Review 1

This paper presents the first remote-sensing-based contrail altitude estimation algorithm. Both the image-level model and the cirrus pixel-by-pixel model are developed and compared, with an evaluation of predictive uncertainty and an assessment of the method's accuracy using individual test data and independent flight data. This study offers valuable insights for further assessing the climate impact of contrail cirrus. The paper is well-organized and well-written. I urge its publication in AMT, with some minor comments provided for the authors' consideration.

We thank the reviewer for their positive evaluation of the paper, and the several specific comments that helped improve the manuscript. The responses to the specific comments are shown below.

Specific comments:

Line 40: Please provide the physical explanations for why the infrared channels are used for estimating cloud top altitude.

We have added the sentence “Fundamentally, these retrieval algorithms utilize the fact that the infrared radiance observed by the satellite instrument is a combination of that emitted by the surface, atmosphere and the cloud itself (Liou, 2002).”.

Figure 7: The plot shows a trend where the CNN generally overestimates contrail altitude compared to the true values from CALIPSO. Are there any potential ideas for this?

We have computed the bias, defined as

$$\text{bias} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i),$$

where n is the number of data points, $\hat{y}_i$ is the mean of the probability distribution output by the CNN for data point i and y_i is the CALIOP value for data point i.

Although the subset of data shown in Figure 7 does have a positive bias of 2.86 flight levels (87 meters), an analysis on the full dataset shows no statistically-significant bias at the $p<0.05$ level. This suggests that the bias shown in Figure 7 is a result of the specific circumstances of those contrails, rather than a systematic bias in the algorithm.

Figure 10: The plot here seems to support my impression from Figure 7 that the contrail altitude can be slightly overestimated. During data collocation, you carefully

considered the advection of aircraft data due to horizontal wind. Then, contrail ice crystals can sediment, which should theoretically reduce the altitude rather than increase it when compared to the flight data. Are there any reasons behind this discrepancy?

This is an interesting point. We think that any discrepancy between the estimated contrail top altitude (in flight levels) and selected "closest" flight can be due to a combination of the following:

- Estimation error by the algorithm
- Conversion error from geometric altitude to pressure altitude, owing to errors in the used geopotential data
- Incorrectly choosing the "closest" flight (i.e. we do not compare the estimated altitude with the flight that actually formed the given contrail)
- Dynamical effects: contrail sinking + deepening due to the aircraft wake, formation of the secondary wake, buoyancy effects, radiative heating/cooling, gravitational settling of contrail ice particles, vertical winds and more.

Firstly, we do indeed find that the bias (i.e. the mean of the blue distribution shown in Figure 10) is non-zero (and positive) at a statistical significance level of $p = 0.01$. The analysis presented in response to the previous comment showed that it cannot be concluded at a statistically significant level (with any p-value lower than 0.29) that the model has a bias. However, this analysis pertains to the entirety of the test set, which contains data points that were randomly picked from all available data spanning the years 2018 to 2022. There does exist the possibility that particular circumstances (combination of season and synoptic conditions) lead to a positive or negative bias in the algorithm's estimates when averaged over a 24 hour period. However, the test data stratified by season does not indicate any seasonal biases at a statistically significant level. Assessment of the impact of specific synoptic conditions on model performance would require more analysis.

The other possible sources of discrepancies between the "closest flight" altitude and the estimated contrail top altitude may also lead to biases of either sign. The reviewer is indeed correct in noting that gravitational settling of ice particles, which is not included in the advection of the flight data, would lead to lower "advected" flight altitudes and would therefore increase the positive bias observed here.

Summarizing, there is indeed a bias present in Figure 10. However, the objective of the analysis presented therein is to show an agreement between the CNN estimates and flight altitudes that could potentially have formed the contrail. Given the lack of ground truth data on which flight formed the contrails whose altitudes are estimated that day, it is not

possible to assess the relative importance of possible contributions (i.e. CNN error, geopotential error, etc.) to the bias. This also drove the decision to perform most of the quantification of the altitude estimation algorithm's performance with CALIOP data, rather than using results from existing flight-to-contrail matching approaches.

We have added the following sentences to the manuscript to reflect the above analysis:

"The altitude estimates by the CNN - when compared to the altitude of the closest flight - do show a positive bias of 2.4 flight levels (statistically significant as determined using a one- sample T-test at p = 0.01). The evaluation with CALIOP test data indicates no statistically significant bias (at p = 0.01) for the CNN estimates, however. Potential other causes of the bias observed in Figure 10 may be the use of geopotential data for the conversion between geometric- and pressure altitudes, the methodology used for constructing the flight altitude distributions, as well as the omission of contrail physics in the advection process."

Conclusion: The RMSE is used as the metric to indicate the accuracy of the algorithm, as emphasized in the abstract. Since the developed contrail altitude retrieval method is the next step due to the biased prediction of ice supersaturation vertical extension in contrail avoidance, would it be better to also show the simple mean bias error or mean absolute error for estimating the contrail altitude?

When comparing the RMSE and the mean absolute error (MAE), the RMSE is more sensitive to outliers than the MAE. We evaluated the MAE for the four different models and found the values to always be lower than the corresponding RMSE values. The conclusions on the relative performance of the four models is the same as when using the RMSE. The results from the mean bias error (which we have used in the response to a previous comment, but simply called it "bias") for the CNN have been discussed already. For the other three models, the mean bias error is found to not be zero (with a maximum mean bias error of 540 meters for the Cirrus MLP). Given the role of the RMSE in the evaluation of the probabilistic component of the CNN (to construct the predictive model with constant uncertainty) and the identical conclusions achieved when using the MAE, we choose to leave the latter metric outside of our consideration.

Technical corrections:

Caption of Figure 1: "Zulu" time is equivalent to "UTC" time. However, I'm not sure if it is widely used in this research field. This applies to the entire text to be consistent with the figure.

We have reviewed the submission guidelines set by the journal and have found that the correct way to indicate this is indeed by use of the "UTC" abbreviation. We have modified this throughout the manuscript.

L90: "a 50km distance of the ground-track of CALIPSO." I assume it should refer to the supplement S1.

We agree that it is helps the reader to refer to the supplementary materials here, in particular to the section mentioned by the reviewer. Thank you! We have added the following to this sentence: *"(see section S1 of the Supplementary Materials for more details)."*

L132: "FlightAware (for times in 2023)". Eventually it appears not to have been used because the focus was on the years 2018-2022.

The FlightAware data has been used in the analysis discussed in Section 4 of the manuscript, as this concerns a day of data analyzed in 2023. However, we realize now that the description of where the two different data sources are used was lacking, so we have added the following sentence to the manuscript.

"The OpenSky data is used for comparison with the contrail altitude estimation performance on the test set in section 3, whereas the FlightAware data is used for analyzing contrail detections and altitude estimates for a full day of data in section 4."

L221: "ISS" instead of "ISSRs".

We think that the use of both ISS and ISSRs is possible here.

L273: tends to be over-confident for probabilities between 0.5 and 0.9, as well as between 0.1 and 0.2.

We have updated the manuscript to more accurately reflect the results shown in Figure 5. We thank the reviewer for noticing this error.

"The CNN tends to be overconfident for most predicted probabilities, with largest deviations occurring for predictions between probability 0.4 and 0.9. For example, when the CNN predicts that 60% of the contrails should be below a particular altitude, Figure 5 indicates that only 50% of contrails will actually be found below this altitude."

Overall, the excellent work presented in this article is acknowledged.

We again thank the reviewer for their positive evaluation of our manuscript.