Journal of Open Aviation Science (2026), Vol.**4**
doi:10.59490/joas.2026.8684

JOAS

EDITORIAL

Reviews and Responses for Generative Short-Term Aircraft Trajectory Prediction with Conditional Flow Matching

Authors: Benoit Figuet, Timothé Krauth, and Steve Barry

Reviewers: Gabriel Jarry, Richard Alligier, and Max Li

Editor: Xavier Olive

1. Original paper

The DOI for the original paper is https://doi.org/10.59490/joas.2026.8468

2. Review - round 1

2.1 Reviewer 1

This paper presents a technically robust and well-illustrated framework for probabilistic aircraft trajectory prediction using Conditional Flow Matching (CFM). The methodology is sound and well motivated, leveraging historical ADS-B data from the OpenSky Network to transform random noise into realistic future trajectories. The use of Conditional Flow Matching is appropriate in this context, as it enables stable regression-based training and efficient sampling while offering improved uncertainty calibration compared to adversarial or diffusion-based generative models. The paper is clearly structured, and the results convincingly demonstrate how generative modeling can move beyond binary collision alerts toward continuous, data-driven risk measures relevant for Air Traffic Management (ATM).

Minor Questions and Suggestions for the Authors

1. **Benchmarking against physics-based models.** The current evaluation relies on a Constant-Velocity (CV) baseline, which is a standard but relatively weak reference. Have the authors considered comparing the proposed CFM approach against a more advanced physics-based baseline? For instance, a BADA-based simulation incorporating a distribution of plausible intents (e.g. standard procedural turns or altitude changes) could provide a more competitive benchmark and help clarify the relative strengths of data-driven generative models versus traditional kinetic approaches.
2. **Physical consistency, flyability, and oscillations.** The discussion acknowledges that a small fraction of generated trajectories exhibit unrealistic oscillations or curvature. Beyond kinematic regularization, how do the authors envision systematically validating that generated trajectories are “flyable” with respect to aircraft performance envelopes? Future work could explore the integration of lightweight physical constraints (e.g. bounded curvature, acceleration limits) or physics-informed learning strategies to ensure that the learned vector field remains consistent with basic aerodynamic limits while preserving trajectory diversity.

3. **Probabilistic calibration and horizontal over-dispersion.** The PIT analysis indicates that the model is over-dispersed in the horizontal (x, y) components, with ensemble spreads exceeding the empirical variability observed in the test data. Could the authors comment on the potential operational implications of this behavior? In practical ATM settings, excessive lateral over-dispersion may translate into overly large uncertainty volumes, potentially increasing controller workload or the rate of nuisance alerts despite the probabilistic formulation.
4. **Statistical robustness of risk estimates.** In the real-world conflict case study, mid-air collision probabilities are estimated using an ensemble of 100 sampled futures per aircraft. Given the rarity of such events and the strict collision proxy employed, is this ensemble size sufficient to ensure stable and converged probability estimates? A brief discussion on the sensitivity of the risk metrics to the ensemble size and on Monte Carlo convergence behavior would strengthen the interpretation of these results.
5. **Scope of ATM applications and model decomposition.** While the paper focuses primarily on deconfliction and mid-air collision proxies, the proposed framework appears broadly applicable. Could the authors elaborate on how the model might be adapted to other ATM use cases, such as high-fidelity ATC simulation or the detection of anomalous flight behaviors? More generally, have the authors considered a hierarchical decomposition in which CFM models high-level intent (the “where” and “what”), while a separate Neural ODE or kinematic model generates the high-fidelity trajectory execution (the “how”)? Such a separation could improve both physical consistency and interpretability for human operators.

2.2 Reviewer 2

Review text

2.3 Reviewer 3

The application of generative modeling to short-term aircraft trajectory prediction is well motivated and highly relevant to both safety and operational risk assessment in ATM. The overall modeling framework is promising, and the use of Conditional Flow Matching for uncertainty-aware forecasting is technically appealing. However, several aspects of the methodology, particularly the construction of the Gaussian and Optimal Transport paths, the interpretation of the probabilistic flow, and the numerical realization of the sampling process, are explained quite cursorily. As a result, it was difficult for me to fully build intuition for the CFM mechanism, especially for those not already deeply familiar with flow-based generative models. Providing additional conceptual explanation, clearer notation, and stronger benchmarking against contemporary learning-based baselines would significantly strengthen the paper’s technical clarity and empirical impact.

1. While the presentation up through Section 2.3 is generally clear and well structured, the transition into the Gaussian and Optimal Transport (OT) path construction would benefit from substantially more intuition. In particular, the introduction of the Gaussian path at the boundary between lines 97 and 98 is quite abrupt. It is not clear how the time-varying mean and variance are selected beyond satisfying the stated boundary conditions, nor what physical or probabilistic interpretation should be attached to these choices in the context of trajectory prediction. Additionally, the discussion of the boundary conditions themselves is somewhat confusing, as it is not immediately apparent what “ends” of the distributional path are being examined in operational terms. Overall, Section 2.4 would benefit from a more careful, intuition-driven explanation that connects these constructions explicitly to the underlying generative modeling objective rather than presenting them primarily as formal mathematical machinery.
2. Several instances of mathematical notation would benefit from earlier and more explicit clarification. For example, when norms are introduced in the loss function, it is implicitly assumed that these correspond to l_2 norms; although this is later defined around lines 193–194, a clearer

and earlier definition would improve readability. Similarly, some numerical techniques are used without adequate explanation or citation. A specific example is the use of Heun's method for ODE integration, which is referenced but not briefly characterized nor supported with a citation. Given the centrality of numerical integration to the sampling process, additional clarification here would meaningfully improve transparency and reproducibility.

3. I am somewhat unclear as to what the primary benchmark or point of comparison is for evaluating the proposed generative method. While constant-velocity extrapolation is included as a baseline, this represents a very weak comparison, from my perspective, relative to other existing data-driven trajectory prediction models. A more informative and fair comparison would include at least one modern learning-based predictor, such as a CNN–LSTM, VAE-based model, or alternative diffusion-style architecture. Without such a comparison, it is difficult to fully assess the relative strengths and weaknesses of the proposed Conditional Flow Matching approach within the broader landscape of generative trajectory prediction methods.

3. Response - round 1

3.1 Response to reviewer 1

We thank the reviewer for the positive assessment and constructive suggestions. Below we address each point.

Response

1. **Benchmarking against physics-based models.** We agree this would be valuable but also make the comparison limited due to the limits of physics based modeling without knowing the intent of the flight trajectory. In this proceedings submission, we keep the baseline as constant-velocity extrapolation because it matches the assumptions commonly used in short-term safety nets and in practical risk modeling. We now clarify this explicitly in the Discussion.
2. **Physical consistency, flyability, oscillations.** We agree and have expanded the Discussion to position systematic flyability validation (bounded curvature/acceleration and performance-envelope constraints) as a key future direction.
3. **Operational implications of horizontal over-dispersion.** We added a short discussion note: over-dispersion may inflate uncertainty volumes and increase nuisance alerts, motivating further calibration work.
4. **Statistical robustness / Monte Carlo convergence.** We added a Clopper–Pearson confidence interval for the mid-air collision probability in the conflict case study (based on 10,000 paired samples).
5. **Scope of ATM applications / model decomposition.** We agree the approach is broadly applicable (e.g. offline encounter triage, scenario generation, anomaly detection). A hierarchical decomposition (intent vs execution) is an interesting direction and we mention it as future work.

3.2 Response to reviewer 2

We thank the reviewer for the careful reading and detailed, actionable suggestions. We address the points below.

Response

1. $\sigma_{\min}$ **disappearing.** We clarified in the Background that while the general Gaussian/OT path can use $\sigma_{\min} > 0$, our implementation uses the common simplifying choice $\sigma_{\min} = 0$ (deterministic endpoint).
2. **Why** 12×7 **and feature redundancy.** The 7 channels reflect a kinematic state derived from ADS–B (position, velocity components, vertical rate, and a turn-rate proxy). While some quantities are

correlated, we found this representation stable and practically convenient for learning both local dynamics and uncertainty; we did not run an ablation study on reduced feature sets in this submission and leave it to future work.

3. **Number of forecast steps.** This is an interesting point. We initially did the opposite and predicted with 1 s resolution, but for practical reasons (primarily training time and clearer visualization in the paper), we switched to 5 s ($K = 12$).
4. **PIT / errors in aircraft-centric coordinates** $(\tilde{x}, \tilde{y})$**.** We agree such diagnostics are insightful. In this revision we have implemented the proposed change for the PIT histogram (Figure 7) but kept the original for the error plot (Figure 6) as we think that it makes it easier to quickly grasp the amount of horizontal error we are obtaining.
5. **Figure 5: meaning of sample vectors / vector field.** We updated the manuscript text for Figure 5 to clarify that the orange sample vectors are flow-time displacements between consecutive integration steps (not physical velocities), and that the purple grid field is obtained by querying $v_\theta(\cdot, t \mid H, c)$ on a spatial grid for a single synthetic token, providing a qualitative 2D slice of the full 12×7 field.
6. **Orange vectors not tangent.** This is expected under the above interpretation; we clarified this explicitly in the manuscript.
7. **“512 test trajectories” ambiguity.** We updated terminology throughout to distinguish *test windows / prediction problems* (N) from *stochastic forecast samples* per window (S), and corrected minor notation inconsistencies (n vs S, t vs τ).
8. **Derivative notation and minor typos.** We clarified that primes denote derivatives with respect to flow time t (holding x_1 fixed).
9. **Improvement of section 3.3.** We implemented the proposed improvements in the text.
10. **Improvement of section 4.** We fixed the notations issues and corrected the “60,s” typo.

3.3 Response to reviewer 3

We thank the reviewer for highlighting clarity gaps. We have made targeted edits to improve intuition, notation, and numerical transparency while keeping the experimental scope unchanged.

Response

1. **More intuition for Gaussian/OT path.** We expanded Section 2.4 with a short intuitive explanation of the endpoints (noise at t=0 to a distribution concentrated near x_1 at t=1) and the roles of μ_t (drift) and σ_t (uncertainty removal).
2. **Notation (norms) and Heun’s method.** We now define $\|\cdot\|$ (Euclidean norm) at its first use in the FM objective, and we briefly characterize Heun’s method as an explicit trapezoidal / second-order Runge–Kutta predictor–corrector integrator in the inference section. We also clarified derivative notation (primes w.r.t. flow time t).
3. **Stronger baselines.** We agree that comparisons against modern learning-based predictors would be valuable. For this proceedings paper we retain a constant-velocity baseline (aligned with operational short-term alerting/risk models) and explicitly position broader benchmarking as future work.