Response to Reviewer 1's Comments

The Authors train a machine learned (ML) model on a classification task (contrail absent, short-lived contrail, persistent contrail) using as model input imagery from aircraft-mounted rear-pointing cameras from two recent contrail study campaigns that involved a chase plane studying contrail properties and measuring ambient weather conditions. The labels for the classification task training come from applying Schmidt-Appleman criteria and RHi > 100% to the chase plane's ambient atmospheric conditions measurements. The Authors achieve F1 scores between 0.83 and 0.93 when evaluating this model in a k-fold heldout cross validation scheme.

In its current state I cannot recommend accepting the manuscript; I am on the fence and defer to the editors about whether subsequent major revisions should have a chance to be reconsidered..

We would like to thank the reviewer for their very constructive feedback. The responses to reviewer's suggestions and the associated revisions in the manuscript are provided below. The comments identify genuine weaknesses in how we framed the scientific contribution, in the transparency of our dataset and evaluation choices, and in our citation practice. In the revised manuscript we (i) re-evaluate the classifier under a cross-validation scheme, (ii) add a human visual-verification study comparing our thermodynamic labels against the imagery, (iii) reframe the class definitions and claims around "persistence-conducive conditions" rather than observed persistence, (iv) add baseline model comparisons and uncertainty estimates, and (v) correct and check the citations.

---

1. The case for scientific significance is not compellingly made:

While the manuscript is often clear enough and reproducible, the scientific significance of these F1 scores is not discussed in a compelling way. The manuscript mentions possible use of aircraft-mounted cameras as possibly decreasing expense and therefore accelerating rollouts and availability of data relevant to contrail weather forecasting, but the dots are not clearly connected: contrail forecasting is an important area to improve, but no evidence is provided (nor citations are made) for the assertion that rear-pointed camera data will improve the accuracy of contrail forecasts.

In retrospect, we agree that the manuscript asserted this benefit without sufficiently constructing the supporting argument, especially in the introduction, and we have rewritten the relevant parts

of the Introduction and Discussion Sections to make the chain of reasoning explicit and referenced. A key goal of this paper is to provide recommendations regarding camera placement for future flight measurement campaigns. We made some of these points in the discussion and some in the conclusions sections separately, but then put more focus on the ML techniques than we probably should have, in an attempt to make our assumptions clear.

The role we envision for the onboard cameras is that of a scalable, low-cost way to gather observational data, by having a fleet of camera-equipped aircraft acts as a distributed network of sensors that report where formation and ice supersaturated conditions are actually encountered, which can be used for forecast validation and for data assimilation studies and in-flight operational decisions. In the revised manuscript, instead of focusing on rear-pointed camera data potentially improving the forecast accuracy, we will focus on using this additiona measurement data that can be used towards contrail prediction and that camera placement onboard the aircraft does make a significant difference, per our observations. We will also clarify our contributions accordingly.

We based our argument on previous published work within the community but did not state our thought process or the final assumption clearly. The limiting factor in contrail forecasting for persistence is the model humidity field. From Gierens (2020) and from Hofer (https://acp.copernicus.org/articles/24/7911/2024/, 2024), predicting ice supersaturation at cruise altitudes from ERA5 humidity yields low ETS. Hofer attributed this to the strong variability in the water vapour field, the low number of humidity measurements at the air traffic altitude, and the oversimplified parameterisations of cloud physics in weather models. As also explicitly described by Hofer et al. " This problem (the challenging prediction of persistent contrails) is intensified by the low number of humidity measurements at cruise levels for data assimilation. Data assimilation is necessary to keep the simulation of a complex system close to measured reality. Therefore, more data on relative humidity at flight levels are urgently needed."

Another point to support our position is that the contrail responds to ambient humidity: its ice crystals sublimate in subsaturated air, while persistence requires ice supersaturation (Gierens et al., 2020). A classifier reading the footage estimates the ice supersaturation state along the flight path, so a fleet with cameras properly placed onboard a commercial fleet could give us the flight level observations mentioned earlier, for forecast verification, better prediction and contrail avoidance (Geraedts et al, 2024 and Sonabend-W et al, 2024).

We also should emphasize the need for observational data. Meijer 2026 reported that with in-situ measurements the SAC explains 98.3 % of observed contrails when using temperature and relative humidity measured by IAGOS versus 92.1 % with ERA5. From Low's work in (2025) using ground-based observations over London, they showed that only 75 % of contrails with observed lifetimes above 10 min had formed in ice supersaturated conditions according to the ERA5 RHi, meaning a quarter persisted even though the reanalysis indicated no ice supersaturation.

We also wanted to make the point that existing ways of obtaining flight-level observations using equipment such as research hygrometers, onboard lidar etc do not scale to the fleet. Having in-situ

humidity instrumentation equips only a small number of aircraft, and onboard lidar faces the certification, economic, and operational barriers we noted in section 1.1. A rear-facing camera is, by comparison, inexpensive, light, and deployable at fleet level.

---

2. The manuscript defines "short-lived contrail" and "persistent" categories not as observed in the imagery but via proxy of chase-plane measured ambient conditions of temperature, SAc and RHi > 100%. This definition glosses over the current open research question about agreement levels between SAc/measured RHi and _observed_ contrails (most recently embodied by Low et al 2025 that is cited elsewhere in this manuscript and https://egusphere.copernicus.org/preprints/2026/egusphere-2026-1171/). Authors do not attempt to visually verify in the imagery (even in a sample) the agreement level between the chase plane SAC/RHi measurements and the camera observations of contrail formation/persistence, and this represents a missed opportunity to improve scientific significance with this manuscript.

We agree that the discrepancy between SAc/measured RHi and observed contrails should be discussed and will fit well in the introduction of the manuscript. Given the ongoing research efforts on capturing their agreement levels, a preliminary qualitative study was conducted by superimposing the results on the aircraft footage for a limited sample across three days. This was omitted in the paper due to the small sample size and software's early development stage at the time. We recognize that the paper would benefit from the inclusion of a visual verification study with the current model across different video segments and conditions.

We are in the process of adding a stratified random sample of frames (balanced across the three classes and across flights) from each campaign, and annotators independently labeled each frame for 1) visible contrail presence and 2) visible dissipation of the contrail within the frame. We will report the agreement between these visual labels and thermodynamic labels as confusion matrices and compare the findings to current literature, such as the reviewer suggested.

---

3. The definition of "persistent" being only RHi > 100% is misleadingly different from most other recent papers in the field which define persistent contrails as confirmed to be existing for at least several minutes (typically 10+ minutes). The vast majority of contrails formed are short-lived in the sense of having little to no climate impact because they exist only as linear contrails for seconds or only a few minutes, not persisting long enough to spread into contrail cirrus and have a radiative forcing effect on a large area. Even for a rear-pointing camera with good viewing angle and otherwise clear conditions, my napkin math indicates it's very unlikely the images could contain evidence of the aircraft's contrail persisting/not beyond 60 seconds from formation, when the aircraft is traveling at cruising speeds. Many recent papers have asserted the case for prioritizing

forecasting the climate-relevant "big hit" contrails (very long-lived contrails, sometimes 12+ hours), and the burden of proof is on the authors to justify the claim that camera observations of SAc/RHi for at most 60 seconds would be useful in forecasting climate-relevant persistent contrail formation.

The field of view and the reliance on SAC/RHi for training are indeed limitations of this methodology. Those will be made more clear in the manuscript along with additional assumptions (e.g. simplification of overall propulsion efficiency) that affect the accuracy of the results. As discussed in the first comment's response, the phasing is being corrected to suggest the use of the proposed method as additional inexpensive real-time information on probable contrail formation that may be used with other methods of forecasting and validation.

---

4. Potentially important dataset details not reported or analyzed:

- At which times was the chase plane inside the contrail/jet-vortex of the lead plane (measuring RHi that is strongly influenced by the jet exhaust), and when was it outside the contrail/vortex (measuring purely ambient conditions)?
- Which specific humidity data streams from the campaigns were used by the authors, whether they suffer from thermal lag, and if so how that might affect the label noise in the Authors' dataset?

Thanks for the suggestion. We are looking into these dataset details and will report back once we analyze these aspects.

---

5. Claiming causality when reporting correlations:

- The authors report the reason for why F1 scores vary across the categories and campaigns is due to camera placement/orientation issues, and it may be so; but no error bars are estimated, and no evidence or statistical/causal analysis is presented to justify this particular cause compared with any other potential cause (such as spurious correlations in the datasets, or propulsion efficiency of one campaign's aircraft being closer to the modeled value of 0.33 compared to the other campaign's aircraft, etc).

At this point in time, the limited test flights and camera placements steered us to a qualitative recommendation of camera placements, as opposed to a statistical/causal analysis. The phrasing could be improved by highlighting other possible causes, as it was not our intention to imply that camera placement is the definitive or sole cause. As in responses to comments 3 and 4, other potential causes will be specified, including differences in RHi measurements, the simplification

of propulsion efficiency, the video durations and number of flight test days, and other possible erroneous correlations in the datasets.

---

6. Machine learning best practices not applied:

- The ML model (autoencoder pretrained on the images, then separately PCA on latent vector followed by 2-layer fully-connected network) is clearly described, but the description is overly verbose for how routine this type of single-image classification task has become, while at the same time failing to justify its choices. Fine-tuning a deep CNN that was trained on ImageNet so it can be domain-adapted to a relatively small classification dataset like this one has been a baseline tutorial-level ML technique for over a decade now (https://research.google/blog/train-your-own-image-classifier-with-inception-in-tensorflow/) and more recently agent-driven optimization of model performance metrics has become possible with open libraries like https://github.com/karpathy/autoresearch and https://github.com/google-research/era ; if these were suitably used the reader would have some reasonable confidence that performance metrics were being maximized among at least a few experimentally attempted alternatives. And again, the more scientifically significant finding would be directly related to the end goal of contrail forecasting or weather re-analysis improvements, for which scores on this classification task are an (as-yet unvalidated) proxy metric.
- It's not fully specified in the manuscript how video frames are assigned to cross validation folds. If they were selected randomly for k-fold cross validation, it can potentially cause over-reporting of metrics. This is because adjacent frames can be extremely similar to each other but end up spanning the cross validation boundary (effectively training on the test set). Using frames restricted to one flight-leg in the campaign as one fold in the k-fold scheme would yield more representative metrics of how the model will perform when applied to new imagery from new flights in the future. See for example https://nsojournals.onlinelibrary.wiley.com/doi/10.1111/ecog.02881

We appreciate the feedback on the ML practices and will refrain from discussing these techniques in such detail. The addition of visual verification should hopefully provide more insight into the model's capability level for contrail forecasting. We have noted the random k-fold cross-validation caution, and we are dividing frames in the folds by flight leg and roll angle ranges to correct the model's performance metrics.

7. Citation support for claims is frequently low quality, a few selected examples are below (but, I strongly recommend all citations need to be audited manually by the human authors of this manuscript):

- Line 56: citation needed
- Line 76: Akhtar Martinez et al is an inappropriate citation for this claim about open-source availability. The "open-source availability" that actually democratized CoCiP and led to its recent "broad adoption" is Shapiro et al 2023 ("pycontrails: Python library for modeling aviation climate impacts").
- Line 110: the paper describing "integrating satellite imagery with meteorological and flight trajectory" is actually https://www.nature.com/articles/s44172-024-00329-7#Sec2
- Table 2: The overall propulsion efficiency (eta) parameter is set to 0.33 for both campaigns' data, as shown in Table 2 without citation; the aircraft in these two campaigns may have a different propulsion efficiency.
- Line 238: Citing Chatterjee and Choudhury 2025 is a strange choice. It would be better to cite the foundational reference for PCA (first reported in the year 1901) or a reference that provides specific precedence/justification for the choice of pretraining an autoencoder, applying PCA, then training another separate classification network on top of the PCA components.
- Line 275: citation needed (CoCiP *does* use SAC…)
- Line 340: 50% accuracy claim is not supported by this citation. There is indeed an erroneous claim made in that IATA report that "Recent trials were successful in identifying the detected cloud as an aircraft-generated cloud 50% of the time when using an artificial intelligence (AI) algorithm, and 80% of the time when the matching was done or assisted by a person [17]" but citation 17 in that IATA report is Geraedts et al 2023, and in Geraedts et al 2023 the actually peer-reviewed scientific claim is "Only 50% of the flights labeled as matches by the automated matching were labeled as matches by the humans" which is a claim about *contrail-to-flight attribution* and NOT about contrail *detection*. Ng et al 2023 (that this work already cites, but elsewhere) is one option for a suitable reference for accuracy of contrail detection algorithms on satellite imagery.

We have additionally performed a full manual audit of every citation in the manuscript, as recommended. Regarding the citations that were flagged by the reviewer:

- Line 56: We added the citation by Teoh (2022)
- Line 76: The open-source implementation of CoCiP that underlies its recent broad adoption is pycontrails (Shapiro et al., 2023, doi:10.5281/zenodo.7776686); We had included it in the very first draft but chose to remove it since it was the Github repo and not a paper. We

now cite it there. We had added the Akhtar Martínez et al. (2025) citation to show an example of pycontrails use but now we are keeping it only where its content (limitations of zero-dimensional contrail models) is actually relevant. We added the Engberg (2025, https://gmd.copernicus.org/articles/18/253/2025/) reference instead.

- Line 110: Included the Sonabend-W (https://www.nature.com/articles/s44172-024-00329-7#Sec2) reference for the integration of satellite imagery with meteorological and flight trajectory.
- Table 2: Bräuer et al. calculates η=0.31 for the ND-MAX campaign (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020GL092166), while Dischl et all use 0.33 in their model for any aircraft that appear in the region https://www.mdpi.com/2226-4310/9/9/485 . For this study, 0.33 was used for both aircraft similarly to Dischl, as stated in Lines 265-268. A distinct value for the B737 can, however, be substituted, though it will still remain an approximation.
- Line 238: Cited Pearson (1901) for PCA. We will also improve our explanation on the revised paper.
- Line 275: Our wording here was wrong and contradicting our earlier statement in line 216. We are citing the 2012 paper by Schumann.
- Line 340: We are removing this citation. When citing the IATA report, there was also confusion among the authors what the report meant.