Referee #1

The paper introduces a method to combine satellite-based contrail observations with weather predictions and in-situ measurements for assessing contrail related weather predictions. The paper demonstrates that this approach improves the agreement between reanalysis data and in-situ measurement, and, hence, improves the framework for evaluating avoidance strategies.

The paper is rather straightforward and can be published soon, after clarifying or correcting some text issues, including the technical corrections identified by the authors in their comment themselves.

The only general critics I want to see reflected by the authors is the following. The method described requires measurements and reanalysis data which are not available quickly after the event considered. What can be done to reduce the time interval between the event to be assessed and the completion of the data analysis? Ideally, a pilot wants to know how well he avoided contrail formation immediately after the flight.

This is an interesting question. The data required for the technique are:

- Weather data. Though reanalysis weather is ideal, it is not often not available for a long time. But Fig. 5b shows that forecast weather also works reasonably well and it is available even before the flight.
- Satellite observations: these are typically available around 10 minutes after the images are taken
- Flight trajectories (from ADSB, need for the contrail-flight attribution): available in near-real-time

Therefore it should in principle be possible to produce this data analysis quickly enough that the pilot can see the result soon after landing. In practice setting up a system to proactively compute as soon as the data is available is a more complicated software engineering problem than just waiting a while until the data is certainly available. For this reason the existing API at https://developers.google.com/contrails/v1/ContrailWatch-description takes a few days between the flight and the data analysis being available.

We have added the following text to the paper: “Computing the `hybrid' prediction requires satellite imagery, flight ADS-B data and weather data. The satellite imagery is available around 10 minutes after the image is taken, and the flight ADS-B data are available in near-real time. The good performance of the forecast weather in Fig. 5b means that in principle the results of the hybrid method could be made available almost immediately after

the flight time, allowing immediate feedback for aircraft operators on whether contrails were formed."

Line 26 -> and so cannot see every flight

changed

Fig 1: the yellow region is hard to see.

and

How can I see that 'Ensemble 0' matches the observed contrail better?

Changed "Our method applies a higher weight to 'Ensemble 0' because it better matches the observed contrail" to "Our method applies a higher weight to `Ensemble 0' because the location of the contrail formation region it predicts better matches the observed contrail"

Line 66: "is the prior" : I miss a subject behind prior.

To explain better for those less familiar with Bayesian statistics, replaced "is the prior" with "is the 'prior probability' (the initial guess at the probability that a persistent contrail will form)"

Line 101: Please explain the 'histogram matching' method.

Added the text "which recalibrates the forecast humidity to better match in-situ observations" to help explain, and also (as we should have done in the beginning) cited the work which introduced the concept.

Line 75: I cannot believe that IFS relies only in ensembles which use the 'pressure-level' data at the rather coarse resolution of 25*hPa.*

My understanding is that the simulations are run at the higher resolution but downsampled on output. We use the downsampled data because downloading historical ensemble forecasts from ECMWF is quite bandwidth limited. Edited to the text to make it clearer that the low resolution is due to constraints on our ability to download the forecast, not the forecast itself.

The two references Driver (2025a,and b) are actually referring to the same paper.

Fixed, thanks for catching.

Referee #2

The paper combines satellite contrail observations with weather ensembles and IAGOS in-situ measurements to assess contrail-formation predictions, and shows the hybrid score agrees better with the in-situ data than weather alone. The approach is sound and the paper is clearly written. I think it can be accepted after a few clarifications.

I also think the broader direction is the interesting part: using contrail detections to inform the weather side, rather than only using weather to predict contrails. That seems worth pursuing well beyond this paper.

Most of my comments are about how in-situ measurements, flights and satellite detections are matched.

First, something I could not reconcile. The chosen distance threshold is 50 km in Eq. 2, but the "Base case" row of Table 4 says 15 km (and 50 km then appears in the separate "Distance" row). These can't both be the configuration used everywhere. Since this threshold defines the neighbourhood that produces the observed/not-observed labels feeding Eq. 1, it matters which one was actually used. Could you confirm the deployed value and make Eq. 2 and Table 4 agree?

The correct value is 15 km, thanks for catching, we have updated Eq. 2. Every time we slightly change the dataset (such as by fixing the index error mentioned in a previous comment) which set of thresholds is ‘best’ slightly changes. This doesn’t really matter (as Fig 7 shows, the results are more or less the same across a number of different thresholds). While updating the manuscript we forgot to update one of the values in Eq. 2.

On the altitude tolerance: 20 m is at the low end of the searched range (10–300 m). The weather is spaced ~10–25 hPa in the vertical, a few hundred metres near cruise, and supersaturated layers are usually deeper than that, so two waypoints 20 m apart sit in the same grid cell and the same layer. What is the 20 m meant to capture? It may well be the conservative choice given the "any nearby contrail" rule, and Table 4 includes a 50 m variant that looks similar, so I am not asking you to change it — just to say in one line why 20 m.

We didn’t choose 20 m explicitly, we chose a point that looked good in Fig. 3 and then used the thresholds that generated that point. As to why the points on the low end were some of

the better performers, we don't know definitively. There's a tradeoff between higher thresholds, which give you more nearby observations, and lower thresholds, which make the observations you do have higher quality. It seems that for altitude especially that tradeoff favors low thresholds. We guess it's because the supersaturated layers are vertically thin so even at a 50 m threshold, some fraction of the time the insitu measurement is inside a supersaturated layer but the flight 50 m above is not. Added the following sentence to the paper

"Fig. 7(a) shows the largest sensitivity to altitude, potentially because the supersaturated layers that drive contrail formation are vertically thin."

The satellite sees the contrail 20–30 min after it forms, so detections are advected back to the flight with ERA5 winds. The advection removes the bulk transport. What is left is the wind error over those 20–30 min, of order a few km for a ~1–3 m/s wind error, so small against the matching distance. But it is spatially correlated, not random noise, and it feeds the observed/not-observed label. A line on its likely size, and on whether the synthetic TPR/FPR calibration already accounts for it, would help.

Added in Sec 2.2: "Note that the advection is uncertain because the wind data used is imperfect and contains errors of around 12.4 km/h (Sonabend 2025) the attribution algorithm accounts for this when it sets the tolerances of what flights can match an observed contrail."

And in Sec 2.5: "Also note that the synthetic data is generated using different wind data than that used during flight attributions, so measurements of the overall system performance include the effects of imperfect wind data."

On §3.1: IFS differs from the ERA5 ensemble in resolution and member count as well as in being a forecast, and you already attribute the mixed result to resolution. So I would just say explicitly that reanalysis vs forecast can't be cleanly separated here. S_nominal(coarse) already isolates resolution and is the obvious place to start for a fuller comparison.

Changed "We suspect this is again due to the relatively low resolution of the ERA5 ensemble data." to "However since the forecast and reanalysis weather have different resolutions we cannot clearly separate the effects of just changing from reanalysis to forecast data. "

One more, minor: the agreement results (~90k IAGOS points, 2019 and 2024) and the energy-forcing results (~430k flights, one day per month, June 2024–May 2025) use

two different datasets over different periods. Worth saying so plainly so the reader doesn't read it as one validation set.