Summary

In the work "Benchmarking and improving algorithms for attributing satellite-observed contrails to flights", the authors demonstrate an improvement in attributing contrail detections to segments of generating flights—by utilising the way that the difference between a contrail's location and the position of an advected flight track compounds as the advection time is increased. As well as enabling attribution of contrails in groups (which minimises detections available for 'false positive' attribution), the offset is used to attribute a group of detections more-confidently to the flight than using 'single-frame' attribution methods, using the fact that the offset (interpretable for a match as 'advection error') tends to zero at the time of the flight.

This algorithm is benchmarked against other attribution algorithms (and tuned) using "synthetic contrail observations"—specifically, linear geometries based on a rasterised optical thickness field using output from CoCiP (Schumann, 2012, A 'Contrail Cirrus Prediction' model).

General comments

This study signficantly furthers attempts at contrail attribution, developing a benchmarking framework, a new, performant model, and an interesting exploration of its physical basis. A comprehensive dataset of attributed contrails is needed to benchmark attribution algorithms, but remains unachievable in the immediate term. This work highlights this issue effectively and constructs an alternative using model outputs, and appears to be the first comprehensive dataset effective in highlighting the limitations of current attribution algorithms, and distinguishing between them.

These methods and associated conclusions are clearly described in the text, including the underlying assumptions, and are well-captured by the abstract and title. The results support the conclusions made.

The manuscript is long, and is frequently repetitive, despite the clarity and importance of the results. In particular, many appendices are included; these are overused and unstructured. Some appendices contain analysis beyond the manuscript or important to include in the body to support discussions. Nonetheless, these are very valuable results, with a methodology which seems largely sound. Although a particularly long list of comments has been included below, most are minor suggestions, and I look forward to seeing the final version. It is recommended for publication following minor revisions.

Specific comments

The larger points of concern are presented here, and are followed by many more minor points and suggestions, several of which are stylistic and are left to the authors' discretion.

1. Many appendices have been presented in an unstructured way (i.e. all part of Appendix A; not listed the same order they are referenced), and often go beyond the analysis or discussion included in the paper, these further analyses should be included in the manuscript body or removed. A number of suggestions have been included in the list below. They should be appropriately structured into sections and subsections. A supplement should be considered for anything that cannot reasonably be included (though a supplement should also be supporting, and not significant analysis and discussion omitted from the body).

2. Lines 268–273 (and Fig. 4, 5(a,b)): Clarify why one needs to group contrails attributed across a flight, rather than starting with Fig. 5(b), wherein only overlapping waypoints are considered. Given the single-frame attributions are the result of advecting flight waypoints, ungrouping overlapping waypoints doesn't represent an important step for the new algorithm. For Fig. 4 it would be more valuable to present ambiguous cases from overlapping waypoints.

3. Line 381–383: The ERA5 ensemble control run is not independent of the ERA5 operational analysis data. Using accurate weather data (i.e. output from the same model) would mean that the metrics were not benchmarking the differences in weather models. Although it is true that this would lead to overestimates of precision, it is already acknowledged that the benchmark is only appropriate for relative comparisons.
   The current approach also means that a future algorithm could falsely perform well only by using the weather data used to produce the benchmarking dataset.
   The applicability of the data is somewhat discussed in Appendix A7—but the errors between models do not occur randomly, so only matching the distribution doesn't mean that the advection error is realistically distributed.
   A better alternative to this approach could be to use the same ERA5 data used for the attribution algorithm, but perturb it by a field representing some estimated error.

4. Lines 478–471: Specify why it is valuable for 'per-flight' recall and precision to be evaluated without requiring attributions to be correct. Will this falsely reward inaccurate attributions in flight-dense regions? In the prose definitions, the precision and recall are hard to differentiate—Appendix A19 is clearer and should be moved to the body of the manuscript.
   Table 1: The per-frame and global quantities struggle to distinguish between relative performance between different algorithms, other than the single-frame algorithm's flight recall. Given the expectation that global quantities are biased towards dense scenes, and the flight recall does not impose accurate attributions, could it be the case that this algorithm is only being rewarded for making false attributions in dense scenes? (This seems to be indicated by Appendix A26's decreased precision seen when including single-frame attributions for ambiguous cases, and Fig. A8's ever-increasing flight recall for the single-frame attribution). The analysis of Appendix A26 should be in the body of the manuscript.

Given the otherwise-similar relative performance of the algorithms, can a single approach (per-frame or global) be presented as the best way to benchmark?

5. Section 3.1: Specify whether all the benchmarked algorithms were tuned on the SynthOpenContrails, and if this would have an impact on their relative performance. Some of the context of Appendix A25 is important for the body of the manuscript, including the fraction of time spans that were used for the tracking algorithm, and whether this has any impact on the results (i.e. how do the other algorithms perform on this subset).

6. Table 1: The standard deviations in the per-frame quantities are significant, as expected based on the variability in Fig. A6. Would this variability be even higher if stratified by 'similar scenes' (Fig. A6 shows similar statistics in consecutive frames with occasional jumps). Perhaps a per-time-span average would better capture variability? Can estimates of uncertainty in precision and recall be made? Are more time spans needed to accurately benchmark?

7. Line 292: Discuss why the slope is an indicator that a fit contains repeated detections of the same contrail—that a real contrail associated with a specific flight would have less error in advection and therefore a shallower slope? Clarify whether a steep negative slope is considered a 'lower Ssc'. Clarify whether Cslope has any impact for Ssc (is it constant across candidate fits?), if its impact is only in Sfit, it could be removed from equation 3 and only included in equation 4. The preference for maximum number of detections above a high Ssc could be clarified by specifying that Ssc is high unless Nslope=max(Nslope). Altogether, could lines 286–297 could be simplified as 'selecting the candidate fit with the shallowest gradient with the largest number of inliers'?

Minor comments and suggestions follow. Many of them are stylistic, which might help to improve clarity. More substantial points are in bold

Abstract

The abstract clearly describes the aims and motivations of the work. It could be useful for the order of the abstract (i.e. benchmark and new algorithm) to be the same as the text.

Line 8: 'synthetic contrail observations'—rephrasing as 'synthetic contrail detections' could make clearer that these are geometry definitions rather than imitating satellite data.

Line 9–10: emphasise that the metrics are appropriate only for relative performance, and are not appropriate for making inference about these 'real world' applications (per Section 2.4.3).

Introduction

Lines 22–58: This part of the introduction motivates contrail attribution using the need to relate contrail observations to the forming aircraft, to be able to relate contrail properties to aircraft properties, for "validating contrail forecasts" (line 49), validate avoidance results, and for contrail altitude estimation.
A shorter version of this part would be sufficient to motivate contrail attribution, and could be structured to highlight the value of attribution more specifically. Further methodological comments are below, although abbreviating this section could negate the need to treat these in the text while keeping the study sufficiently motivated.

Line 25: The frequency and morphology of ISSRs indicates that trajectory modifications are minor, and could lead one to conclude that few flights cause persistent contrails, but it's not clear from this that the contrail impact that does occur could be largely mitigated from deviating only a small subset of contrail-forming flights in practice—this is a modelling result that may differ by region (e.g. Teoh et al., 2020).

Lines 29–35: Here, the modelling results establishing the physical basis for trajectory modification (which assume that meteorological data allows for contrail forecasts to be correct in terms of the statistics of contrail occurrence), are conflated with the application of the data for trajectory avoidance in practice (which may require a "perfect" forecast for success in individual cases).

Lines 63–76: The differences between these algorithms could be highlighted and their deficiencies relative to this study, rather than briefly summarising the full methodology which is similar in each case. In particular differences between this study and the Gryspeerdt et al. (2024) approach are not made clear.

Lines 62, 70, 75: NWP weather data should be disambiguated from reanalysis weather data (such as ERA5).

Lines 88–112: This discussion of contrail avoidance monitoring and validation is an effective and simpler motivating application than supporting model validation. This part of the introduction might be better placed earlier in the introduction, with the other motivating aims.

Line 114: Does the new algorithm improve on the scalability of Geraedts et al. (2024)?

Lines 114–115: It doesn't follow that scalability and lack of relative benchmarking limit the application of attribution, it seems likely that the performance of attribution (and detection) methods are a more significant factor. If anything, existence of applications indicates that if an algorithm worked well, it would see use.

Lines 115–121: This outline at the end of the introduction could be expanded by including 'signposting' reference to specific sections, rather than the current approach of outlines at the top of each section, which makes the manuscript feel repetitive.

Methods

This section could benefit from restructuring. In particular, work for which a suitable citation exists is described in more detail than necessary in section 2.2.2. It may also benefit from separating into separate sections: Section 2 being the Contrail-to-Flight Attribution algorithm containing Section 2.2, and any necessary information from 2.4.3 (i.e. that it can be tuned), then a separate section 3 to introduce the synthetic contrail detections. This would enable e.g. the splitting of Section 2.2.3 into simpler subsections.

Lines 123–125, 134–135: These section introductions outside of subsequent subsections aren't necessary, and lead to repetition. Signposting at the end of the introduction is the most appropriate place for this context.

Section 2.1: This section is a little confusing when separated from the context where the notation is used, and the notation is not so complex that it is required—it could be removed, with symbols introduced at the time they are used, or including one of the appendices' symbols tables instead.

Contrail to flight attribution algorithm

Sections 2.2.1 and 2.2.2: These sections largely follow Geraedts et al. (2024). It could be worth highlighting this even more strongly at the beginning of Section 2.2.1. Section 2.2.2 describes the algorithm in considerable detail which might not be needed to understand or reproduce the improvements shown here. As long as the W parameter were sufficiently introduced as the cross-track offset between the advected flight track and detected contrail, relying on the Introduction's description of this algorithm may even be sufficient, particularly if adapted to include Fig. 2. The discussions of lines 177–181 and 209–215 are valuable in motivating this study and should not be removed in simplifying this section—if the authors decide not to significantly simplify here, these could at least be brought to be more prominent.

Line 139: This is the best-achieved spatial resolution of GOES-R's ABI and occurs at nadir—clarify that this won't be achieved over the region studied.

Line 143: 'several time spans'—specify how many.

Line 164: Refine 'separated by roughly 10 hPa'—this is not true of ERA5 L137 model levels.

Line 190–191: The use of u and w as coordinate names is slightly confusing as they are conventionally associated with wind speeds, but their use is understandable for consistency with Geraedts et al. (2024). It might be worth particularly specifying them as being spatial.

Line 200: clarify W and V are distances, not translations and that θ is an angle.

Line 207: As discerning them is a key strength of CoAtSaC, it would be valuable to show how often the previous algorithm was forced to rely on the 'additional logic' to disambiguate detections, or discuss this in the context of the per-contrail recall.

Section 2.2.3: As mentioned above, this section (which introduces the new insights of CoAtSaC) could benefit from restructuring to feature near the start of its own main body section (potentially enabling contrail grouping to be separated from the full algorithm). The section itself is a little repetitive, for example, the 'fitting' methodology is described in brief (ln. 218–220), in figure captions (Figs 3,5), in a description of the figure in the text (ln 225–245), and in the implementation (ln 268–312). While the visualisations are effective and valuable, as long as they have clear captions, one clear in-text explaination should be sufficient.

Fig. 3: Panel (a) may be clearer if the individual frames were plotted on separate smaller axes, especially in clarifying cases such as Contrail D, whose associate flight track is a full-frame over-advected, but similar colours make this hard to distinguish.

Line 233: It might be worth noting whether the corresponding waypoints of flight 2 occurred before or after contrail 1 was first detected.

Lines 247–250: 'we acknowledge that we do not know the ages of the advected contrails'—this is confusing following the description of Fig. 3 because implied contrail age is already clear to be the time information necessary, as each analysis is working with a specific hypothesis flight whose time information is known, as is the time of the detection. Clarify this introduction of 'implied contrail age'.

Fig. 4: May benefit from a different palette from Fig. 3. (a) Clarify whether 1 or 2. (c) Clarify whether additional causes could be evolution of the particular contrail, advection error in the v direction from a later contrail, and whether this is a match in CoAtSaC. (d) Clarify whether Sattr threshold is actually ignored in this work/whether speaking of 'available to match to other flights' under different algorithms. Clarify that speaking of 'contrail detections' rather than contrails, particularly (g)–(i).

Fig. 5: (e) clarify whether the offset parameter should ever be decreasing when the contrail ages (as is the case for the maroon fit), and if not, whether such contrails could be removed—clarify the benefit of instead using a 'fit score'. The inclusion of this fit (and reliance on a conflicting attribution) seems to be at odds with the claim of lines 242–244, that the near-zero W-intercept is fundamental to the algorithm.

Line 257: Clarify if this slope condition is equivalent to ensuring that no two simultaneous, spatially-separated contrail detections can be attributed to the same generating flight?

Line 280: More meaningful notation than m and b could be used. Clarify if m and dW/dt are the same quantity.

Line 305: clarify that 'each flight' is 'each group of flight waypoints with overlapping single-frame attributed contrails'.

Fig. 6: Does Fig. 3 not highlight a similarly rejectable scenario for contrail 1 as matched to flight 2? If so, this figure may not be necessary.

Line 307: What is the impact of including Ssc to this score? Although enhancing confidence in this being a single contrail, shallow slopes reduce the confidence in the 'low W-intercept' condition.

Line 340: If scalability is critical, it would be good to clarify if this approach is any more scalable than the algorithms mentioned in the introduction, specifically Geraedts et al. (2024).

Lines 344–345: Clarify if any other attribution attempts propose such a methodology.

Synthetic contrail benchmark dataset

Line 347: The value of the benchmark dataset goes beyond tuning the hyperparameters—the segue could be removed.

Line 364: The name 'SynthOpenContrails' is chosen to parallel the 'OpenContrails' dataset of Ng et al. (2023). Make explicit that the new dataset is not suitable for benchmarking contrail detection algorithms, only attribution algorithms, and that a performant contrail detection is assumed.

Line 374–375: Clarify 'flight loading purposes'.

Line 380: Clarify 'weather data to use' for what.

Tuning and Benchmarking

Lines: 440–442: This introduction is not necessary.

Line 455: Could whether or not it is the case that multiple sets of parameters match the real data be specified with more confidence? Is there any motivation to the specific set chosen?

Line 459: The content of Appendix A21 would be valuable in the body of the manuscript. Comment on whether the train split was used for manual tuning, and why a better match could not be obtained in terms of contrail pixels and linear contrails.

Line 453: Specify what it means for the metrics to be computed globally, the region of study is the Contiguous US.

Line 474: Fig. A6 might be valuable to include in the manuscript body. Suggest that this is then used to justify calculating per-frame quantities when results are introduced rather than here, to avoid 'teasing' the results.

Lines 481–483: Clarify this. It might be better placed as part of an applicability discussion after the results are presented.

Lines 485–489: This is a repeated introduction in abstract, and isn't necessary.

Line 498: The results of the sensitivity could be very briefly summarised here. Given the relatively straightforward outcome, the appendix could be removed. If the contents is kept, specify which precision is meant for the result of the single-frame algorithm (line 941)

Results

Section 3.2: Although performance relative to each other is approximately constant, the relative performance between bins of different contrail properties is interesting and could be expanded by including some of Appendix A26—which includes significant analysis and discussion beyond the contents of the manuscript's body. The discussions of contrail density and altitude are especially important. Given the error in meteorological data is hypotehsised as having significant impact, could this be resolved? The discussion of contrail age and width is also valuable—does this speak to limitations of using linear objects as contrails?

Technical corrections

In Algorithm 1, using e.g. italicised text rather than mathematics text may improve legibility.

Line 431: 'subrouting' should read 'subroutine'.

References

Geraedts, Scott et al. (2024). "A Scalable System to Measure Contrail Formation on a Per-Flight Basis". In: *Environmental Research Communications* 6.1, p. 015008. ISSN: 2515-7620. DOI: 10.1088/2515-7620/ad11ab.

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, p. 084059. ISSN: 1748-9326. DOI: 10.1088/1748-9326/ad5b78.

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

Schumann, U. (2012). "A Contrail Cirrus Prediction Model". In: *Geoscientific Model Development* 5.3, pp. 543–580. ISSN: 1991-9603. DOI: 10.5194/gmd-5-543-2012.

Teoh, Roger et al. (2020). "Mitigating the Climate Forcing of Aircraft Contrails by Small-Scale Diversions and Technology Adoption". In: *Environmental Science & Technology* 54.5, pp. 2941–2950. ISSN: 0013-936X. DOI: 10.1021/acs.est.9b05608.