Comments on "Predicting Aviation Contrail Occurrence Using Bayesian Population Statistics From Reanalysis Data"
by Daniel A. Williams, Cyril J. Morcrette, and James M. Haywood

The objective of this study is to develop a method for contrail prediction based on population-level statistics and Bayesian methods. The manuscript combines manually labelled contrail observations from the OpenContrails dataset, which contains satellite imagery from the GOES-16 Advanced Baseline Imager, with ERA-5 reanalysis data to learn the meteorological conditions associated with contrail formation. The authors use Bayesian methods to construct a predictive tool for contrail occurrence at given humidity and pressure heights, based on the conditions associated with previously confirmed contrails. The approach aims to provide a scalable and interpretable alternative to existing highly parameterised models or poorly explainable machine learning approaches.

The authors evaluate the performance of the prediction tool using binary classification metrics. Furthermore, the comparison between predicted contrail probability maps and satellite observations demonstrates that the approach can reproduce regions with enhanced contrail formation likelihood. The authors suggest that this method could support future contrail avoidance strategies by providing an explainable and scalable prediction framework.

I have several comments. After addressing these points, the manuscript may be suitable for publication.

Comments:

- L. 3/4: "Contrails and contrail induced cirrus clouds contribute an estimated 57% to the sector's total effective radiative forcing (ERF)." → The authors may consider briefly mentioning the uncertainty associated with this estimate, as the contribution of contrails and contrail-induced cirrus clouds to aviation ERF remains subject to considerable uncertainty.

- L. 32: "At present, the majority of aviation ERF comes from non-CO2 effects" → However, it is worth pointing out the greater uncertainties associated with non-CO_2 effects compared to CO_2 forcing.

- L. 48/49/50: "During daytime, the two effects are opposite and broadly comparable, while at night contrails provide a strong warming effect (Penner et al., 1999; Digby et al., 2021; Quaas et al., 2021; Ortiz et al., 2024), an order of magnitude higher than that due to CO2." → The authors should clarify whether this comparison refers to instantaneous radiative forcing or to the integrated climate impact over different timescales.

- L. 90/91/92: "ML methods can overcome this barrier since a model can 'learn' any required parametrisations and these can be adjusted in the light of new data. They however often suffer from the 'black-box' problem, which industry partners and policymakers may be more reluctant to make use of if decisions cannot be adequately justified." → Which specific aspects of the lack of interpretability in ML models pose a problem for stakeholders? Does this mainly concern the traceability of individual model

decisions, the quantification of uncertainties, or regulatory requirements for predictive models?

- L. 110/111: “Images included in the dataset are pre-filtered according to criteria set out in Ng et al. (2024), ensuring that each tagged contrail had an associated real flight path.” → Could the pre-filtering criteria introduce a systematic bias in the OpenContrails dataset, for example by favouring certain flight regions, seasons, or meteorological conditions where contrails are easier to identify and associate with flight paths?

- L. 130/131: “Use the DBSCAN (Ester et al., 1996) clustering algorithm on all skeletonised vectors in an image to create a set of aggregated contrails, with the requirement that they are labelled by multiple taggers, as demonstrated in Figure 1.” → How were the DBSCAN hyperparameters (e.g., the maximum distance between points to be considered neighbours and the minimum number of points required to form a cluster) selected, and how sensitive are the extracted contrail aggregates to these parameter choices?

- L. 132: “Using the endpoints (boundary) for each aggregated identified contrail, construct the great-circle path linking them” → How frequently do actual aircraft trajectories deviate significantly from the assumed great-circle path, particularly in highly congested airspaces where air traffic control constraints may lead to route deviations?

- L. 147/148/149: “When applied to over 100,000 contrail endpoints however, we not only minimise the impact of each of these assumptions, but can begin to learn population level behaviours by significantly reducing any signal-to-noise ratio.” → The statement regarding reducing the signal-to-noise ratio appears to be unclear, as reducing this ratio would normally indicate a decrease in the quality of the signal relative to noise. Did the authors intend to describe an increase in the signal-to-noise ratio or a reduction of noise relative to the extracted signal?

- L.181- 185: “Since the Cramér-von Mises test provides a measure of the difference in relative humidity distributions between the contrail-biased and stochastic samples, we can use the test statistic normalised across the pressure heights as an effective proxy for the heights at which contrails are observed. This provides the informed prior for our contrail predictor, which we formulate as a 2D look-up table using ERA-5 pressure levels and relative humidity binned into 5% increments.” → The informed prior appears to be derived from the same contrail-biased sample that is also used to estimate the likelihood. The authors should clarify how potential double-counting of information is avoided in this framework and provide further justification for interpreting the normalised Cramér–von Mises test statistic as a prior distribution.

- L. 209: The definition of the F1-score appears inconsistent. If the left-hand side of Equation 4 is used, the transformation to the right-hand side is incorrect. The result would be: TP/(2TP + FP + FN). The standard definition of the F1-score as the harmonic mean of precision and recall gives F1 = 2 (precision * recall)/(precision + recall) = (2TP)/(2TP + FP + FN).

- L. 195–200 and 268–280: The classes used in the confusion matrix are derived from the internal and external projection points along the extrapolated flight path. To what extent could the reported values of precision, recall, and F1-score be influenced by the fact that the ground-truth labels themselves rely on assumptions inherent to the projection method (e.g., approximately linear flight trajectories and a smooth or monotonic variation of relative humidity near the contrail boundary)? The authors should discuss the sensitivity of the classification results to these assumptions and clarify whether an independent validation using actual flight trajectories or other external observations would be feasible.

- L. 320/321: “Whilst efforts have been made to consider various sources of uncertainty in the predictions, some cannot be ignored.” → Which uncertainties have the greatest impact on forecast quality, and have these been assessed quantitatively? Prioritising the sources of uncertainty (e.g. annotation, ERA-5 resolution, advection, model assumptions) would strengthen the conclusions.