_i^2}_{\text{fit term}} + \underbrace{C_{\text{shift}}(V^2 + W^2)}_{\text{shift term}} + \underbrace{C_{\text{angle}}(1 - \cos\theta)}_{\text{angle term}} + C_{\text{age}} \tag{2.2}$$

where C constants are created so that each term would be approximately 1. Equation 2.2 combines four terms:

- **Fit term:** penalizes deviations from linearity in the candidate trajectory, since non-linear candidates are unlikely to generate linear contrails
- **Shift term:** accounts for systematic displacement due to wind bias in meteorological data

- **Angle term:** captures local variations in wind error that can cause rotations of advected trajectories
- **Age term:** adjusts the tolerance of the shift and angle terms as contrails age, reflecting that older contrails accumulate larger displacements

The resulting cost function provides a match error, with lower values indicating a closer correspondence between a contrail and a candidate flight. Following Geraedts et al. (2024), a candidate is considered a match if the match error is below 3. In cases where multiple candidates are matched to the same contrail, the lowest-error candidate is taken as the true match, while other candidates within one unit of the minimum score are retained as possible alternatives.

2.4. Physically Consistent Contrail-to-Flight Matching Algorithm

The baseline method in Section 2.3 treats each contrail independently and relies exclusively on geometric similarity. In contrast, the algorithms proposed here introduce physical consistency constraints across all contrails observed within the same scene. The underlying assumption is that, within a limited spatial and temporal domain, residual wind errors that displace advected flight trajectories from their corresponding contrails should be approximately uniform in both magnitude and direction. Accordingly, the algorithm selects matches such that the resulting advection error vectors are mutually consistent.

2.4.1. Problem Setup

Let $\mathcal{C} = \{1, \dots, N\}$ denote the set of contrails and $\mathcal{F} = \{1, \dots, M\}$ the set of candidate flights. For each pair $(i, j) \in \mathcal{C} \times \mathcal{F}$, a two-dimensional *advection error vector* $\mathbf{e}_{ij} \in \mathbb{R}^2$ is precomputed, representing the displacement required to align candidate j with contrail i. The attribution problem is then formulated as a mixed-integer linear program (MILP), which assigns at most one candidate to each contrail while simultaneously encouraging the selected error vectors to align with an area-specific background advection vector.

2.4.2. Contrail Clusters

To accommodate spatially varying advection errors, contrails are first partitioned into K spatial areas. Let N be the number of contrails and $\{\boldsymbol{c}_i\}_{i=1}^{N}$ their geographic centroids. Define:

$$D_{\max} \;=\; \max_{1 \le i < j \le N} \; d(\boldsymbol{c}_i, \boldsymbol{c}_j) \tag{2.3}$$

where $d(\boldsymbol{c}_i, \boldsymbol{c}_j)$ is the great-circle (Haversine) distance between the two centroids (in kilometres). The number of areas K is chosen using:

$$K \;=\; \max\!\left(\text{round}\left(\sqrt{N}\right),\; \text{round}\left(D_{\text{max}}/s\right) \right) \tag{2.4}$$

where $s > 0$ is the maximum distance from one side of the cluster to the other (chosen to be $s = 100$ km based on experiments). There are two rules that make up the clustering equation above:

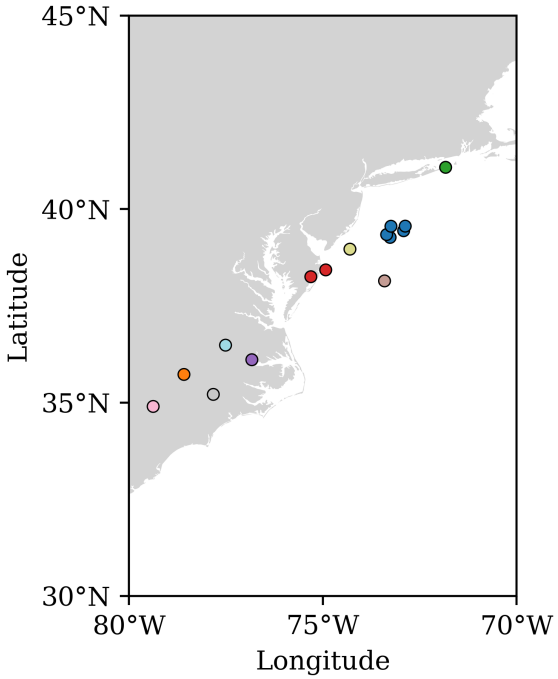
- **Count-based:** When N is very small, there is no need for a large number of clusters; however, when N is very large, the number of clusters should grow, but not directly proportional to N, as that would result in clusters of single contrails.
- **Distance-based:** In some cases, contrails can be spread widely across a given region. To prevent the situation where all contrails are grouped together over a large area, this second rule enforces a minimum number of clusters based on spatial proximity.

The number of clusters is then chosen to be the maximum value of the results of the two rules. There is one possible scenario in which the number of clusters resulting from the equation would be larger than the number of contrails in the dataset. In this case, the number of clusters would become the number of contrails directly. However, this would not be favourable for the algorithms, as they are based on the hypothesis that contrails in the same clusters should share similar advection errors.

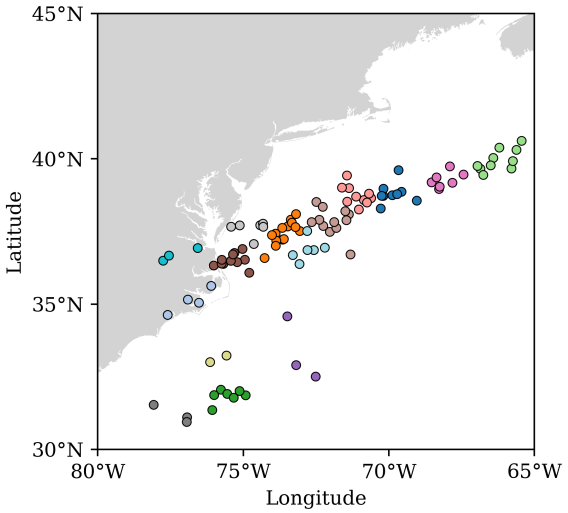
Given K, contrails are assigned to clusters by K-means (Jin et al., 2011), and each contrail i is labelled with an area index $a(i) \in \{1, \dots, K\}$. Each area k is associated with a background advection vector

$\boldsymbol{\mu}_k \in \mathbb{R}^2$, estimated during optimization so that the error vectors of matches in an area k are mutually consistent.

The results of the clustering method can be seen in Figure 2.7. For Figure 2.7a, the contrails' centroids are represented by the coloured circles where the colour represents the cluster they are attributed to. It can be seen that the situation described earlier (some clusters contain only one contrail) appears here. In Figure 2.7b, it can be seen that all clusters contain at least two contrails. Having only one contrail per cluster opposes the main hypothesis of the thesis (that the implied advection error should be consistent for neighbouring contrails). Therefore, for the latter parts of the thesis, datasets are created with this in mind.



(a) Clustering results for 15 contrails from the previously presented dataset



(b) Clustering results for 107 contrails from a dataset used later in the performance assessment

Figure 2.7: Clustering results for two datasets

2.4.3. Advection Error Vector Computation

For a contrail–candidate pair (i, j), let the two trajectories be c_i (for the contrail) and f_j (for the candidate). The advection error vector $\mathbf{e}_{ij} \in \mathbb{R}^2$ measures the 2D displacement that best represents how far c_i is from f_j.

All coordinates are stored in geographic form but can be projected to a metric system for distance computations. Optional contrail-age scaling of displacement vectors is added to all computation methods to represent the wind error vector explicitly.

In the process of finding an efficient and physically consistent vector computation method, several variations were formulated and tested with the algorithm. Each of these formulations expresses the advection error similarly, but they differ in their approach to the problem. For a better understanding of each error vector computation method, Figure 2.8 and Figure 2.9 are provided.

Geo

The simplest option is to take the displacement between the mean geographic positions of the contrail-candidate pair. While stable and computationally cheap, this measure ignores curvature and relative alignment along the trajectories, which proved unreliable when contrails and candidates had different lengths or shapes.

Hausdorff
This measure selects the displacement at the contrail point farthest from the candidate trajectory, paired with its nearest candidate point. It is robust to local noise and captures global misalignment, but was found to be overly sensitive to single endpoints or sharp turns.

Symmetric Hausdorff
Here, the Hausdorff distance is computed in both directions (contrail-to-candidate and candidate-to-contrail), and the larger is retained. This avoids bias toward one curve but remains governed by an extreme point, which can exaggerate errors caused by isolated segments.

Balanced Hausdorff
To mitigate the influence of extremes, this method averages the displacement vector at the Hausdorff point with that at the closest-aligned point, both taken in the contrail-to-candidate direction. This reduced sensitivity to outliers, but the method was not symmetric and could still bias results depending on the chosen reference curve.

Balanced, Symmetric Hausdorff
This method reduces the dominance of extreme points inherent in the pure Hausdorff formulation while avoiding the directional bias of the non-symmetric balanced version. It captures both global misalignment and local curve agreement, making it more stable than the symmetric Hausdorff approach in the presence of short outliers or partial overlaps.

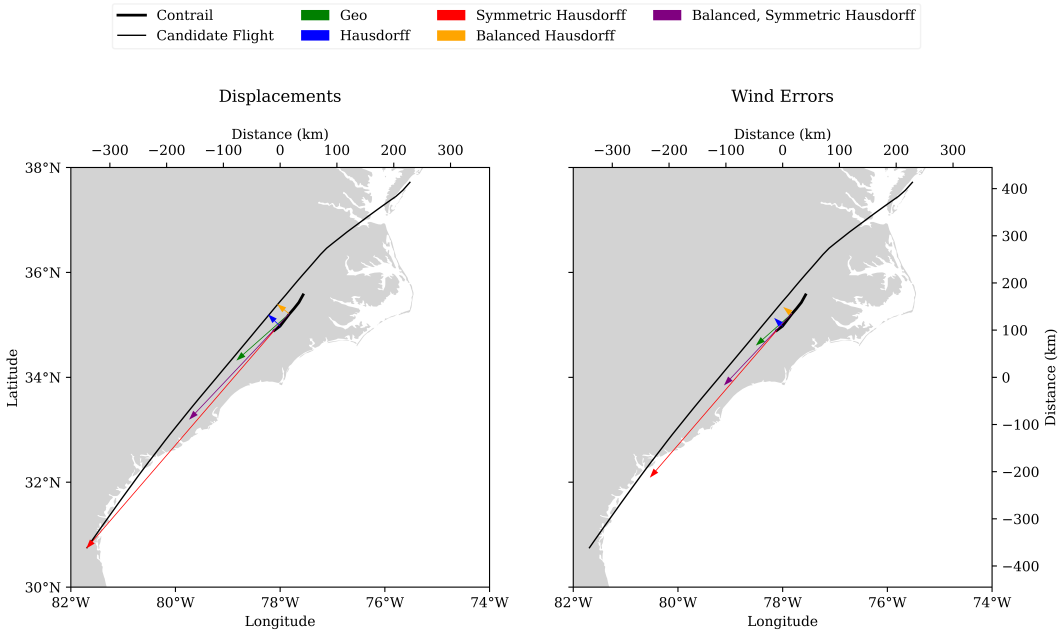


Figure 2.8: One example of the types of advection error vectors computation methods

Figure 2.8 shows a contrail-candidate pair that is actually a correct match in the synthetic dataset (i.e., the contrail and the candidate originate from the same initial flight) and the error vectors between the two computed with each of the five computation methods described earlier. The left plot shows the displacements between the trajectories, while the right plot presents the accumulated wind errors by using the contrail-age scaling method described earlier. From this figure, it can be seen that the *Geo* and *Hausdorff* methods retrieve the expected magnitude of both displacement and wind error between the two trajectories when the candidate is advected from the whole flight.

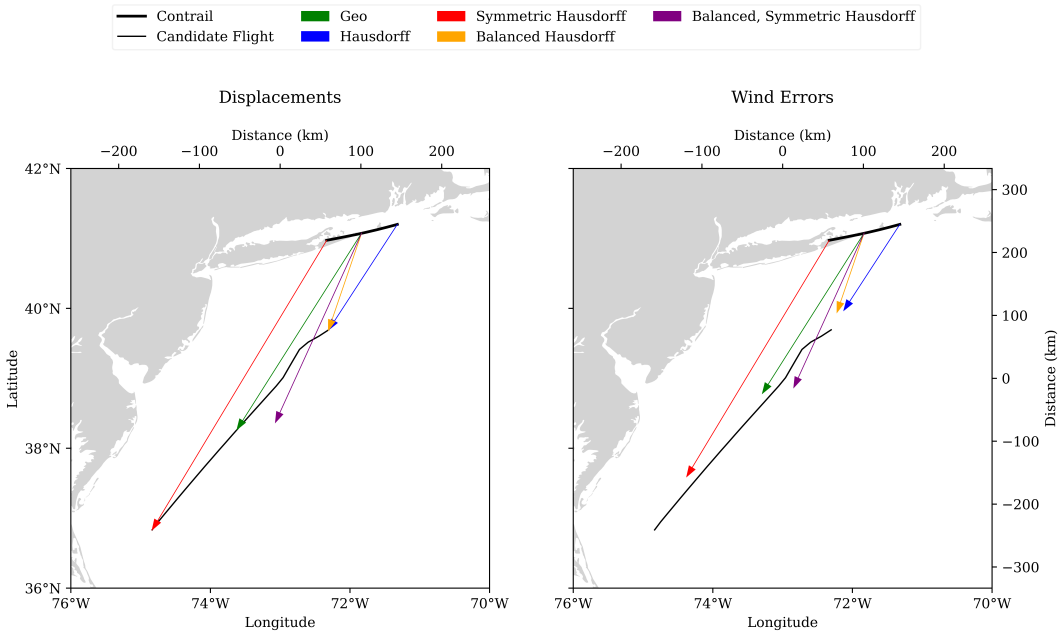


Figure 2.9: One more example of the types of advection error vectors computation methods

Figure 2.9 presents two other trajectories that are not actually labelled as a match in the synthetic dataset. It can be seen that, in this case, the *Geo* and *Hausdorff* methods are quite different. The latter manages to also retrieve the shift in the heading of the trajectories as well as the correct magnitude of the vectors.

2.4.4. Problem Formulation

Given the precomputed advection error vectors $\mathbf{e}_{ij}$, the altitude differences between candidates and their area-level ISSR height reference, and the optional baseline geometric scores, the contrail-to-flight attribution problem is formulated as a weighted mixed-integer linear program.

Each contrail $i \in \mathcal{C}$ must be matched with exactly one candidate flight $j \in \mathcal{F}$, and each candidate may be used at most once. Let $x_{ij} \in \{0, 1\}$ denote whether candidate j is assigned to contrail i. Let $\boldsymbol{\mu}_{a(i)}$ be the background advection vector for the area to which the contrail i belongs. The following objective function is used:

$$\min_{\{x_{ij}\}} \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{F}} \left(\lambda_1 \cdot \underbrace{\|\mathbf{e}_{ij} - \boldsymbol{\mu}_{a(i)}\|}_{\substack{\text{advection error} \\ \text{consistency penalty}}} + \lambda_2 \cdot \underbrace{\max\left(0, \frac{|\, z_j - \zeta_{a(i)} \,| - h}{h}\right)}_{\text{altitude consistency penalty}} + \lambda_3 \cdot G_{ij} \right) x_{ij} \tag{2.5}$$

subject to:

$$\sum_{j \in \mathcal{F}} x_{ij} = 1 \quad \forall\, i \in \mathcal{C}, \qquad \sum_{i \in \mathcal{C}} x_{ij} \leq 1 \quad \forall\, j \in \mathcal{F}, \qquad x_{ij} \in \{0, 1\} \tag{2.6}$$

where z_j is the altitude of candidate flight j, $\zeta_{a(i)}$ is the mean of the altitudes of all candidates that were previously assigned to contrails in the area that contains contrail i, and h is a scaling factor chosen as the average height of an ISSR (500 m). G_{ij} represents the match score between contrail i and candidate flight j resulted from the baseline algorithm implementation. λ_1, λ_2, λ_3 are variable weights used to give more or less importance to each objective function term. The objective favours assignments whose implied displacement vectors cluster tightly around their area vector. This MILP is solved with PuLP (*Optimization with PuLP — PuLP 3.3.0 documentation* 2025) - a programming modeller written in Python.

The area background vectors $\boldsymbol{\mu}_k$ are not known in advance but are estimated iteratively together with the contrail–candidate assignments. The procedure alternates between two steps:

- **Assignment step:** Given a current estimate of the background vectors $\{\boldsymbol{\mu}_k\}$, the MILP is solved to produce a set of assignments $\{x_{ij}\}$
- **Re-estimation step:** For each area k, the background vector is updated as the average of the advection error vectors corresponding to the currently assigned pairs:

$$\boldsymbol{\mu}_k = \frac{1}{|\mathcal{S}_k|} \sum_{(i,j)\in\mathcal{S}_k} \mathbf{e}_{ij}, \quad \mathcal{S}_k = \{(i,j) \mid i \in \mathcal{C},\ j \in \mathcal{F},\ a(i) = k,\ x_{ij} = 1\} \tag{2.7}$$

These two steps are repeated until convergence, which is defined as no further changes in assignments or reaching a maximum iteration limit. In practice, convergence is typically achieved in fewer than three iterations for the case studies considered in this thesis.

The structure of the algorithm is also presented in Figure 2.10.

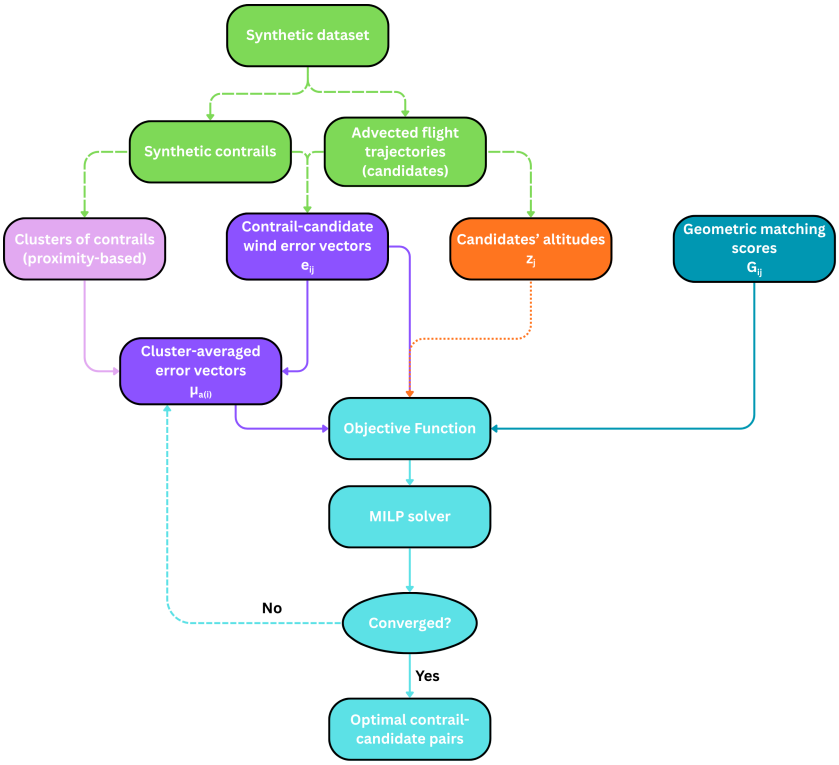


Figure 2.10: Optimization flowchart

Seeing as there are two different types of synthetic datasets, there is also a need to see how each optimization parameter might affect the performance on each of the dataset types. Therefore, in the following section, several parameters will be varied to see the influence on performance.

2.4.5. Parameter Sensitivity and Optimization

The physically consistent attribution algorithm contains several customizable components that influence its behaviour: the choice of advection error computation method, the use of advection age or conversion to the metric system for the computation method, the ISSR height approximation, and the three weights of the objective function terms. To identify parameter combinations that yield stable and accurate matches for different types of trajectory data, a structured grid search was performed across both full candidate trajectories and segmented candidate trajectories.

The ISSR height parameter was tested for values of 300 and 700 m as well. All the advection error computation methods were tested together with their respective choices of advection age scaling or metric system conversion. The three penalty weights were varied in the range $[0, 1]$ with the constraint

that they sum to 1. What is more, the individual weights were also allowed to be 0 to test how the algorithms would perform without each penalty term. This idea will be tested in more detail in chapter 3 through an ablation study.

The tests were performed on the dataset described in Subsection 2.2.2. For both types of datasets, the advection error computation method described in Subsection 2.4.3 proved to yield the most accurate results in combination with advection age scaling and the metric system conversion. ISSR height had minimal impact on performance so it was chosen as a constant 500 m.

As expected, when the algorithm was tested against the whole trajectories dataset, the baseline-derived scores impacted the performance negatively due to it being designed to work with 10-minute segments at most. Therefore, in this case, the corresponding weight is set to 0. In the case of working with segmented trajectories, this weight proved useful and is equal to the advection error penalty ($\lambda_1 = \lambda_3 = 0.2$), while the altitude penalty weight is set to 0.6.

2.5. Evaluation

In this section, the choice for evaluation datasets is explained. Later, it is described how the performance metrics are computed and what I believe to be important when analysing the performance of these algorithms.

2.5.1. Evaluation Datasets

To evaluate and compare the performance of the attribution algorithms, eight synthetic test datasets, which differ in geographical location, season, and time of day, were created. These choices were guided by two considerations. Firstly, I restricted the spatial domain to two smaller regions within the larger study area defined in *Appendix A* of Sarna et al. (2025), which is also the region used by Geraedts et al. (2024). This ensures consistency and comparability with existing algorithms. Secondly, I aimed to investigate how environmental variability (both seasonal and diurnal) affects attribution performance, since contrail formation depends strongly on atmospheric conditions.

Geographic Regions

Two domains were selected within the broader region used in previous literature:

- **Region 1:** Longitude of $-100°$ to $-90°$ and latitude of $20°$ to $30°$ (Gulf of Mexico)
- **Region 2:** Longitude of $-80°$ to $-70°$ and latitude of $30°$ to $40°$ (East Coast of the United States of America)

Seasonal and Diurnal Selections

Contrail coverage varies substantially across seasons and times of day due to changes in air traffic density, temperature, and the distribution of ice-supersaturated regions. As shown in Figure 2.11, observed contrail coverage typically peaks during the morning hours (around 9 AM local time) and remains moderately high during the late afternoon and early evening.

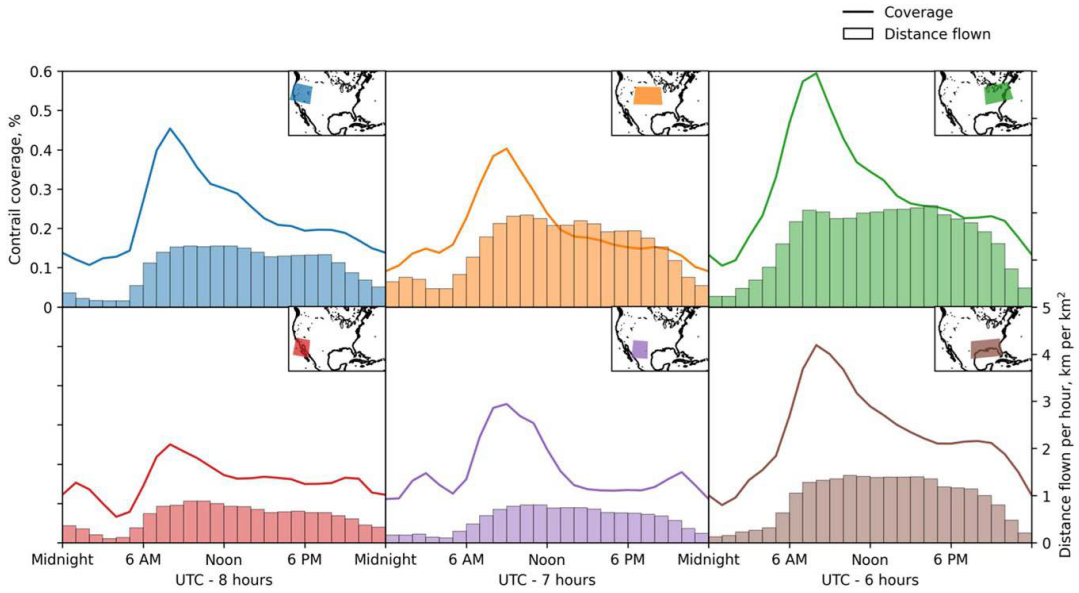


Figure 2.11: Diurnal variation in contrail coverage across several regions in the United States of America (Meijer, Kulik, et al., 2022)

To account for seasonal variation, each region was sampled during:

- **Winter:** 21-22 January 2025
- **Summer:** 11-12 July 2025

For each season, two time-of-day intervals were chosen (one at noon and one in the evening) when contrails are typically present in satellite imagery. These periods capture distinct atmospheric states and diurnal air traffic patterns while ensuring that the synthetic contrail-generation process produces a sufficient and diverse number of contrails for evaluation. The resulting intervals are as follows:

- **Winter, Noon:** 15–18 UTC (10 AM to 1 PM local US time)
  Frame time: 18 UTC (1 PM local US time)
- **Winter, Evening:** 23–02 UTC (6 PM to 9 PM local US time)
  Frame time: 02 UTC (9 PM local US time)
- **Summer, Noon:** 14–17 UTC (10 AM to 1 PM local US time)
  Frame time: 17 UTC (1 PM local US time)
- **Summer, Evening:** 22–01 UTC (6 PM to 9 PM local US time)
  Frame time: 01 UTC (9 PM local US time)

These periods fall within medium-to-high contrail coverage windows for the selected regions and provide a balanced representation of atmospheric variability relevant to flight attribution.

2.5.2. Evaluation Metrics

The attribution algorithms are evaluated against the synthetic datasets described previously using metrics adapted from Sarna et al. (2025). Evaluation is carried out both on a per-contrail basis, where the unit of analysis is the observed contrail, and on a per-flight basis, where the unit of analysis is the candidate flight. This distinction is important because contrail attribution seeks to be consistent with both observed contrails and the originating flights that produced them.

Attribution Outcomes

For each candidate–contrail pair, outcomes are categorized into the following cases:

A: Flight X formed contrail Y (reality) and flight X was attributed to contrail Y (result)

B: Flight X did not form contrail Y (reality) but flight X was attributed to contrail Y (result)

C: Flight X formed contrail Y (reality) but flight X was not attributed to contrail Y (result)

D: Flight X formed a contrail (reality) and flight X was attributed to a contrail (result)

E: Flight X did not form any contrail (reality) but flight X was attributed to a contrail (result)

F: Flight X formed a contrail (reality) but flight X was not attributed to any contrail (result)

For an easier understanding of each outcome, Sarna et al. (2025) provides the following table:

		flight x formed linear contrail y	
		Yes	No
flight x attributed to linear contrail y	Yes	**A**	**B**
	No	**C**	

		flight x formed any linear contrail	
		Yes	No
flight x attributed to any linear contrail	Yes	**D**	**E**
	No	**F**	

Figure 2.12: Contingency table for performance metrics computation (Sarna et al., 2025)

Per-Contrail Metrics

Per-contrail metrics quantify the ability of the algorithm to attribute individual contrails to their correct originating flights:

$$\text{Recall}_{\text{contrail}} = \frac{A}{A+C} \times 100 \quad [\%] \tag{2.8}$$

$$\text{Precision}_{\text{contrail}} = \frac{A}{A+B} \times 100 \quad [\%] \tag{2.9}$$

Per-Flight Metrics

Per-flight metrics measure the algorithm's ability to detect whether flights produced contrails:

$$\text{Recall}_{\text{flight}} = \frac{D}{D+F} \times 100 \quad [\%] \tag{2.10}$$

$$\text{Precision}_{\text{flight}} = \frac{D}{D+E} \times 100 \quad [\%] \tag{2.11}$$

Both types of metrics are important for a contrail-to-flight attribution algorithm. For example, the algorithm could achieve relatively high values in per-contrail metrics (50%) but perform poorly on per-flight metrics. In such a case, it would mean that the algorithm over-predicts the number of contrail-forming flights, proving inefficient for the goal of contrail mitigation.

Computation Time

In addition to attribution accuracy, computational efficiency is a key evaluation criterion. For each algorithm, we report the total runtime per dataset study. Computation time does not include generating the synthetic datasets (i.e., advecting the initial trajectories into contrails and candidates). What is more, this metric will be calculated differently for each type of dataset. When working with whole trajectories, the computation time of the baseline algorithm will be compared with the raw computation time of the physically consistent algorithm. However, in the case of segmented trajectories, the latter's computation time will be calculated as the sum of the base time and the computation of the baseline algorithm because it uses the scores generated by it.

All computation time measurements were obtained on a laptop equipped with an Intel(R) Core(TM) i7-1260P 2.10 GHz CPU, with 32 GB RAM, running Windows 11 Pro.

2.6. Toy Problem

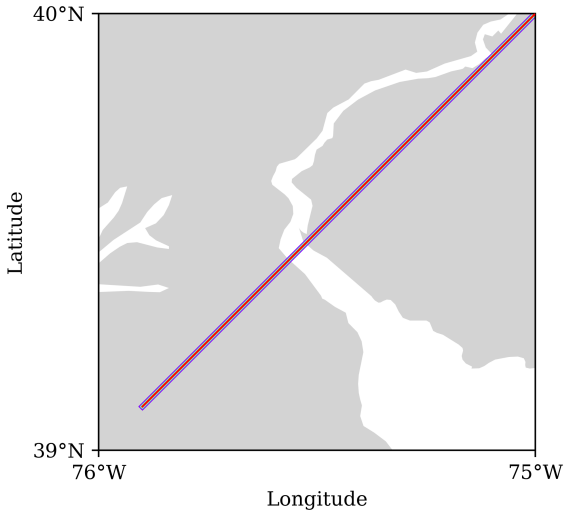
An example run of the algorithm was performed to illustrate the limitations of purely geometric attribution and the effect of the physical consistency constraints. A simple toy example was constructed with three flights that fly the same trajectory at different times in a small region near the East coast of the United States of America.

Setup

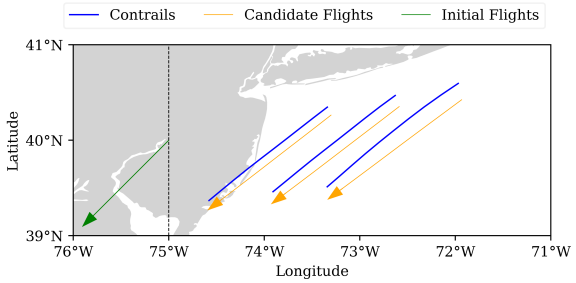
The scene is observed at a single frame time on 15 April 2025 at 10 UTC. Three flights (10 waypoints, 10 minutes of flight time) follow identical ground tracks but occur at different times, approximately 20 minutes apart:

- **Flight 1:** 08:31 to 08:40 UTC
- **Flight 2:** 08:50 to 08:59 UTC
- **Flight 3:** 09:11 to 09:20 UTC

The flights are synthetically created and advected into contrails and candidates using the methodology described in Section 2.2. The only difference is that the whole flights are advected into contrails, disregarding the ISSR/SAC filter for simplicity. The flight and the synthetic dataset can be observed in Figure 2.13.



(a) Initial flights



(b) Synthetic dataset of 3 contrails and 3 advected trajectories created from the same initial trajectories

Figure 2.13: Clustering results for two datasets

Figure 2.13b shows how the difference in time between the initial trajectories results into 3 different contrails on the same frame time because of longer advection times.

Performance

For each contrail–candidate pair, the attribution algorithm also implies an advection error (i.e., the displacement that would have to be explained by wind errors in order for the candidate to have produced the observed contrail). Figure 2.14 compares these implied vectors for the baseline algorithm and for the physically consistent algorithm.

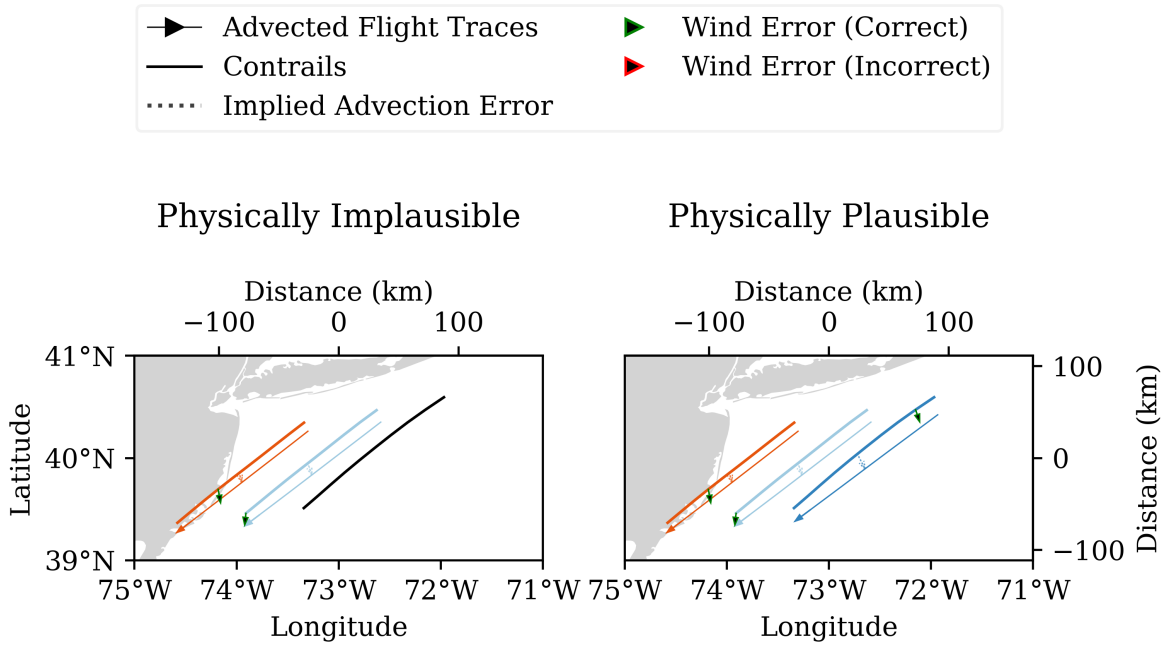


Figure 2.14: Physical consistency comparison between the baseline algorithm and the physically consistent algorithm

In Figure 2.14, the contrails are linked by colour to the resulting attribution, and the centres of the trajectories are linked with dotted vectors of the same colour. The black arrows represent the implied wind error between the two trajectories. In the *physically implausible* case, the baseline algorithm matches two contrails to the correct candidates, while the other contrail is not attributed to any flight. In contrast, in the *physically plausible* case, the algorithm produces a set of advection error vectors that are consistent in both magnitude and direction.

Wind Error

The inconsistency would arise because an incorrect match would require a wind error that is incompatible with other implied wind errors in the same region. Figure 2.15 shows the wind-speed-difference field at the frame time and local wind-speed-difference vectors between the two meteorological datasets used for contrail and candidate advection. The advection error vector associated with the correct match aligns with the local wind-speed-difference pattern. Figure 2.16 further illustrates the calibrated ensemble member wind field used for the contrails and the ensemble mean wind field used for the candidates: the spatial gradients between these fields are consistent with the wind fields shown in Figure 2.15 as expected.

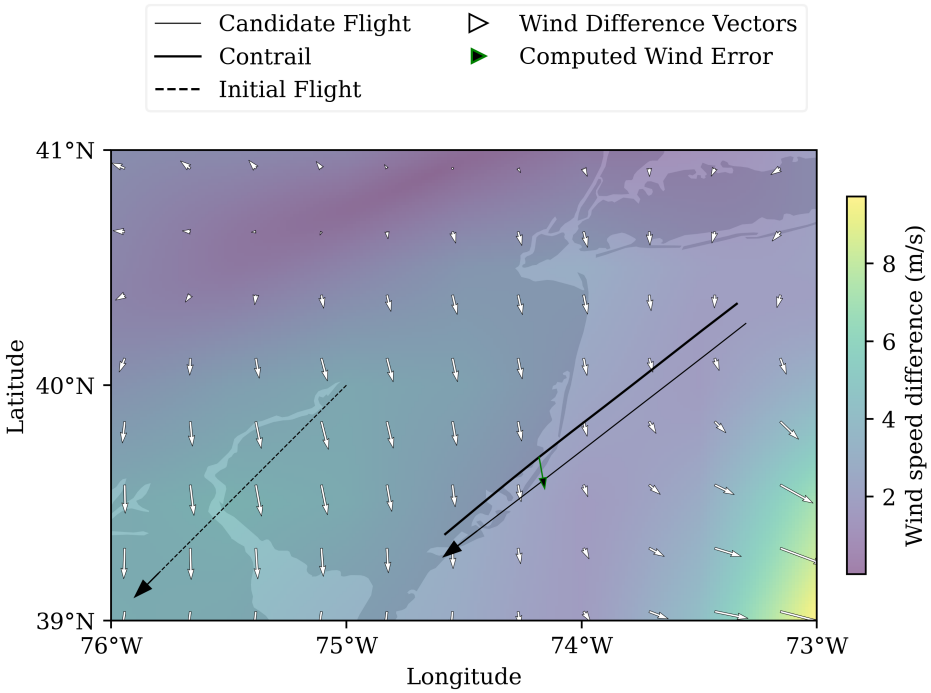


Figure 2.15: Wind-speed-difference field

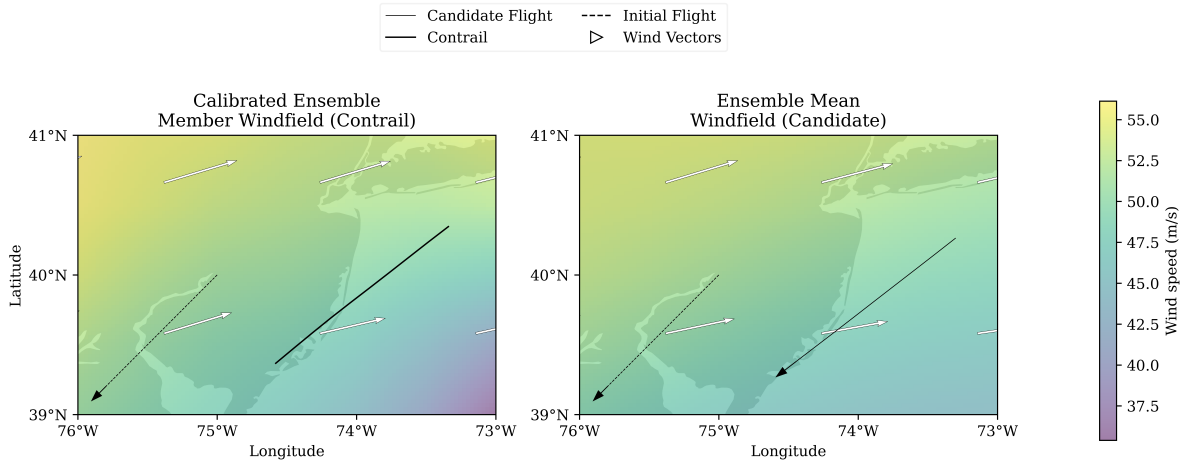


Figure 2.16: Wind fields of the calibrated ensemble member used for contrail advection (left) and wind field of the ensemble mean used for candidate flight advection (right)

Chapter 3

Results

This chapter presents a comparative evaluation of the baseline geometric attribution algorithm and the physically consistent attribution algorithms across all synthetic datasets. Section 3.1 introduces an ablation study that isolates the impact of individual components, such as physical consistency constraints, advection errors, altitude information, and geometric scoring. Section 3.2 provides a global performance analysis, comparing precision and recall across regions, seasons, times of day, and contrail properties for both whole-trajectory and segmented synthetic datasets. Finally, Section 3.3 examines specific scenarios to show how trajectory length, segmentation choices, geometric scoring, and clustering rules influence attribution outcomes and may lead to incorrect or inconsistent matches.

3.1. Ablation Study

To understand how each term of the objective function (Equation 2.5) contributes to the performance of the physically consistent attribution algorithm, an ablation study was conducted on the segmented trajectories testing dataset (Subsection 2.2.2). Segmented trajectories represent the scenario most comparable to the baseline method, as that algorithm also operates on fixed-length flight segments. Therefore, this dataset type is the appropriate type for evaluating configurations that include a non-zero baseline geometric score weight.

3.1.1. Setup

Four weighting configurations were prepared for this study:

- **Normal:** $\lambda_1 = 0.2$, $\lambda_2 = 0.6$, and $\lambda_3 = 0.2$; this configuration is representative of the algorithm described in Subsection 2.4.5
- **Advection error penalty only:** $\lambda_1 = 1.0$ and $\lambda_2 = \lambda_3 = 0.0$
- **Advection error and altitude penalty:** $\lambda_1 = 0.3$, $\lambda_2 = 0.7$, and $\lambda_3 = 0.0$;
- **Advection error penalty and geometric scores:** $\lambda_1 = 0.5$, $\lambda_2 = 0.0$, and $\lambda_3 = 0.5$;

This design of the ablation study isolates the influence of each individual penalty term and identifies whether the geometric score from the baseline algorithm is beneficial when used together with physical consistency constraints.

3.1.2. Performance

Across all configurations, the physically consistent algorithm outperforms the purely geometric baseline algorithm. However, the degree of improvement varies substantially depending on the combination of weights used in the objective function.

Table 3.1: Results of the ablation study

	Per-contrail Recall [%]	**Per-contrail Precision [%]**	**Per-flight Recall [%]**	**Per-flight Precision [%]**
Baseline	20	30	33.3	55.6
In-house (Base)	33.3	33.3	40	40
In-house ($\lambda_1 = 1.0$ and $\lambda_2 = \lambda_3 = 0.0$)	26.7	26.7	33.3	33.3
In-house ($\lambda_1 = 0.3$, $\lambda_2 = 0.7$, and $\lambda_3 = 0.0$)	33.3	33.3	40	40
In-house ($\lambda_1 = 0.5$, $\lambda_2 = 0.0$, and $\lambda_3 = 0.5$)	33.3	33.3	40	40

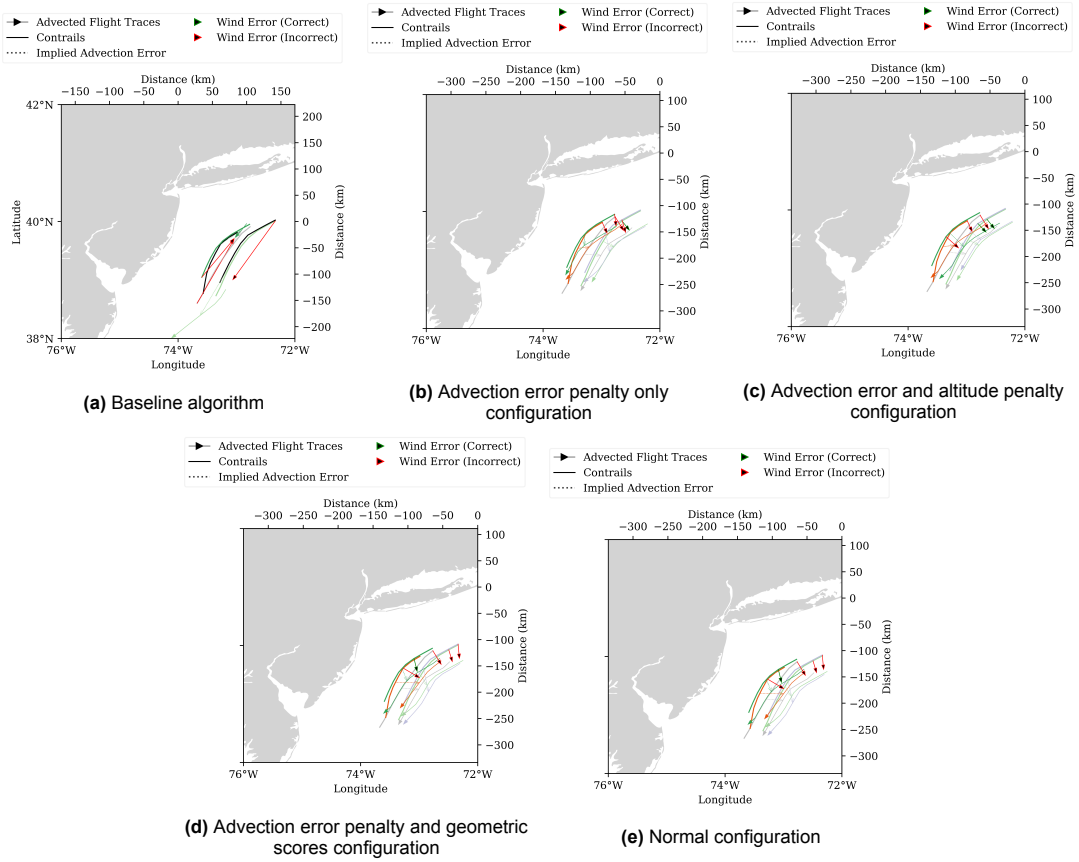
The advection-only setup already shows a clear improvement over the baseline, increasing the number of correct attributions from 3 to 4 out of 15. This confirms that enforcing physical consistency in the implied advection error rather than relying solely on geometric matching provides meaningful physical information that helps filter out some implausible matches. However, despite outperforming the baseline algorithm, the advection-only configuration consistently underperforms the other configurations.

Both two-penalty-term algorithms achieve the highest performance among the previously tested configurations, each reaching 5 correct attributions out of 15. The fact that both variants achieve the same peak performance indicates that physical consistency benefits from an additional constraint (either altitude differences or geometric matching can act as that secondary term). In other words, a single physical term might be insufficient, but combining it with one additional dimension results in a more robust attribution.

The normal configuration also achieves 5 correct matches, placing it on par with the best-performing setups in terms of performance. However, the metric differences between the normal configuration and the two earlier-mentioned, high-performing configurations are negligible. This suggests that while the normal configuration does not clearly surpass the best variants in terms of performance metrics, it may still perform better in specific ambiguous scenarios. Therefore, a more detailed inspection of individual cases is necessary to determine whether the normal configuration provides advantages that are not fully reflected in the performance metrics.

3.1.3. Detailed Analysis

To better understand how each ablation study setting affects the physical plausibility of the resulting contrail–flight assignments, two representative clusters from the dataset are examined. The first cluster contains several contrails, which makes it a suitable example for evaluating how well each configuration resolves ambiguity under realistic geometric and physical constraints. The second cluster contains only one contrail, allowing for a more focused inspection of how the different penalty setups behave when the geometric structure is simple.



(a) Baseline algorithm

(b) Advection error penalty only configuration

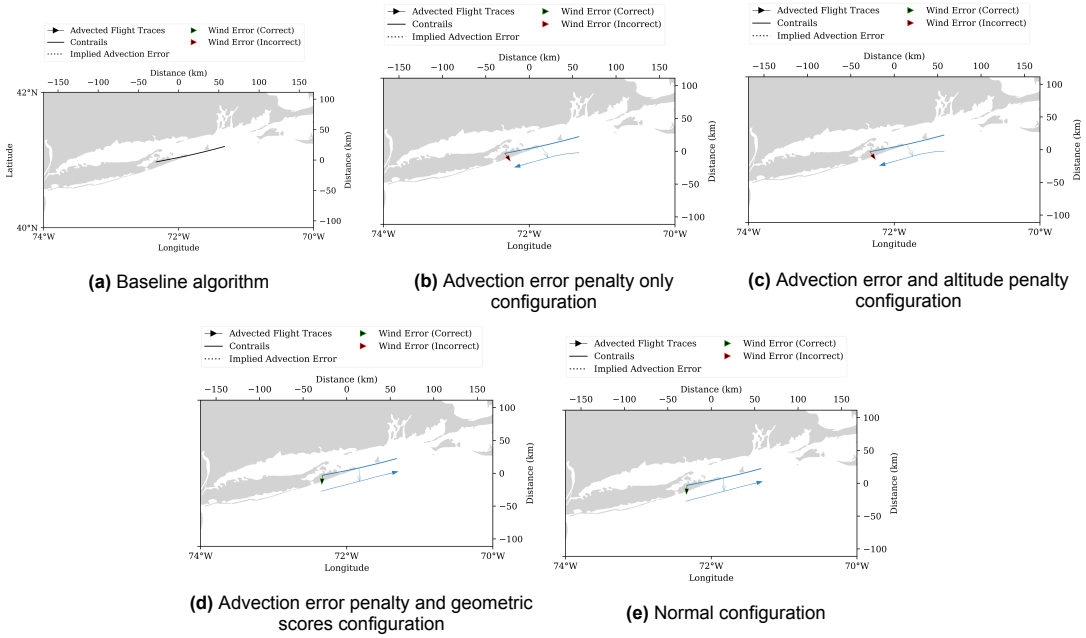
(c) Advection error and altitude penalty configuration

(d) Advection error penalty and geometric scores configuration

(e) Normal configuration

Figure 3.1: One cluster of contrails attributed to their respective candidates, together with the implied advection errors

Figure 3.1 shows one of the contrail clusters created in the ablation study. In this case, the baseline algorithm (Figure 3.1a) produces physically inconsistent pairings, indicated by red implied advection error vectors that clearly diverge from the true wind-error direction. The advection error penalty configuration (Figure 3.1b) improves upon the baseline by penalizing one inconsistent displacement vector, resulting in one correct assignment but more coherent assignments overall (in terms of wind error consistency). The advection and altitude penalties configuration (Figure 3.1c) performs better in this scenario: by integrating altitude similarity with the advection penalty, it captures two correct matches within the cluster, reducing local mismatches while maintaining physically realistic error vectors. Similarly, the advection error penalty and geometric scores configuration (Figure 3.1d) also suppresses implausible assignments by combining geometric matching with physical constraints. Finally, the normal configuration (Figure 3.1e), which combines all three penalties, yields physically coherent implied advection-error vectors. Although consistent in terms of performance metrics, it does not necessarily outperform the two-term configurations in this particular cluster. Overall, this example illustrates how adding physically meaningful penalties increases the robustness of the matching process in ambiguous multi-contrail settings.



(a) Baseline algorithm

(b) Advection error penalty only configuration

(c) Advection error and altitude penalty configuration

(d) Advection error penalty and geometric scores configuration

(e) Normal configuration

Figure 3.2: Another cluster of contrails attributed to their respective candidates, together with the implied advection errors

Figure 3.2 shows a second, simpler example consisting of a cluster with a single contrail. The baseline algorithm (Figure 3.2a) does not attribute the contrail to any candidate. The advection error penalty and the advection error penalty and altitude penalty configurations (Figure 3.2b and Figure 3.2c) improve the result by attributing the contrail to a candidate but still fail to identify the correct candidate. In contrast, both the advection error penalty and geometric scores configuration (Figure 3.2d) and the normal configuration (Figure 3.2e) successfully recover the correct match. It is important to note, however, that this is a single-contrail cluster (inconclusive in the prospect of verifying physical consistency through implied advection errors). Nonetheless, this example shows that integrating geometric matching from the baseline algorithm's scores with physically consistent advection penalties can be particularly effective in scenarios where advection errors would not give much insight into the problem.

To conclude the ablation study, the advection error penalty configuration consistently outperforms the baseline, confirming that enforcing wind-error consistency is effective in improving the results of the baseline algorithm. However, its performance remains below that of the mixed-penalty configurations, indicating that advection information alone is not sufficient in cases where geometric ambiguity or altitude differences play an important role. Although the normal configuration does not consistently achieve higher metric values across all clusters (compared to the two-penalty configurations), the two figures show that it delivers stable and physically plausible matches across diverse scenarios.

3.2. Overall Performance

To evaluate the performance of the proposed attribution algorithms, all experiments were conducted using the complete collection of synthetic datasets described in Subsection 2.5.1. Across all datasets, the simulation contains a total of 3 775 flights, from which 304 contrails were generated. For each contrail, the algorithms compare its shape against candidate flights. In the whole-trajectory dataset setting, there are 3 775 candidate flights, whereas, in the segmented-trajectory dataset setting (where each flight is divided into multiple smaller segments), there are 13 365 candidate flight segments. These datasets provide the basis for comparing the performance and computational efficiency of the baseline and physically consistent algorithms.

To systematically assess how each algorithm behaves across different operational conditions, this section analyses performance along several dimensions. First, computation time is compared in relation to dataset size because the contrail-set size differs substantially between whole and segmented trajectories, and it directly impacts the efficiency of the algorithm. Next, overall attribution performance will be

compared between the algorithms using the metrics introduced in Subsection 2.5.2. To understand how environmental variability influences performance, the results are further broken down by geographical region, time of day, and season. Finally, following the evaluation framework proposed by Sarna et al. (2025), the performance is analysed as a function of contrail age and contrail altitude. Together, these analyses provide a comprehensive overview of the strengths, limitations, and behavioural differences between the baseline and physically consistent attribution algorithms.

Figure 3.3 shows how computation time scales with the number of contrails across the four algorithm–dataset combinations. The markers show the computation times for each run, while the dotted curves are exponential fits to the data-points, included to visualise the scaling trend with the number of contrails.

Several trends are immediately visible. Firstly, the physically consistent algorithm operating on whole trajectories exhibits the lowest overall computation time, staying below 5 minutes even for the largest dataset of 107 contrails. This behaviour is expected, as it evaluates a relatively small set of candidates and is not using the match scores from the baseline algorithm. When switching to segmented trajectories, the runtime of the physically consistent algorithm increases substantially due to the increase in candidate count. The curve grows with dataset size, reaching nearly one hour for the largest scenario. This confirms that segmentation leads to a significant computational time increase.

The baseline algorithm operating on whole trajectories shows a similar trend, but with an increasing rate. This is largely attributable to the structure of the algorithm, which compares each contrail to all valid candidates before deciding on an attribution. The second-most computationally intensive configuration is the baseline algorithm applied to segmented trajectories. The curve grows rapidly and almost reaches one hour for the largest dataset.

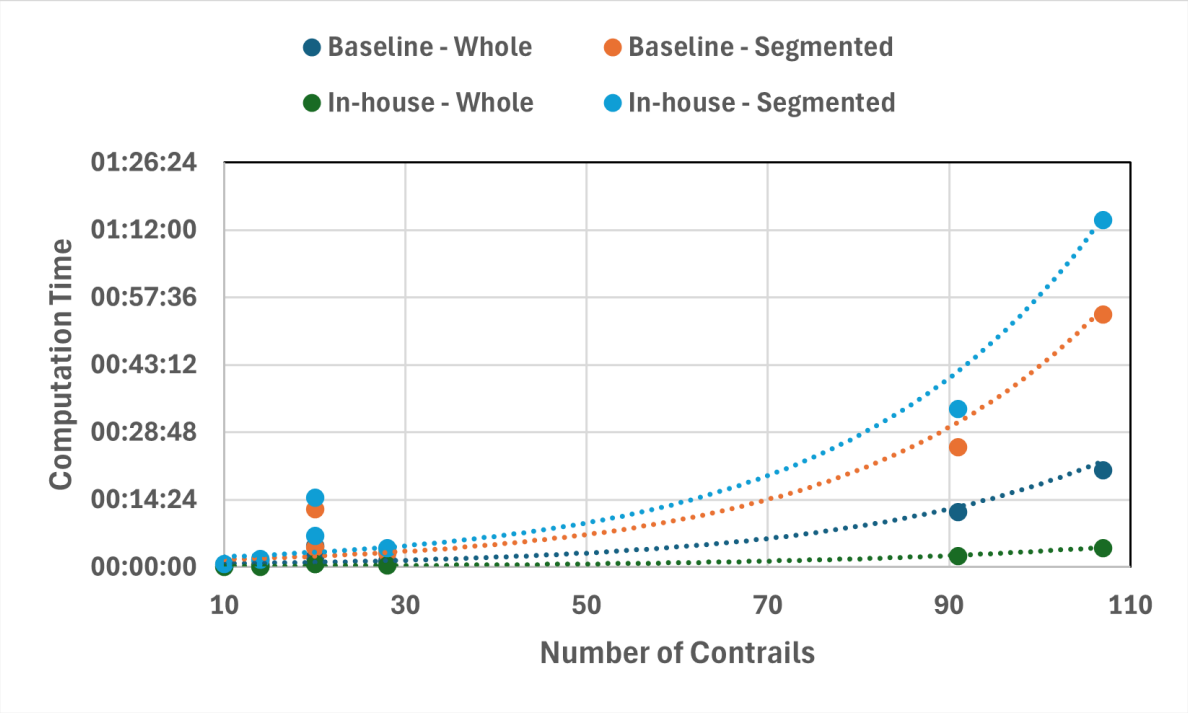


Figure 3.3: Computation times across datasets varying in size (by number of contrails)

Figure 3.4 summarizes the overall attribution performance across all datasets for both algorithms and both types of datasets (whole and segmented trajectories). Several patterns clearly emerge. Firstly, the physically consistent algorithm applied to whole trajectories achieves the strongest performance across all four metrics. Per-contrail recall and precision both reach 36.5%, which represents a substantial improvement over the baseline whole-trajectory method (7.6% recall and 13.9% precision). A similar gain is observed in per-flight metrics, where the physically consistent whole-trajectory configuration

obtains 54.9% recall and precision.

Using segmented trajectories improves the baseline algorithm considerably, especially for per-flight recall, which increases from 19.1% (whole trajectories) to 53.3% (segmented). However, segmentation also introduces more candidates, which leads to lower per-contrail precision (8.9%) and per-flight precision (41.3%) compared with the physically consistent whole-trajectory configuration.

When applied to segmented trajectories, the physically consistent algorithm performs poorly in terms of recall but drops in precision relative to the whole-trajectory case. This is expected, as the number of candidate segments increases almost fourfold, leading to more confusion. Nevertheless, the physically consistent segmented configuration still maintains competitive per-flight recall (23.7%) and substantially improves over the baseline whole-trajectory setup.

Overall, these results demonstrate that the physically consistent whole-trajectory algorithm is the best configuration in terms of accuracy, while segmentation benefits the recall of the baseline method but comes at the cost of reduced precision. The interaction between physical consistency and trajectory segmentation will be further explored in the region-specific and time-of-day analyses that follow.

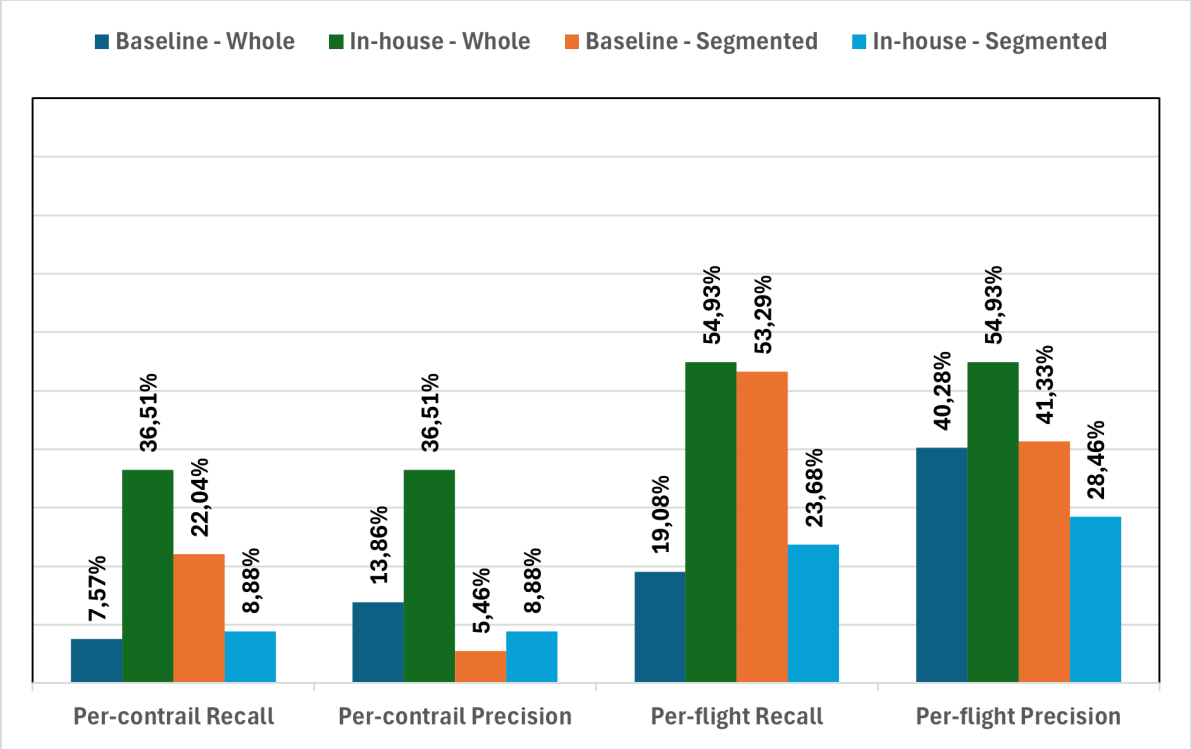


Figure 3.4: Overall performance metrics across all datasets

For context, Sarna et al. (2025) report global metrics for the single-frame baseline of Geraedts et al. (2024), the tracking algorithm of Chevallier et al. (2023), and their own method, Contrail Attribution Sample Consensus (CoAtSaC), on a much larger benchmark dataset. The physically consistent whole-trajectory algorithm reaches contrail precision/recall of 36.5% and flight precision/recall of 54.9%. In terms of contrail recall, this is higher than both the single-frame (33%) and tracking (28.6%) baselines and matches CoAtSaC (36.6%), although the in-house algorithm's contrail precision remains below all three methods, especially CoAtSaC (66.9%). For the per-flight metrics, the in-house method improves on the single-frame and tracking algorithms' precision (41.4% and 50.3%, respectively), but falls short of CaAtSaC (68.4%). In terms of recall, the single-frame baseline' performance (68.4%) overshadows the in-house algorithms' per-flight recall of 54.9%. Because these benchmark results are obtained on a substantially larger dataset than the synthetic sets used in this study, the comparison should be interpreted qualitatively rather than as a strict ranking. Nonetheless, it indicates that the physically consistent whole-trajectory algorithm is competitive with current state-of-the-art contrail–flight attribution

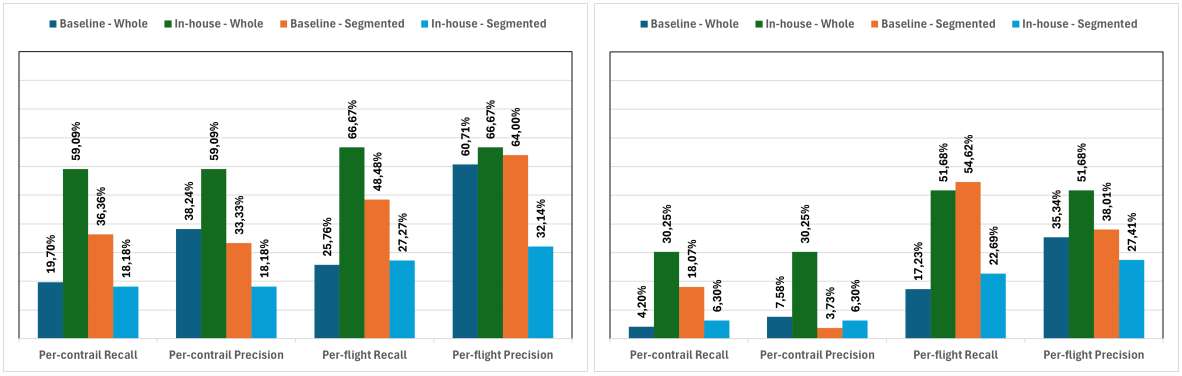
methods.

Figure 3.5 shows the attribution performance separately for the two regions studied. Several consistent patterns emerge. In both regions, the physically consistent algorithm applied to whole trajectories remains the best overall configuration, achieving the highest precision and recall across most metrics. This mirrors the global trend observed in the full-dataset evaluation and indicates that incorporating physical constraints improves results independently of geographical context.

The baseline algorithm benefits substantially from trajectory segmentation in both regions, particularly in terms of per-flight recall. In the first region, segmentation brings the baseline close to the physically consistent whole-trajectory configuration in recall, although at a cost of lower precision. The second region shows an even stronger recall boost for the segmented baseline, suggesting that its geometric scoring aligns well with the contrail structures present in this area. However, precision remains consistently lower than that of the physically consistent method.

The physically consistent algorithm applied to segmented trajectories exhibits reduced performance in both regions. While it retains approximately similar recall in both regions, its precision decreases notably, confirming that segmentation has a negative impact on the physically consistent approach.

Overall, these regional comparisons demonstrate that the physically consistent whole-trajectory method is geographically robust, while segmentation primarily benefits recall for the baseline model.



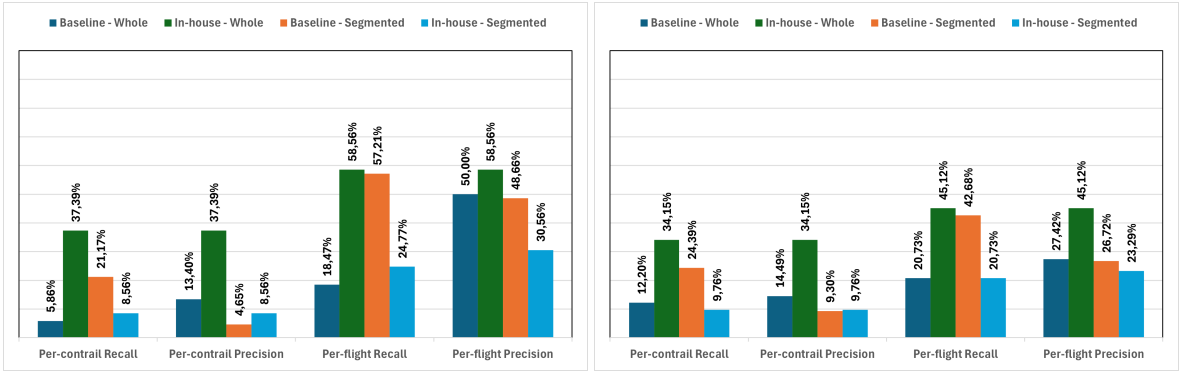
(a) First region (Gulf of Mexico)

(b) Second region (East Coast of the United States of America)

Figure 3.5: Performance metrics across all datasets per region studied

Figure 3.6 shows the attribution performance broken down by season, allowing us to examine whether the algorithms behave differently. Several patterns can be seen. In winter, the physically consistent algorithm applied to whole trajectories achieves the strongest performance across all metrics, with per-contrail recall and precision both reaching 37.39% and per-flight recall and precision around 59%. The baseline algorithm also benefits from segmentation in winter, particularly in per-flight recall, although precision drops, indicating that segmentation expands the search space but introduces more ambiguity due to the larger number of candidates.

Summer results differ in structure. The in-house whole-trajectory approach remains the strongest, but the gap between whole and segmented datasets decreases. The baseline segmented algorithm achieves recall values comparable to those of the in-house whole-trajectory method. However, precision remains higher for the physically consistent whole-trajectory method, indicating that physical constraints play a stabilizing role independently of seasonal variance.

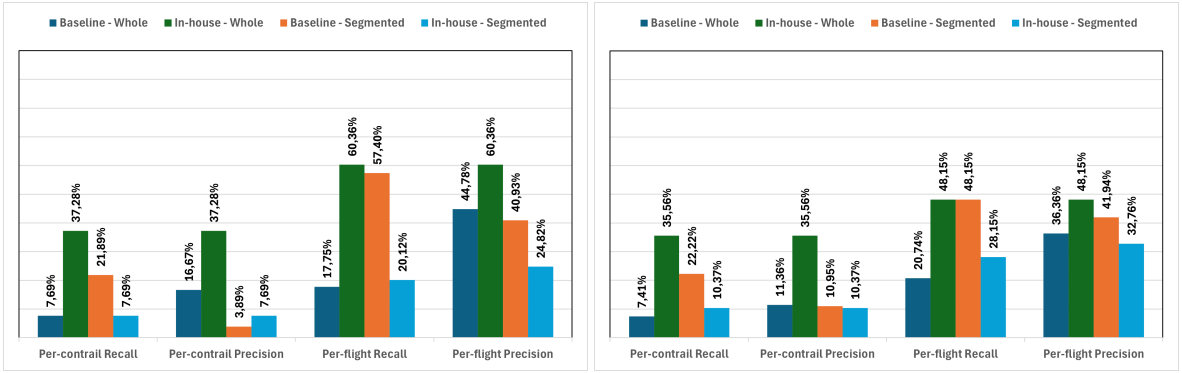


(a) Winter **(b)** Summer

Figure 3.6: Performance metrics across all datasets per season studied

Figure 3.7 shows how performance varies across different periods of the day. The noon datasets (Figure 3.7a) show the same overall pattern observed in previous analyses: the physically consistent whole-trajectory configuration again outperforms all other setups across recall and precision. The baseline method benefits from segmentation, boosting per-flight recall to levels comparable with the in-house whole-trajectory configuration, but this gain is accompanied by a noticeable drop in precision. The physically consistent segmented configuration performs weakest in this time window.

In contrast, the evening datasets (Figure 3.7b) exhibit a more balanced performance distribution across algorithms. Attribution quality is generally lower in the evening across all configurations. Nevertheless, the physically consistent whole-trajectory method continues to deliver the highest precision and competitive recall. The baseline segmented configuration maintains strong recall but again shows degraded precision, reinforcing the trend that segmentation amplifies the recall–precision trade-off. The in-house segmented algorithm improves slightly relative to its noon performance, but it remains significantly behind the whole-trajectory variant. Overall, these results highlight the robustness of the physically consistent method across diurnal variance.



(a) Noon **(b)** Evening

Figure 3.7: Performance metrics across all datasets per period of the day

Sarna et al. (2025) analyse how their algorithm and the baseline algorithm perform with varying contrail attributes such as age and altitude. In their study, contrail age seemed to be the most relevant property, showing higher precision and recall for younger contrails. In this study, we compare the eight algorithm-dataset combinations in a similar way. Figure 3.8 presents the performance for each contrail age bin as well as the number of contrails in that bin. Figure 3.9 has the same structure but groups contrails based on altitude.

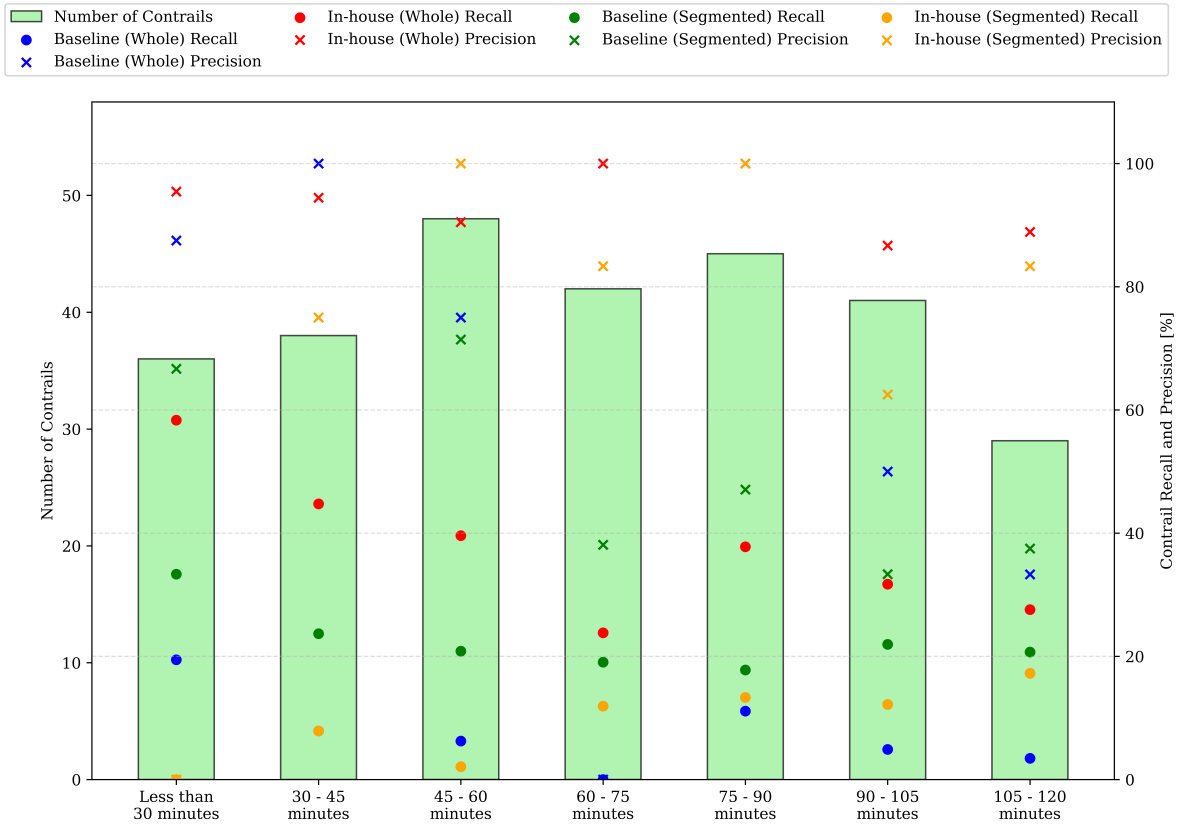


Figure 3.8: Attribution performance per contrail age group

Figure 3.8 demonstrates that attribution performance exhibits a dependency on contrail age, a trend also reported Sarna et al. (2025). Similar to their results, younger contrails are generally easier to attribute correctly to flights, while older ones show degraded recall and, to a lesser extent, precision. The number of contrails peaks between roughly 45 and 105 minutes, where similar mid-range age groups dominated the dataset. One first observation to make is that the dataset used for this study is significantly smaller than the one used in Sarna et al. (2025) and might not provide the data variability to be able to perfectly compare these results with theirs.

The physically consistent whole-trajectory method outperforms all other configurations across almost all age bins in terms of precision. Precision for this algorithm almost always reaches 90–100% for contrails younger than 90 minutes. Recall declines with age in both studies (with the exception of the in-house segmented algorithm). This shows that, even though older contrails are generally harder to attribute to flights because of their variable size and shape, the physically consistent (whole) algorithm can still recover the correct attributions due to its dependency on wind fields.

Overall, the relationship between age and attribution performance in our data is consistent with the pattern reported by Sarna et al. (2025) for the baseline algorithm. Younger contrails provide clearer attribution signals, while older ones become considerably more difficult to match.

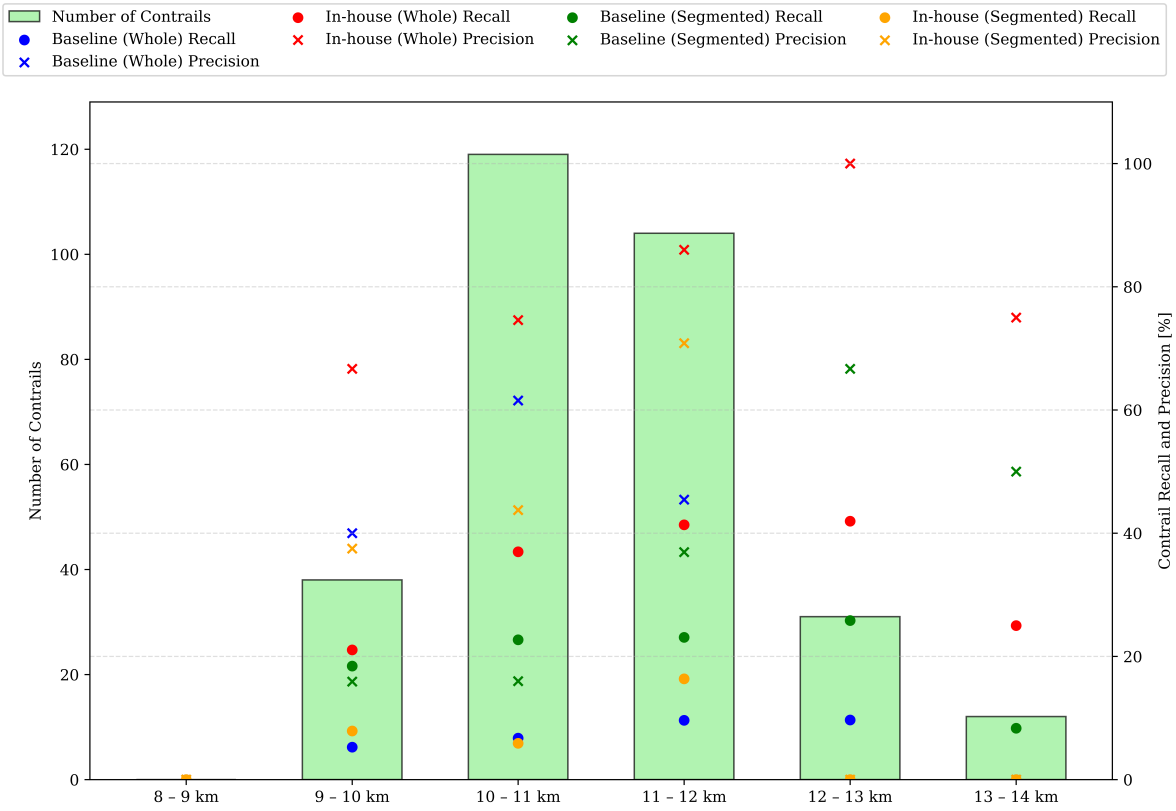


Figure 3.9: Attribution performance per contrail altitude group

Figure 3.9 compares performance across altitude bins following the same methodology as. As expected, most contrails are concentrated between 10 and 12 km of altitude, matching typical cruise altitudes and the distribution shown in Sarna et al. (2025).

The physically consistent whole-trajectory method again displays the highest precision across almost all altitude groups. Precision values are especially stable in the 10–12 km range, often exceeding 85–95%. Recall exhibits moderate variation across altitudes, with the physically consistent method performing best at mid-range altitudes. At the lowest and highest altitude bins (below 9 km and above 13 km), both studies observe small sample sizes.

Segmentation again benefits the baseline recall at all altitudes. However, the cost is substantially lower in precision because of the increased number of potential matches. Finally, the physically consistent algorithm applied to segmented data performs poorly.

Overall, in all evaluated subsets, the physically consistent algorithm on whole trajectories maintains high precision and, in most cases, higher recall than the baseline on whole trajectories.

At the same time, the baseline algorithm shows a different behaviour from that reported by Sarna et al. (2025). In their large-scale evaluation, the baseline algorithm achieved significantly higher recall and precision. In contrast, the baseline–segmented configuration in this study performs relatively poorly, particularly in precision. Even though segmentation boosts recall, this comes at the expense of introducing far more false positives. The discrepancy between these results and those of Sarna et al. (2025) is therefore notable and suggests the need for a more detailed and realistic analysis. A plausible explanation is that the datasets used here are substantially smaller than in Sarna et al. (2025). This could be investigated by using the algorithm implemented here on the synthetic dataset created by (Sarna et al., 2025).

A final observation concerns the large performance gap between the physically consistent whole-trajectory and segmented variants. While the whole-trajectory configuration is consistently strong, segmentation dramatically reduces precision and recall. This degradation arises from the sharp increase

in candidate count: segmentation multiplies the number of available flight segments by nearly a factor of four, greatly expanding the search space. Unlike the baseline method, which gains recall from segmentation, the physically consistent algorithm relies on global coherence of advection (both magnitude and direction). Segmenting flights disrupts this coherence and makes the inferred displacement vectors noisier.

Baseline Algorithm Performance

This subsection explains why the performance of the implementation of the algorithm presented in Geraedts et al. (2024) is slightly different than the performance reported by (Sarna et al., 2025).

Both (Geraedts et al., 2024) and (Sarna et al., 2025) work exclusively with linear contrails. In this implementation, contrails are used regardless of their shape and orientation. Since the objective function used by (Geraedts et al., 2024) is highly dependent on this representation, it is expected that some of the contrails used in this study (which are highly irregular in shape) will not be correctly attributed to their forming flight when the baseline algorithm is used. The following figures are representative of two cases: a linear contrail (Figure 3.10), two irregular contrails that were not matched by the baseline algorithm (Figure 3.11), and one case where the contrail is irregular and it was attributed to the wrong flight (Figure 3.12).

As explained earlier, the baseline algorithm finds the best contrail-to-flight match by aligning the candidate with the contrail in a rotated coordinate system based on the contrail geometry. In essence, the best match would be a candidate that aligns best with the w-axis. Figure 3.10 shows a case in which both the baseline algorithm and the physically consistent algorithm recover the same contrail-candidate pair due to the linearity of the contrail.

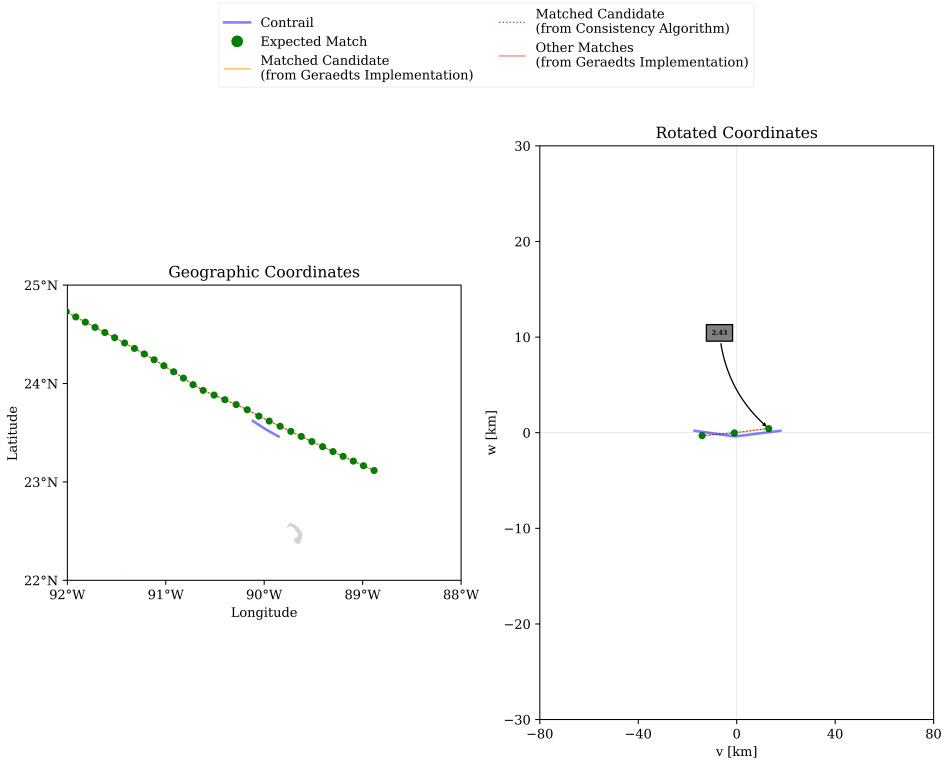
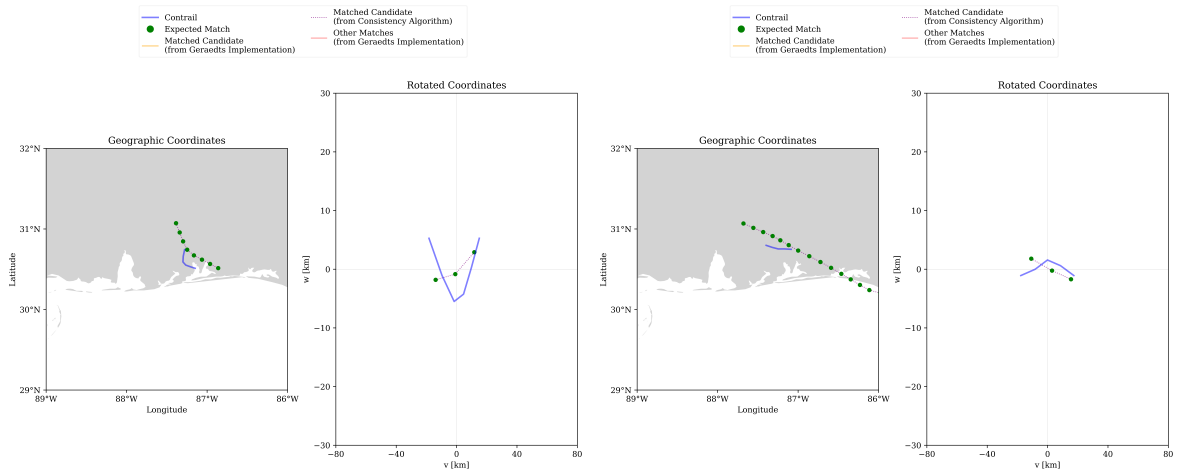


Figure 3.10: Linear contrail rotated using the baseline algorithm implementation

Figure 3.11 shows two contrails with noticeable curvature. In both examples, the baseline algorithm fails to return any attributions. The curvature prevents the construction of a meaningful straight reference axis.



(a) First example **(b)** Second example

Figure 3.11: Irregular contrails rotated using the baseline algorithm implementation (no baseline attributions)

Figure 3.12 does not recover the correct match as the second algorithm does because of the contrail's irregularities. The implementation of the baseline algorithm chooses another candidate, which is better aligned with the w-axis.

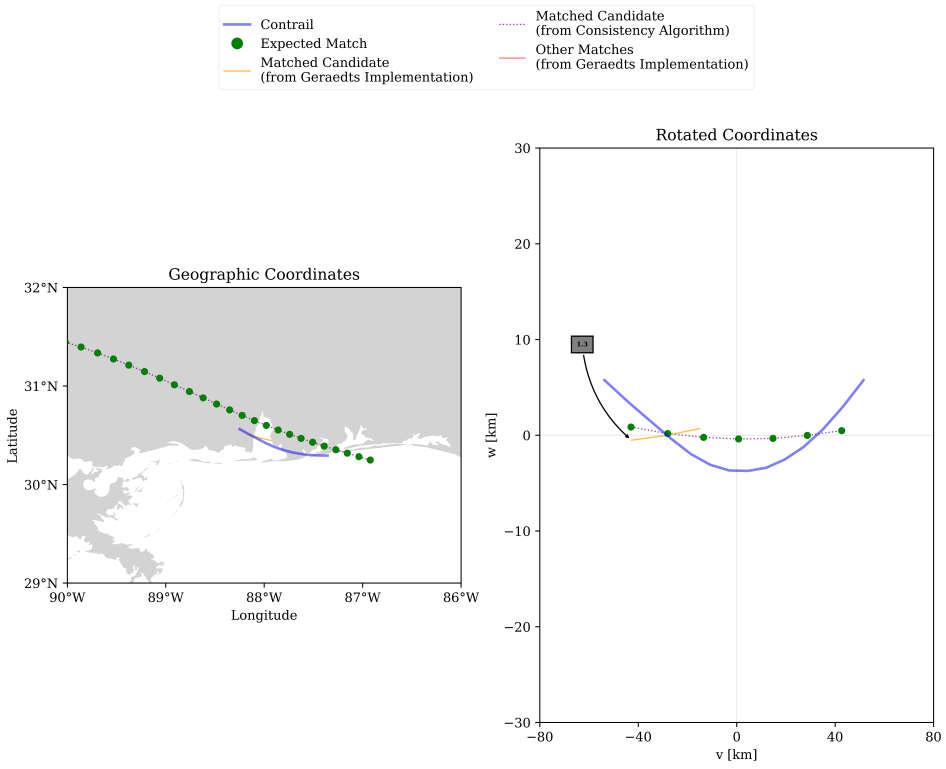


Figure 3.12: Irregular contrail rotated using the baseline algorithm implementation (attributed incorrectly)

Moreover, another difference between the implementation of the baseline algorithm and the original paper is in how the correct match is defined (as stated in Section 2.3).

A natural way to test the hypothesis that non-linearity affects the baseline algorithm would be to introduce a quantitative measure of contrail linearity, filter out strongly non-linear contrails, and rerun the algorithm on this new subset. Implementing such a dedicated filter would be useful in directly proving

this hypothesis, but the analysis of performance as a function of contrail age (Figure 3.8) already provides indirect support for this explanation. Contrails are expected to lose linearity as they age, so, as seen in the figure, the recall for the baseline method also decreases with contrail age.

3.3. Specific Scenario

To illustrate in detail how the baseline and physically consistent algorithms behave under challenging conditions, this section analyses one concrete multi-contrail scenario. The aim is to understand why the two algorithms diverge in their predictions, how segmentation affects their behaviour, and how the clustering rule may introduce inconsistencies. The scenario consists of two contrails (and later three) evolving under similar meteorological conditions and positioned close enough in space to be clustered together.

Figure 3.13 compares the baseline algorithm (Geraedts et al., 2024) with the physically consistent in-house algorithm when both operate on whole flight trajectories. In this configuration, the in-house approach correctly recovers the expected matches for both contrails, whereas the baseline algorithm fails to identify any attribution at all.

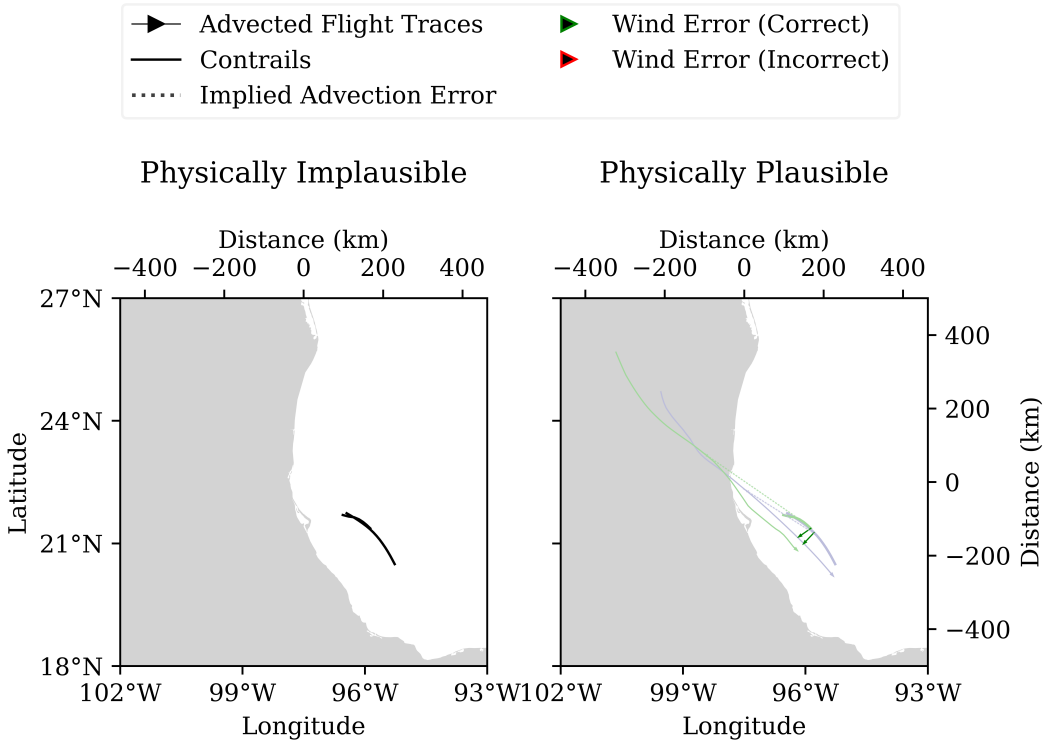


Figure 3.13: Attributions for both the baseline and the physically consistent algorithm on the whole-trajectory datasets

The main reason for this discrepancy lies in the trajectory length asymmetry. The baseline algorithm relies heavily on the geometry of the rotated coordinate system, in which the contrail defines the w-axis and matching candidates are expected to align with it. When two candidate flights differ markedly in length from the contrail—or when the contrail exhibits even mild curvature—the baseline score penalizes trajectories that do not align closely with the constructed axis. The in-house algorithm, by contrast, evaluates matches based on the induced wind error vector and altitude differences. This makes it robust to geometric discrepancies and enables it to recover the correct match despite the length difference.

When the same scenario appears on the segmented trajectories datasets, the behaviour of both algorithms changes substantially. Figure 3.14 shows that the baseline algorithm still fails to return any attribution.

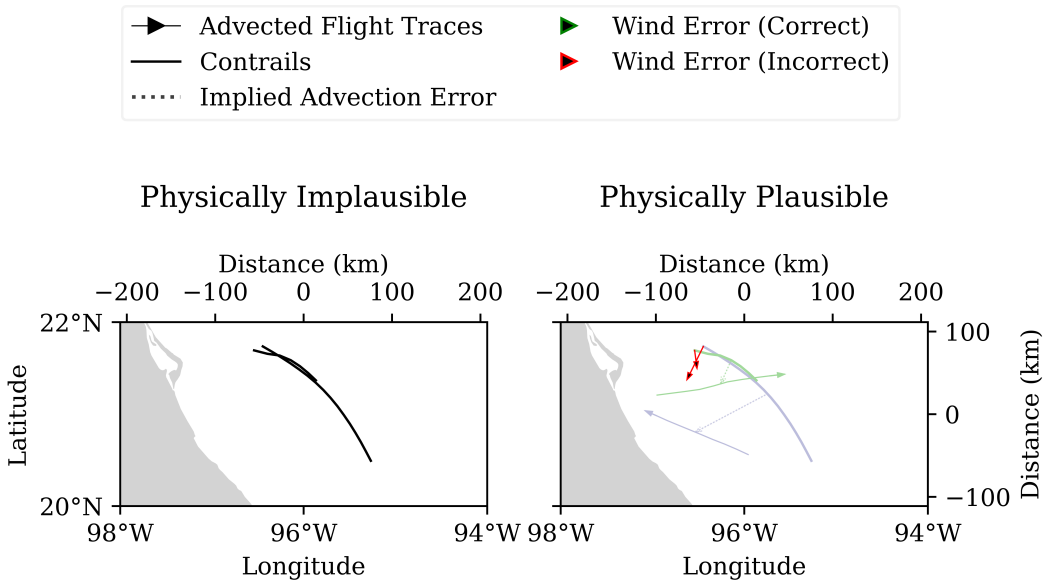


Figure 3.14: Attributions for both the baseline and the physically consistent algorithm on the segmented-trajectory datasets

This aligns with the global findings from Figure 3.2: segmentation does not solve the geometry problem for contrails exhibiting curvature or inconsistent local direction, because each segment still requires a well-defined linear reference axis for the match score to be meaningful. Even short curved segments may violate this assumption.

More surprisingly, the in-house segmented algorithm also fails to recover the correct match, despite having successfully identified the correct flights in the whole-trajectory case. In this scenario, one candidate segment (indicated as the indigo trajectory) exhibits a slightly better geometric alignment with the contrail segment, causing it to gain a better match score even though its implied advection error is less physically plausible. The remaining contrail is then forced to match the second candidate through the physical consistency rule, but this results in a wind error vector that is inconsistent with the scene and, therefore, an incorrect second attribution.

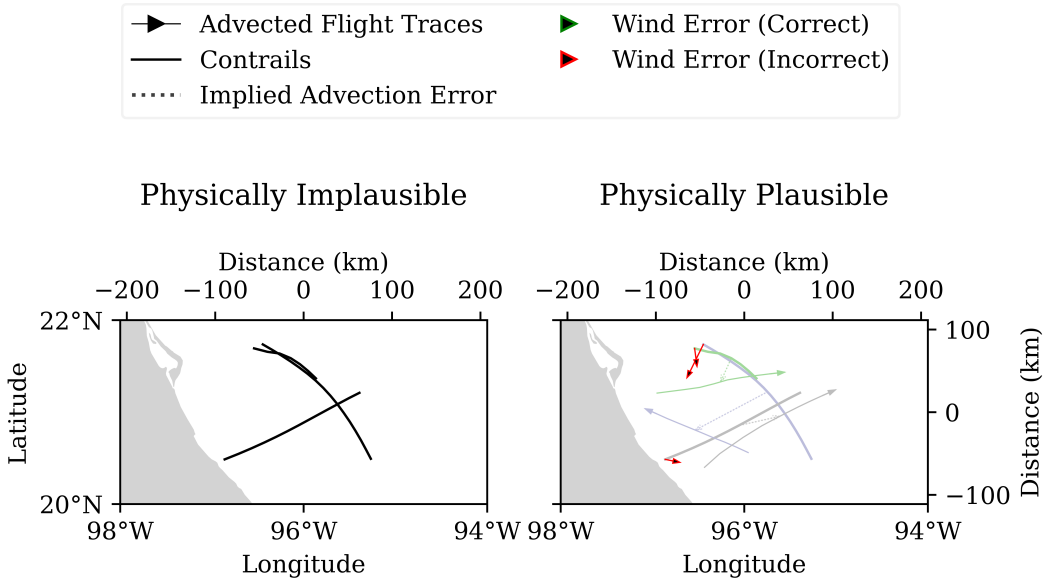


Figure 3.15: Attributions for both the baseline and the physically consistent algorithm on the segmented-trajectory datasets (different cluster)

Another complication emerges when examining how the contrails are grouped into clusters for consistency. In the whole-case (Figure 3.13) and the first segmented-case (Figure 3.14), the two contrails form a two-element cluster. However, under a slightly modified segmentation, the clustering logic assigns the same contrails to a three-contrail cluster, as illustrated in Figure 3.15. This additional contrail alters the optimization problem: the required consistency is now enforced across three wind error vectors, making it more difficult to satisfy all constraints simultaneously. As a result, the algorithm selects a different combination of matches, again driven partly by the geometric score's preference for certain candidate segments. To address the issue of inconsistent clustering over multiple runs, a new clustering rule or strategy could be implemented, one that would not change the contrails in each cluster if the conditions are the same.

3.4. Statistical Analysis of Performance Differences

The precision and recall values reported in the previous sections are computed on a finite benchmark dataset. Consequently, part of the observed differences between the baseline geometric algorithm of Geraedts et al. (2024) and the physically consistent method may simply be due to sampling variability. To assess whether the performance gaps are compatible with random fluctuations, a simple statistical significance analysis was carried out.

For each algorithm configuration, the performance metrics were summarized in terms of a confusion matrix over all the candidate flights:

- True positives (TP): a flight formed a contrail and the algorithm assigned a contrail to the flight
- False positives (FP): a flight did not form a contrail, but the algorithm assigned a contrail to the flight
- True negatives (TN): a flight did not form a contrail and the algorithm did not assign a contrail to the flight
- False negatives (FN): a flight formed a contrail, but the algorithm did not assign a contrail to the flight

From this confusion matrix, the following parameters are computed:

$$H = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad F = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \qquad S = \frac{\mathrm{TP}+\mathrm{FN}}{n} \tag{3.1}$$

where H is the hit rate (recall), F is the false-alarm rate, S is the base rate of contrail-forming flights in the sample, and n is the total number of candidate flights. Precision and recall are then calculated as such:

$$\mathrm{precision} = \frac{HS}{HS + F(1-S)}, \qquad \mathrm{recall} = H \tag{3.2}$$

For the whole-trajectory experiments, these equations yield $\mathrm{recall} = 0.19$ and $\mathrm{precision} = 0.40$ for the baseline algorithm, and $\mathrm{recall} = 0.55$ and $\mathrm{precision} = 0.55$ for the physically consistent algorithm (based on the 3 775 candidate flights). For the segmented-trajectory experiments, the corresponding values are $\mathrm{recall} = 0.53$ and $\mathrm{precision} = 0.41$ for the baseline and $\mathrm{recall} = 0.24$ and $\mathrm{precision} = 0.29$ for the physically-consistent method (based on 13 365 candidate flight segments).

The so-called null hypothesis is that two algorithms being compared (e.g., baseline and physically consistent algorithms on whole trajectories) actually have the same performance on the same dataset. Under this hypothesis, they share a common base rate (S_*), hit rate ($(H_*$), and false-alarm rate ($(F_*$). These shared parameters are estimated by pooling the confusion matrices of the two algorithms:

$$H_* = \frac{H_1 S_1 n_1 + H_2 S_2 n_2}{S_1 n_1 + S_2 n_2}, \qquad F_* = \frac{F_1(1-S_1)n_1 + F_2(1-S_2)n_2}{(1-S_1)n_1 + (1-S_2)n_2}, \qquad s_* = \frac{S_1 n_1 + S_2 n_2}{n_1 + n_2} \tag{3.3}$$

where subscripts 1 and 2 refer to the two algorithms that are being compared. Given some results of Equation 3.3, many experiments are simulated in which:

1. For each algorithm, the confusion matrix is simulated according to the shared base, hit, and false-alarm rates, which yields simulated TP, FP, FN, and TN.

2. From these, precision and recall are computed for each algorithm, and the differences are recorded.

This procedure is repeated $m = 10\,000$ times and produces empirical sampling distributions for the recall and precision differences of two algorithms under the null hypothesis. The observed differences from the real algorithm runs (see Figure 3.4) are also recorded.

Figures Figure 3.16 and Figure 3.17 summarize the results of the simulations for the whole and segmented trajectories datasets, respectively. In each panel, the blue histogram shows the sampling distribution of the difference in recall or precision under the null hypothesis that the two algorithms have identical underlying performance. The red dashed line marks the actually observed difference from the actual tests performed previously.

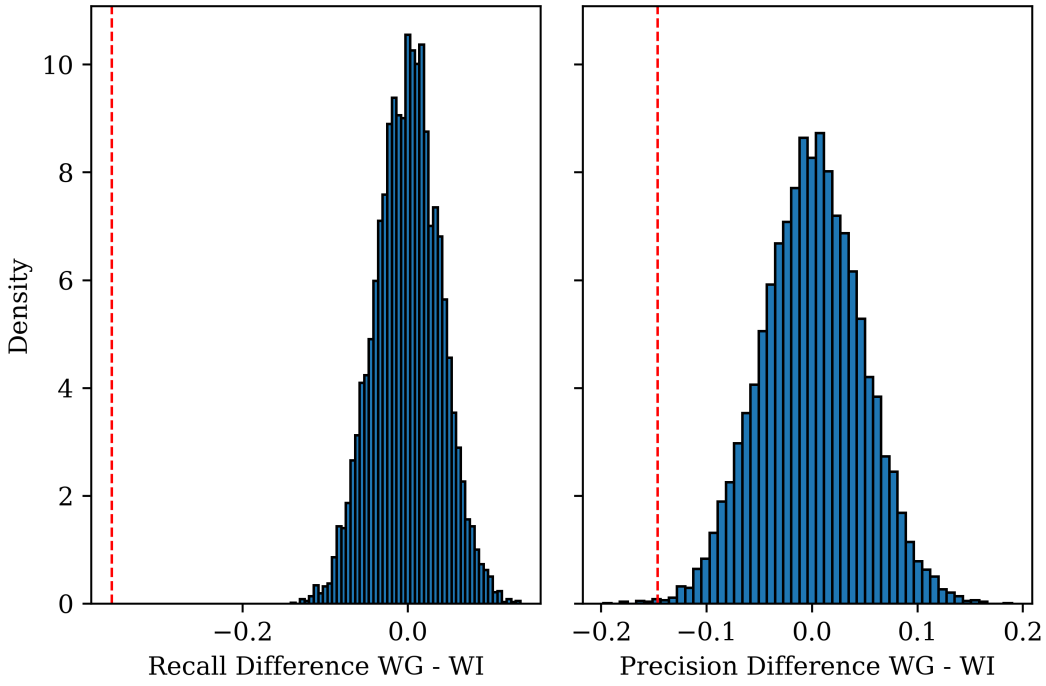


Figure 3.16: Sampling distributions of the difference in recall (left) and precision (right) between the whole-trajectory baseline algorithm (WG) and the whole-trajectory physically consistent method (WI)

For whole trajectories (Figure 3.16), the blue histograms are centred close to zero, as expected under the equal-performance null model, while the red dashed line lies far out in the left-hand tail for both recall and precision. Only 3 out of $10\,000$ simulations produce a precision difference at least as large as the observed one in favour of WI, and none of the simulations show a recall difference of comparable magnitude. In practical terms, it is extremely unlikely that the advantage of the MILP method in the whole-trajectory setting is just a consequence of sampling noise.

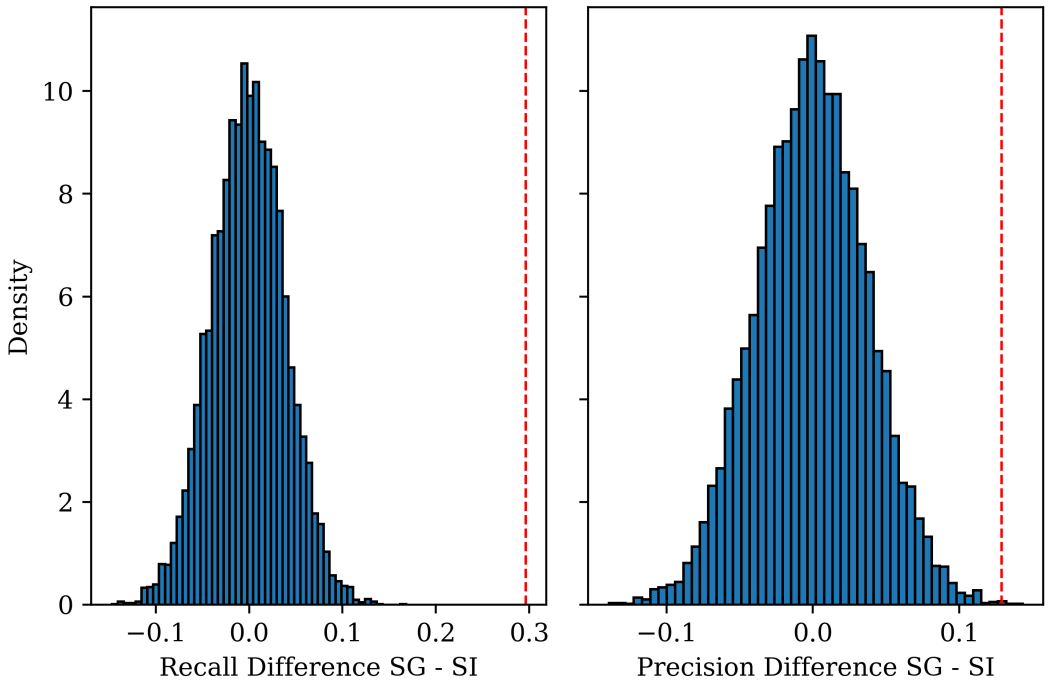


Figure 3.17: Sampling distributions of the difference in recall (left) and precision (right) between the segmented-trajectory baseline algorithm (SG) and the segmented-trajectory physically consistent method (SI)

For segmented trajectories (Figure 3.17), the simulated differences cluster tightly around zero, but the red dashed line now lies far out in the right-hand tail for both metrics, indicating a strong advantage for the baseline method. Only 4 out of $10\,000$ simulations generate a precision difference as large as the actual result in favour of SG, and none reach the observed recall difference. Thus, under the equal-performance null model, the probability of obtaining such strong segmented-performance gaps purely by chance is again extremely small.

A useful visual feature of these plots is that the red dashed line can fall outside the region where the bulk of the blue bars appear. This simply reflects the fact that none of the simulated experiments under the null hypothesis reproduce a difference as extreme as the actual result. Taken together, the figures indicate that the qualitative ordering of the algorithms (the physically consistent algorithm outperforms the baseline method for whole trajectories, and the baseline outperforms the in-house method for segmented trajectories) is very unlikely to be caused by the dataset size and variability only.

Chapter 4

Conclusion and Discussion

This thesis presented a physically consistent contrail-to-flight attribution algorithm based on the assumption that adjacent contrails must exhibit comparable advection error vectors, indicating spatially related wind field uncertainties. The algorithm showed that maintaining physical consistency among all contrail–candidate pairs in a scene can significantly enhance attributions (for the overall results, the physically consistent algorithm used on whole trajectories correctly attributed 20% more contrails than the baseline algorithm used on segmented trajectories) while remaining computationally efficient, achieving runtimes that are about one hour shorter than the baseline. Internal performance evaluations indicated that the whole trajectory datasets result in minimal computation times (less than 30 minutes for datasets containing up to 100 contrails), rendering this algorithm type appropriate for contrail mitigation applications.

4.1. Conclusion

This thesis introduced a physically consistent contrail-to-flight attribution algorithm built on the premise that neighbouring contrails should share similar advection error vectors, reflecting spatially correlated wind field uncertainties. The algorithm demonstrated that enforcing physical coherence across all contrail–candidate pairs within a scene can substantially improve attributions while being computationally efficient. In-house performance tests showed that, for typical scene sizes, the whole-trajectory datasets achieve low computation times, making this type of algorithm suitable for large-scale applications.

The ablation study confirmed that all three components of the objective function (the advection error penalty term, the altitude consistency rule, and the geometric score) contribute positively to overall attribution accuracy. Configurations missing any component degraded performance, whereas the full weighted combination consistently yielded the best results across all datasets. This reinforces the idea that contrail attribution requires not only geometric similarity but also physical consistency across nearby contrails.

At the global level, the physically consistent algorithm outperformed the baseline geometric method. When evaluated on the whole-trajectory synthetic datasets constructed in this thesis, the algorithm achieved high per-contrail precision and recall, while also significantly improving per-flight precision—a critical metric for contrail-mitigation strategies. Mitigation requires identifying which flights actually produce contrails, and high per-flight precision reduces false positives that could otherwise lead to unnecessary re-routing or operational penalties.

The synthetic datasets revealed performance patterns comparable to those reported by Sarna et al. (2025), though overall results were slightly lower. This is attributed to the presence of a substantial number of non-linear contrails in the synthetic data used in this thesis, whose curvature makes geometric alignment more difficult. While their dataset focuses exclusively on linear contrails, this thesis includes a broader and more challenging set of contrail shapes, which explains the performance gap (Sarna et al. (2025) report similar per-flight metrics but compared to the baseline algorithm used on segmented trajectories, but differences of up to 35% in per-contrail metrics).

At the same time, the evaluation still relies on a relatively small number of contrails: the synthetic datasets together contain only 304 contrails, and individual scenes often include just a few events.

This inevitably inflates the sampling uncertainty of the absolute performance metrics and makes one-to-one quantitative comparisons with large-scale studies such as Sarna et al. (2025), who use tens of thousands of contrails, only partially meaningful. The statistical analysis in Section 3.4 shows, however, that the relative differences between the algorithms studied here are statistically robust within this dataset, with empirical one-sided probabilities below 4×10^{-4} for all key comparisons. In other words, while exact values of precision and recall might need to be interpreted with caution, the qualitative ordering of the methods is unlikely to be an effect of sampling variability.

Finally, the results highlight that trajectory segmentation degrades physical-consistency performance. Segmenting trajectories introduces additional geometric and physical ambiguity. The physically consistent algorithm relies on context across contrails and across the full flight trajectory; segmentation reduces both. The choice of clustering rule also proved important: although the proposed method generally grouped nearby contrails effectively, suboptimal cluster definitions (especially clusters containing only one contrail) can weaken the physical-consistency constraint and reduce attribution quality.

Overall, the proposed algorithm successfully demonstrates that incorporating physical structure across the entire contrail field leads to more reliable and physically consistent attribution results. The framework provides a promising direction for future contrail tracking and mitigation strategies.

4.2. Discussion

While the physically consistent attribution algorithm shows strong performance, several directions for improvement remain. First, the synthetic dataset allows controlled experiments, but the true usefulness of the algorithm requires evaluation using actual contrails detected in satellite images and corresponding flight and weather data. Applying the method to real GOES-16 scenes would expose challenges such as noise in contrail detection, more complex appearance-lag effects, and potentially different traffic patterns.

Due to strict SAC/ISSR filtering and observation-window requirements, many scenes in the synthetic experiments contain relatively few contrails; in total, the evaluation datasets used in this thesis comprise 304 contrails, whereas Sarna et al. (2025) rely on tens of thousands. This difference in scale implies a larger statistical uncertainty for the absolute performance metrics reported here. Nonetheless, the dedicated statistical study in Section 3.4 indicates that the main performance gaps between algorithms in this study are highly unlikely to arise from sampling noise alone. The primary limitation is therefore not statistical significance within the synthetic settings, but real-life validity: it remains to be shown to what extent the observed advantages transfer to real-world attribution scenarios.

The current clustering rule avoids extremely large cluster sizes but may occasionally generate clusters containing only one contrail, removing the physical-consistency constraint. More advanced approaches - e.g., density-based clustering (Shah, 2012) - could produce more optimal spatial groupings. Ensuring that each cluster contains multiple contrails is particularly important for achieving stable background advection vectors. What is more, Section 3.3 has shown that, given the same datasets, multiple clustering configurations could be generated. To also account for this, the new approach should respect the following additional guidelines:

- Each cluster must contain at least 2 contrails
- Each cluster should always contain the same contrails if conditions are the same

The altitude consistency penalty uses a simple window based on average cluster altitude. A more sophisticated altitude rule (e.g, incorporating probabilistic ISSR height distributions) may improve discrimination between plausible and implausible matches.

Results indicated that non-linear contrails are inherently more difficult to attribute by geometric-based attribution algorithms. A dedicated evaluation on datasets containing only regular, linear contrails would more directly compare with the benchmarks of Sarna et al. (2025).

The present framework operates on a single frame. Combining it with multi-frame tracking, similar to Chevallier et al. (2023) or Sarna et al. (2025), could incorporate temporal coherence and further reduce misattributions, especially for older contrails.

References

Geraedts, Scott et al. (2024). “A scalable system to measure contrail formation on a per-flight basis”. In: *Environmental Research Communications* 6.1. Publisher: IOP Publishing, p. 015008. ISSN: 2515-7620. DOI: 10.1088/2515-7620/ad11ab. URL: https://dx.doi.org/10.1088/2515-7620/ad11ab.

Lee, D. S. et al. (2021). “The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018”. In: *Atmospheric Environment* 244, p. 117834. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2020.117834. URL: https://www.sciencedirect.com/science/article/pii/S1352231020305689.

Meijer, Vincent R. (2024). “Satellite-based Analysis and Forecast Evaluation of Aviation Contrails”. Accepted: 2024-06-27T19:46:44Z. Thesis. Massachusetts Institute of Technology. URL: https://dspace.mit.edu/handle/1721.1/155350.

Sarna, Aaron et al. (2025). “Benchmarking and improving algorithms for attributing satellite-observed contrails to flights”. In: *EGUsphere*. Publisher: Copernicus GmbH, pp. 1–58. DOI: 10.5194/egusphere-2024-3664. URL: https://egusphere.copernicus.org/preprints/2025/egusphere-2024-3664/.

Meijer, Vincent R., Luke Kulik, et al. (2022). “Contrail coverage over the United States before and during the COVID-19 pandemic”. In: *Environmental Research Letters* 17.3. Publisher: IOP Publishing, p. 034039. ISSN: 1748-9326. DOI: 10.1088/1748-9326/ac26f0. URL: https://dx.doi.org/10.1088/1748-9326/ac26f0.

Brewer, A. W. (1946). *Condensation Trails*. Wiley Online Library. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/j.1477-8696.1946.tb00024.x.

Appleman, H. (1953). “The Formation of Exhaust Condensation Trails by Jet Aircraft”. In: Section: Bulletin of the American Meteorological Society. DOI: 10.1175/1520-0477-34.1.14. URL: https://journals.ametsoc.org/view/journals/bams/34/1/1520-0477-34_1_14.xml.

Schumann, U. (1996). “On conditions for contrail formation from aircraft exhausts”. In: *Meteorologische Zeitschrift* 5. URL: https://www.osti.gov/etdeweb/biblio/233640.

Gierens, Klaus, Sigrun Matthes, and Susanne Rohs (Dec. 2020). “How Well Can Persistent Contrails Be Predicted?” In: *Aerospace* 7.12. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute, p. 169. ISSN: 2226-4310. DOI: 10.3390/aerospace7120169. URL: https://www.mdpi.com/2226-4310/7/12/169.

Spichtinger, Peter and Martin Leschner (2016). “Horizontal scales of ice-supersaturated regions”. In: *Tellus B: Chemical and Physical Meteorology* 68.1, p. 29020. DOI: 10.3402/tellusb.v68.29020. URL: https://doi.org/10.3402/tellusb.v68.29020.

Kärcher, B. et al. (2018). “Contrail Formation: Analysis of Sublimation Mechanisms”. In: *Geophysical Research Letters* 45.24. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL079391, pp. 13, 547–13, 552. ISSN: 1944-8007. DOI: 10.1029/2018GL079391. URL: https://onlinelibrary.wiley.com/doi/abs/10.1029/2018GL079391.

“Anthropogenic and Natural Radiative Forcing” (2014). In: *Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by Intergovernmental Panel on Climate Change (IPCC). Cambridge: Cambridge University Press, pp. 659–740. ISBN: 978-1-107-05799-9. DOI: 10.1017/CBO9781107415324.018. URL: https://www.cambridge.org/core/books/climate-change-2013-the-physical-science-basis/anthropogenic-and-natural-radiative-forcing/63EB1057C36890FEAA4269F771336D4D.

Schumann, Ulrich (2005). “Formation, properties and climatic effects of contrails”. In: *Comptes Rendus Physique*. Aircraft trailing vortices 6.4, pp. 549–565. ISSN: 1631-0705. DOI: 10.1016/j.crhy.2005.05.002. URL: https://www.sciencedirect.com/science/article/pii/S1631070505000563.

Märkl, Raphael Satoru et al. (Mar. 27, 2024). “Powering aircraft with 100% sustainable aviation fuel reduces ice crystals in contrails”. In: *Atmospheric Chemistry and Physics* 24.6. Publisher: Coperni-

cus GmbH, pp. 3813–3837. ISSN: 1680-7316. DOI: 10.5194/acp-24-3813-2024. URL: https://acp.copernicus.org/articles/24/3813/2024/.

Hofer, Sina, Klaus Gierens, and Susanne Rohs (July 11, 2024). “How well can persistent contrails be predicted? An update”. In: *Atmospheric Chemistry and Physics* 24.13. Publisher: Copernicus GmbH, pp. 7911–7925. ISSN: 1680-7316. DOI: 10.5194/acp-24-7911-2024. URL: https://acp.copernicus.org/articles/24/7911/2024/.

Ng, Joe Yue-Hei et al. (2023). *OpenContrails: Benchmarking Contrail Detection on GOES-16 ABI*. DOI: 10.48550/arXiv.2304.02122. arXiv: 2304.02122[cs]. URL: http://arxiv.org/abs/2304.02122.

Chevallier, Rémi et al. (2023). “Linear Contrails Detection, Tracking and Matching with Aircraft Using Geostationary Satellite and Air Traffic Data”. In: *Aerospace* 10.7. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, p. 578. ISSN: 2226-4310. DOI: 10.3390/aerospace10070578. URL: https://www.mdpi.com/2226-4310/10/7/578.

Suo, Chunnan et al. (2024). “Quality Assessment of ERA5 Wind Speed and Its Impact on Atmosphere Environment Using Radar Profiles along the Bohai Bay Coastline”. In: *Atmosphere* 15.10. Publisher: Multidisciplinary Digital Publishing Institute, p. 1153. ISSN: 2073-4433. DOI: 10.3390/atmos15101153. URL: https://www.mdpi.com/2073-4433/15/10/1153.

Gryspeerdt, Edward et al. (2024). “Operational differences lead to longer lifetimes of satellite detectable contrails from more fuel efficient aircraft”. In: *Environmental Research Letters* 19.8. Publisher: IOP Publishing, p. 084059. ISSN: 1748-9326. DOI: 10.1088/1748-9326/ad5b78. URL: https://dx.doi.org/10.1088/1748-9326/ad5b78.

Barbosa, Maria Paula (2024). “Relationship between synoptic scale meteorology, aircraft parameters, and observable contrails”. PhD thesis. Massachusetts Institute of Technology.

Zhang, Damao et al. (2018). “Ice particle production in mid-level stratiform mixed-phase clouds observed with collocated A-Train measurements”. In: *Atmospheric Chemistry and Physics* 18.6. Publisher: Copernicus GmbH, pp. 4317–4327. ISSN: 1680-7316. DOI: 10.5194/acp-18-4317-2018. URL: https://acp.copernicus.org/articles/18/4317/2018/.

Strohmeier, Martin, Xavier Olive, and Junzi Sun (2022). “Evading the Public Eye: On Astroturfing in Open Aviation Data”. In: *Engineering Proceedings* 28.1. Publisher: Multidisciplinary Digital Publishing Institute, p. 7. ISSN: 2673-4591. DOI: 10.3390/engproc2022028007. URL: https://www.mdpi.com/2673-4591/28/1/7.

Contrails API (2025). Contrails API. URL: https://apidocs.contrails.org/index.html.

Sun, Junzi (2021). *The 1090 Megahertz Riddle: A Guide to Decoding Mode S and ADS-B Signals*. TU Delft OPEN Books. ISBN: 978-94-6366-402-8. URL: https://books.open.tudelft.nl/home/catalog/book/11.

Meijer, Vincent R., Sebastian D. Eastham, et al. (Oct. 23, 2024). “Contrail altitude estimation using GOES-16 ABI data and deep learning”. In: *Atmospheric Measurement Techniques* 17.20. Publisher: Copernicus GmbH, pp. 6145–6162. ISSN: 1867-1381. DOI: 10.5194/amt-17-6145-2024. URL: https://amt.copernicus.org/articles/17/6145/2024/.

Aircraft Performance Database (2025). Aircraft Performance Database. URL: https://contentzone.eurocontrol.int/aircraftperformance/.

Hersbach, Hans et al. (2020). “The ERA5 global reanalysis”. In: *Quarterly Journal of the Royal Meteorological Society* 146. Publisher: Wiley ADS Bibcode: 2020QJRMS.146.1999H, pp. 1999–2049. ISSN: 0035-9009. DOI: 10.1002/qj.3803. URL: https://ui.adsabs.harvard.edu/abs/2020QJRMS.146.1999H.

Wilks, Daniel S. (2006). *Statistical Methods in the Atmospheric Sciences*. Google-Books-ID: _vSwyt8_OGEC. Academic Press. 650 pp. ISBN: 978-0-12-751966-1.

Gierens, K. and P. Spichtinger (2000). “On the size distribution of ice-supersaturated regions in the upper troposphere and lowermost stratosphere”. In: *Annales Geophysicae* 18.4. Publisher: Copernicus GmbH, pp. 499–504. ISSN: 0992-7689. DOI: 10.1007/s00585-000-0499-7. URL: https://angeo.copernicus.org/articles/18/499/2000/.

Jin, Xin and Jiawei Han (2011). “K-Means Clustering”. In: *Encyclopedia of Machine Learning*. Springer, Boston, MA, pp. 563–564. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_425. URL: https://link.springer.com/rwe/10.1007/978-0-387-30164-8_425.

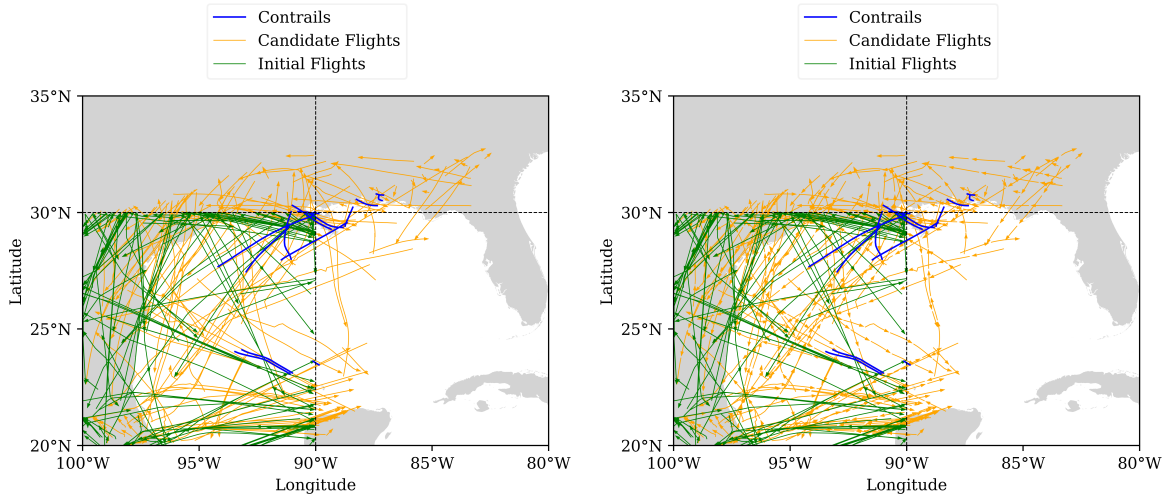
Optimization with PuLP — PuLP 3.3.0 documentation (2025). URL: https://coin-or.github.io/pulp/.

Shah, Glory H. (2012). “An improved DBSCAN, a density based clustering algorithm with parameter selection for high dimensional data sets”. In: *2012 Nirma University International Conference on Engineering (NUiCONE)*. 2012 Nirma University International Conference on Engineering (NUiCONE). ISSN: 2375-1282, pp. 1–6. DOI: 10.1109/NUICONE.2012.6493211. URL: https://ieeexplore.ieee.org/document/6493211.

Appendix A

Evaluation Datasets

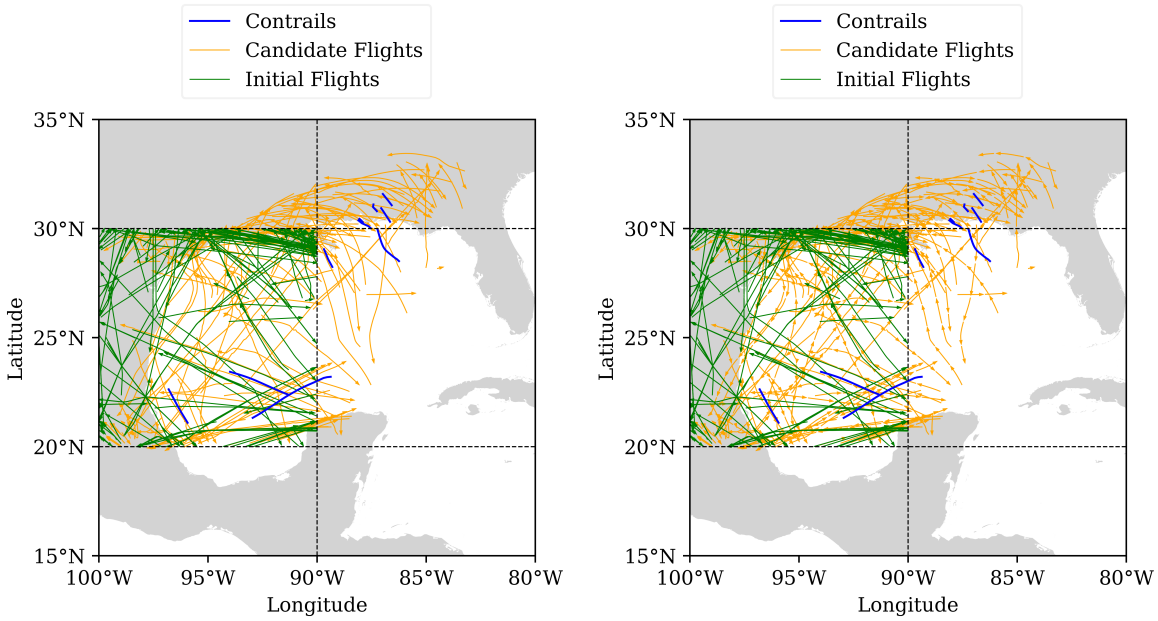
This appendix presents all the synthetic datasets used to evaluate the physically consistent contrail-to-flight attribution algorithm. Each dataset corresponds to one combination of geographic region, season, and time-of-day interval defined in Subsection 2.5.1. The two selected regions span distinct parts of the East coast of the United States of America, while the winter and summer sampling windows capture contrasting atmospheric states and contrail-formation environments. For each season, both a noon and an evening interval were chosen to reflect typical diurnal patterns in contrail coverage and air-traffic density. The figures in this appendix display, for every configuration, the resulting synthetic datasets generated under ISSR/SAC-valid conditions for both the whole-trajectory and segmented-trajectory types of datasets.



(a) Whole trajectories dataset

(b) Segmented trajectories dataset

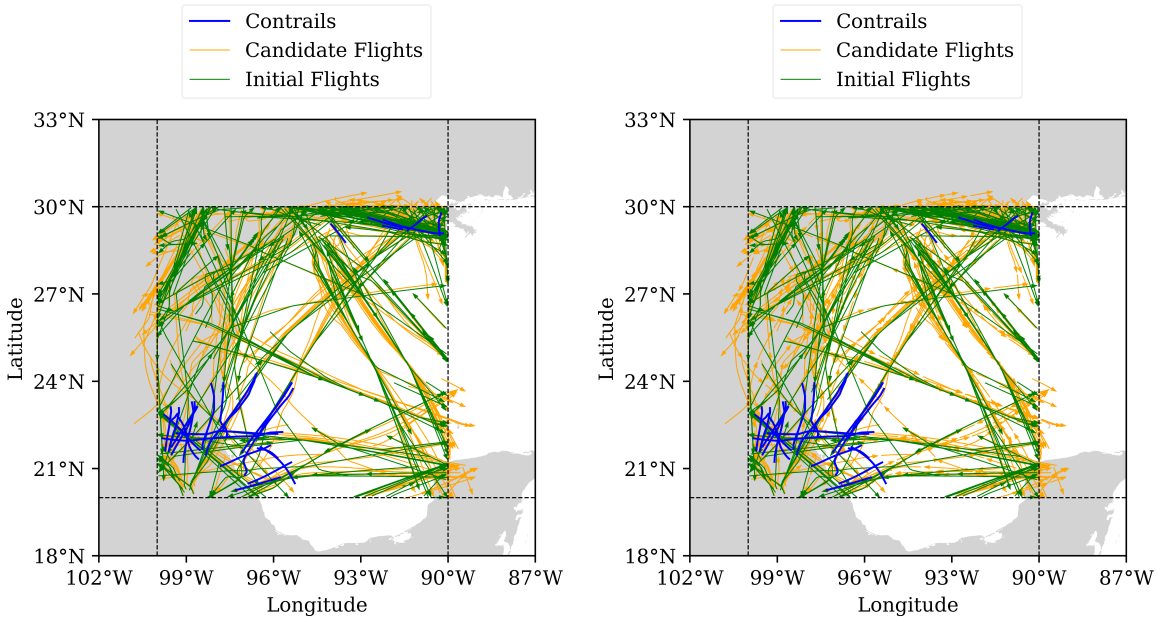
Figure A.1: First region, winter, noon



(a) Whole trajectories dataset

(b) Segmented trajectories dataset

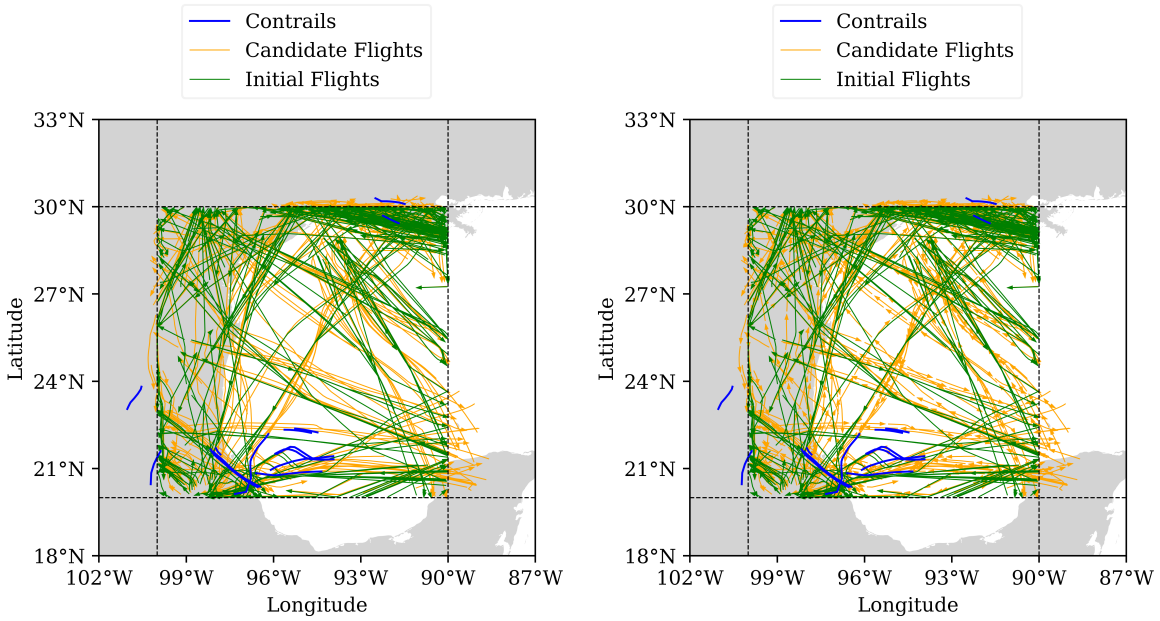
Figure A.2: First region, winter, evening



(a) Whole trajectories dataset

(b) Segmented trajectories dataset

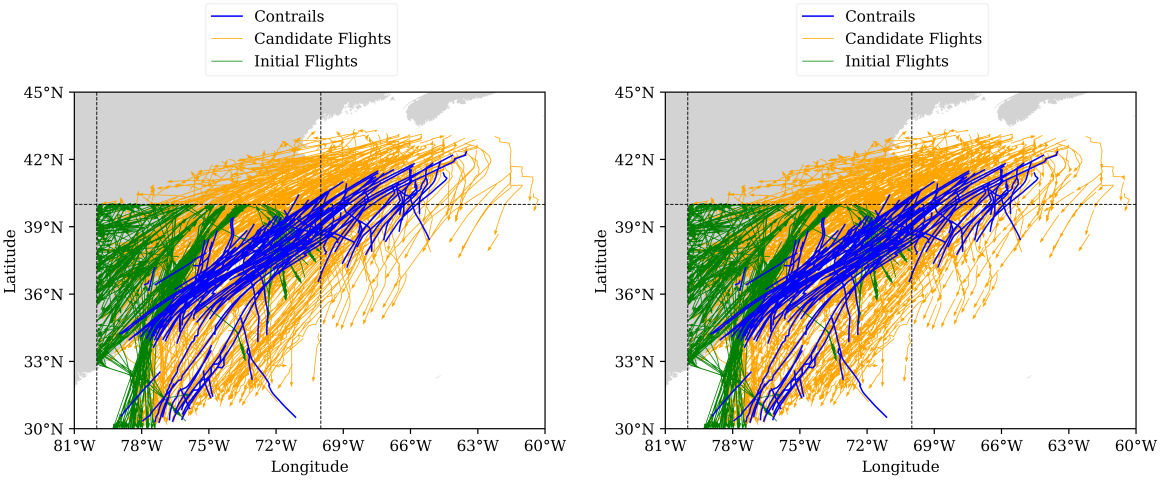
Figure A.3: First region, summer, noon



(a) Whole trajectories dataset

(b) Segmented trajectories dataset

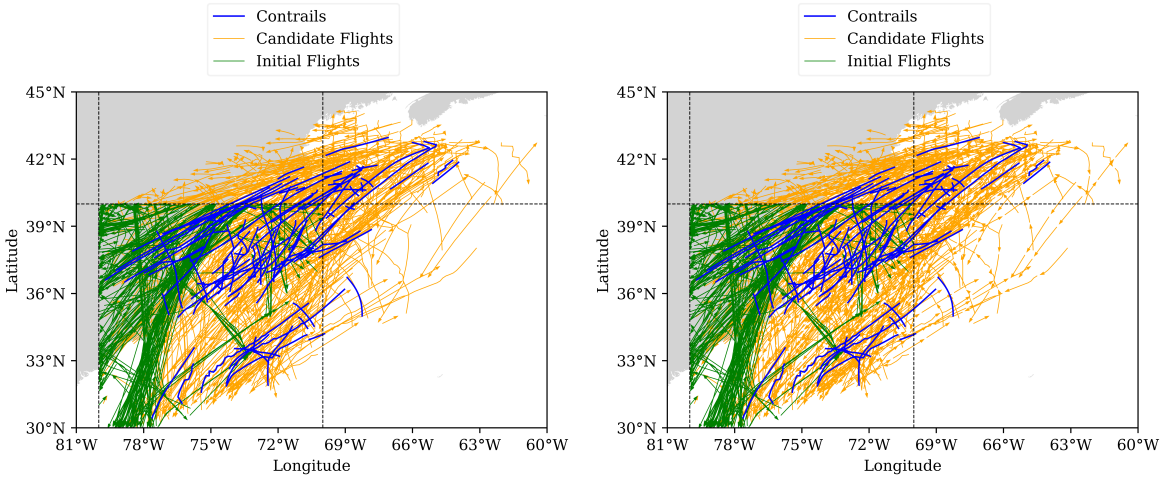
Figure A.4: First region, summer, evening



(a) Whole trajectories dataset

(b) Segmented trajectories dataset

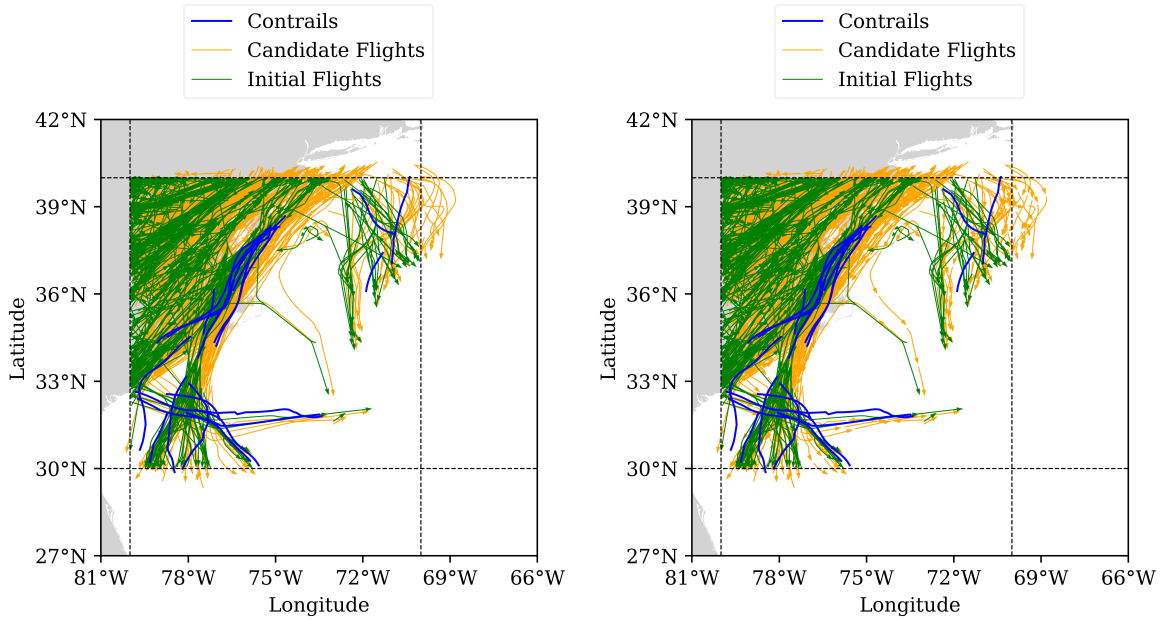
Figure A.5: Second region, winter, noon



(a) Whole trajectories dataset

(b) Segmented trajectories dataset

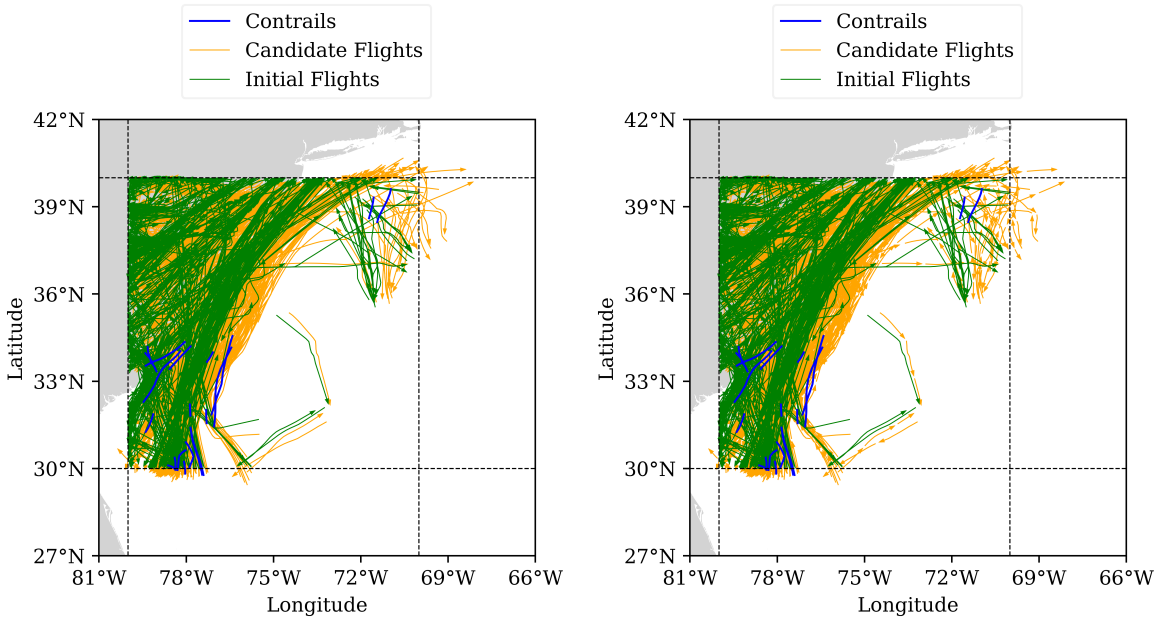
Figure A.6: Second region, winter, evening



(a) Whole trajectories dataset

(b) Segmented trajectories dataset

Figure A.7: Second region, summer, noon



(a) Whole trajectories dataset

(b) Segmented trajectories dataset

Figure A.8: Second region, summer, evening