Journal of Open Aviation Science (2026), Vol.**4**
doi:10.59490/joas.2026.8813

JOAS

EDITORIAL

Reviews and Responses for Review of ADS-B Data Usage with the Focus on Data Cleaning

Authors: Ruolan Ren, Jingcheng Zhong, Dizhi Guo, Ruixin Wang, and Christophe Hurter

Reviewers: Ramon Dalmau and Enrico Spinielli

Editor: Martin Strohmeier

1. Original paper

The DOI for the original paper is https://doi.org/10.59490/joas.2026.8467.

2. Review - round 1

2.1 Reviewer 1

This paper addresses a relevant problem in aviation data by examining how data cleaning procedures affect downstream algorithm performance. The comprehensive literature review categorizing ADS-B applications is useful, and the attempt to systematically evaluate cleaning impacts through an autoencoder case study represents a worthwhile research direction. However, the experimental validation is too limited to support the paper's broad claims about cleaning strategies, and several methodological choices limit the generalizability of the findings. The paper requires major revisions before it can make a credible contribution to the field.

Major Issues

- The experimental design suffers from significant methodological limitations that prevent generalizable conclusions. The paper claims to investigate "the relationship between data cleaning and algorithmic performance in ADS-B analytics," yet tests only a single autoencoder architecture on a single reconstruction task. The authors themselves acknowledge that "the impact of noise strongly depends on the model's structure and learning mechanism," which directly limits the extent to which general cleaning guidance can be drawn. The conclusion that "noise suppression should prioritize smoothing and interpolation optimization" cannot be supported by evaluating only one model type. Different algorithms may exhibit fundamentally different sensitivities to specific noise characteristics.
- The baseline selection methodology introduces model-dependent circularity that compromises the validity of the evaluation. Section 4.2 uses autoencoder reconstruction error to identify "clean" trajectories as the baseline, then tests the same autoencoder on corrupted versions of these trajectories. This creates a self-referential evaluation in which the model is assessed on data it already reconstructs well.
- The artificial noise injection lacks validation against real ADS-B error patterns. While Section 4.3 describes three noise types, the paper provides no quantitative evidence that the synthetic

noise parameters correspond to the statistical characteristics or magnitudes of real ADS-B errors documented in the cited literature. Without such validation, conclusions about model sensitivity to Gaussian, drift, or spike noise may not translate to real-world ADS-B data.

- Critical data quality issues discussed earlier in the paper are excluded from the experimental evaluation. Although missing data and interpolation are emphasized throughout the review sections, the authors explicitly state that the autoencoder "did not further investigate the effect of missing-value errors". This omission is significant, as interpolation to handle missing data is among the most common preprocessing steps described in the paper. The experimental framework should include missing-data scenarios and compare different interpolation strategies.

Medium Issues

- Figure 8 presents boxplots comparing reconstruction error across noise conditions, but no statistical significance testing is reported. The claim that the model "shows greater tolerance to drift and spike noise" would benefit from quantitative statistical support beyond visual inspection. In addition, "normalized RMSE" is referenced without specifying the normalization procedure in the figure caption or main text.
- Several key experimental parameters lack justification. The autoencoder uses a 32-dimensional latent representation for a 200-dimensional input, but no rationale is provided for this compression ratio. More critically, the noise injection parameters (e.g., Gaussian sigma values, drift rates, spike probabilities in Figure 8) are not justified relative to realistic ADS-B error magnitudes reported in the literature.
- The experimental scope is narrow. Using data from a single airport (Zurich) over a two-month period and selecting only 100 trajectories as the baseline raises questions about representativeness. It is unclear whether these trajectories adequately capture the diversity of aircraft types, flight phases, and operational scenarios discussed in Section 2.
- The paper correctly identifies a lack of transparency in the literature regarding data-cleaning procedures, yet its own experimental methodology is incompletely specified in the main text. Critical implementation details for the autoencoder and noise injection are deferred to a GitHub repository. Given the paper's emphasis on methodological clarity and reproducibility, these details should be fully documented within the manuscript.

Minor Issues

- The paper states that results are averaged over 10 repetitions, yet Figure 8 boxplots suggest a larger number of samples. The figure caption should clarify whether the distributions represent variation across trajectories, repetitions, or both.
- Section 3.3 provides a thoughtful discussion of the limitations of cleaning methods, but these points could be strengthened by explicitly linking them to specific studies cited earlier in the literature review.
- In Table 1, entries in the "Role of ADS-B Data" column such as "Core data source" are vague. More specific descriptions of how data quality requirements differ across application categories would better motivate why cleaning strategies may need to vary.

Recommendation: Revisions Required

2.2 Reviewer 2

Very interesting paper. I would suggest to review the cited references given the detected mismatches.

General: some acronyms are not listed, i.e. 4D (needed?), LSTM, MILP, MCNN, GMM, ML, ETA, PCA, ACAS, BADA, FDR, (RACE?), RDP, UTC (not needed probably?), MSE (Line 409)…

Good reproducibility: code available in Github repo. Code sparsely but sufficiently commented.

- Lines 11–14: the references seem to be incorrect. For example at line 12 "flight phase identification [3]" but [3] M. Schlosser, H. Brassel, H. Fricke "Analysis of Ground Trajectories…" deals with ground portions of a flight. Maybe J. Sun, J. Ellerbroek, J. M. Hoekstra "Large-Scale Flight Phase Identification from ADS-B Data Using Machine Learning Methods" (https://repository.tudelft.nl/record/uuid:af67a6bd-d812-474d-a304-7a594991390b) would be a better example. Also the other references seem ill-fitting, i.e. "airport operations optimization [6]" while [6] is E. Roosenbrand, J. Sun, J. Hoekstra. "Contrail Altitude Estimation Based on Shadows Detected in Landsat Imagery". Similarly for the others?
- Line 36: add space before [7], i.e. "risks[7]." → "risks [7]."
- Line 41: "…via on-board equipment [8]."; it seems to me that citing [8] is just related to its section 2 about the explanation of ADS-B data sources, not really about the topic of the paper which is filtering techniques for ADS-B trajectory preprocessing. Maybe it should be stated such as "…via on-board equipment, see Section 2 of [8]." or better citing the "The 1090 Megahertz Riddle" (https://mode-s.org/1090mhz/) book.
- Line 90: what is intended for 'keywords'? Maybe just change "…and author keywords;" → "and authors;"
- Lines 100–103: use a numbered list (or number them inline) for the eight major domains. (ok clearer from Table 1.)
- Line 115: "…followed by a Multi-Cell Neural Network…" → "followed by a Multi-Cell Neural Network (MCNN)"
- Line 121: add space "optimization.ADS-B data…" → "optimization. ADS-B data…"
- Lines 150–152: the cited work by Schulz et al. [22] is not using "FDR and ADS-B data to model fuel consumption and operational efficiency using machine learning methods"…There is definitely something amiss with citations.
- Line 164: "…EUROCONTROL PRU initiative." → "…EUROCONTROL PRC initiative." Also add PRC acronym, Performance Review Commission.
- Figure 2: pie charts should be ordered decreasingly from 12 o'clock, see also these guidelines (https://data.europa.eu/apps/data-visualisation-guide/guidelines-for-pie-charts). After all keeping the order of the eight domains doesn't mean anything semantic.
- Section 3 title: unfortunately falls at the end of the page…typographically, it is not great.
- Line 219: "discrepancies between barometric and geometric altitude" these are structural for en-route because barometric altitude is based on standard air pressure which varies with local meteo conditions…so it is not an issue of accuracy!
- Figure 3: it is quite far away from where it is cited…It should be positioned alone closer to citation and to occupy less vertical space it could be designed as an horizontal pipeline.
- Line 259: DBSCAN has already been defined as an acronym in Line 114.
- Line 264: AE has already been defined in the Abstract (like ADS-B which has not been repeated).

- Line 275: add space "(1)Interpolation" → "(1) Interpolation".
- Line 286: PCHIP is used once, so maybe no need to define the acronym.
- Line 301: add space "…trajectory features.The representative" → "…trajectory features. The representative".

Recommendation: Revisions Required

3. Response - round 1

3.1 Response to reviewer 1

Major Issue 1. The experimental design suffers from significant methodological limitations that prevent generalizable conclusions. The paper claims to investigate "the relationship between data cleaning and algorithmic performance in ADS-B analytics," yet tests only a single autoencoder architecture on a single reconstruction task. The authors themselves acknowledge that "the impact of noise strongly depends on the model's structure and learning mechanism," which directly limits the extent to which general cleaning guidance can be drawn. The conclusion that "noise suppression should prioritize smoothing and interpolation optimization" cannot be supported by evaluating only one model type. Different algorithms may exhibit fundamentally different sensitivities to specific noise characteristics.

Response

We agree that relying on a single autoencoder architecture limited the generalisability of the original conclusions. In the revised manuscript, we have expanded the experiments to include three distinct autoencoder architectures: a fully connected autoencoder (FC-AE), an LSTM-based autoencoder (LSTM-AE), and a GRU-based autoencoder (GRU-AE). The FC-AE treats the entire trajectory as a flat vector and learns global spatial patterns through feedforward layers, while the LSTM-AE and GRU-AE process the trajectory as a temporal sequence and capture dependencies across time steps through gating mechanisms. These three architectures were chosen specifically because they represent fundamentally different learning mechanisms, spanning both feedforward and recurrent paradigms, and thus provide a meaningful test of whether noise sensitivity patterns are architecture-dependent.

Each architecture is evaluated under identical noise conditions (Gaussian, drift, spike, and missing data) across all four airport datasets (Zurich, Harbin, Guangzhou, and Hangzhou), using the same baseline trajectories, noise parameter ranges, and evaluation metrics. This yields a total of 12 architecture–dataset combinations, each swept across the full range of noise intensities for all four noise types.

Our results show that, within each dataset, all three architectures produce concordant noise sensitivity rankings. That is, the relative ordering of noise type impact is consistent regardless of whether a feedforward or recurrent model is used. However, this ranking varies across datasets, indicating that noise sensitivity is primarily driven by the geometric and operational characteristics of the airport environment rather than by the choice of model architecture. We have accordingly revised the conclusions to avoid architecture-specific cleaning recommendations and instead emphasise the dataset-dependent nature of noise impact. The updated experimental design and results are presented in Sections 4.4 and 4.5.

Major Issue 2. The baseline selection methodology introduces model-dependent circularity that compromises the validity of the evaluation. Section 4.2 uses autoencoder reconstruction error to identify "clean" trajectories as the baseline, then tests the same autoencoder on corrupted versions of these trajectories. This creates a self-referential evaluation in which the model is assessed on data it already reconstructs well.

Response

We appreciate the reviewer raising this methodological concern. The baseline selection is not intended as a performance claim but as a controlled-variable design choice. Trajectories with inherently high reconstruction error—due to unusual geometry or extreme manoeuvres—would mask noise-induced degradation: the δRMSE signal from injected noise becomes indistinguishable from the pre-existing reconstruction floor. By selecting trajectories with low clean RMSE, we ensure that subsequent degradation under noise injection is attributable to the noise itself rather than to intrinsic trajectory complexity. Our analysis metric, $\Delta\text{RMSE} = \text{RMSE}_{\text{noisy}} - \text{RMSE}_{\text{clean}}$, further isolates the noise effect by subtracting each trajectory's own baseline, so the absolute reconstruction quality at the starting point does not enter the comparison. The consistent and strongly differentiated degradation patterns across noise types—spike δRMSE remaining near zero while missing data δRMSE increases by an order of magnitude—confirm that the experimental design captures genuine noise sensitivity rather than a selection artifact. The revised manuscript clarifies this rationale in Section 4.2. The revised manuscript clarifies the role of baseline selection in Section 4.2.

Major Issue 3. The artificial noise injection lacks validation against real ADS-B error patterns. While Section 4.3 describes three noise types, the paper provides no quantitative evidence that the synthetic noise parameters correspond to the statistical characteristics or magnitudes of real ADS-B errors documented in the cited literature. Without such validation, conclusions about model sensitivity to Gaussian, drift, or spike noise may not translate to real-world ADS-B data.

Response

We thank the reviewer for this important point. In the revised manuscript, we now explicitly calibrate our synthetic noise parameters against documented real-world ADS-B error characteristics:

- **Gaussian noise:** Our parameter range is calibrated against position accuracy figures reported by NATS (2015) and Schäfer (2014), which document typical ADS-B position errors on the order of tens to low hundreds of metres. We map these to equivalent magnitudes in our normalized [0, 1] coordinate space using each dataset's spatial extent.
- **Drift noise:** Our constant-offset parameters correspond to systematic biases documented in Olive & Sun (2024), who characterise persistent positional shifts arising from avionics or transponder errors.
- **Spike noise:** Spike magnitudes and frequencies are informed by outlier characteristics reported in Nur et al. (2018) and Olive & Sun (2024), reflecting sudden erroneous position jumps observed in operational data.
- **Missing data:** Gap lengths and frequencies are calibrated against reported message loss rates in operational ADS-B reception (NATS, 2015; Mohleji & Wang, 2010).

A detailed mapping between our normalised noise parameters and their real-world equivalents (in metres/degrees) is now provided in Section 4.3 of the revised manuscript.

Major Issue 4. Critical data quality issues discussed earlier in the paper are excluded from the experimental evaluation. Although missing data and interpolation are emphasized throughout the review sections, the authors explicitly state that the autoencoder "did not further investigate the effect of missing-value errors". This omission is significant, as interpolation to handle missing data is among the most common preprocessing steps described in the paper. The experimental framework should include missing-data scenarios and compare different interpolation strategies.

Response

We fully agree with the reviewer that the omission of missing data from the original experiments was a significant gap. The revised manuscript now includes missing data as the fourth noise type alongside Gaussian, drift, and spike noise. Missing data is simulated by randomly removing a fraction of trajectory points (missing rate from 0 to 0.50) and reconstructing them via linear interpolation before encoding, mirroring the most common preprocessing strategy in ADS-B applications. The results reveal that missing data is among the most impactful noise types, ranking first in degradation on datasets with complex approach geometries (Zurich) and third on datasets with more regular trajectory patterns (Hangzhou, Guangzhou). This finding directly supports the reviewer's intuition that interpolation quality is a critical factor in downstream algorithm performance. Full results are reported in Section 4 and Appendix.

Medium Issue 1. Figure 8 presents boxplots comparing reconstruction error across noise conditions, but no statistical significance testing is reported. The claim that the model "shows greater tolerance to drift and spike noise" would benefit from quantitative statistical support beyond visual inspection. In addition, "normalized RMSE" is referenced without specifying the normalization procedure in the figure caption or main text.

Response

We thank the reviewer for this suggestion. In the revised manuscript, we now apply the Wilcoxon signed-rank test to all pairwise comparisons between noise types, for each of the 12 model–dataset combinations. The table summarises the ΔRMSE (defined as RMSE at maximum noise level minus RMSE under clean conditions) for each noise type, along with the resulting sensitivity ranking:

Table 1. Summary of ΔRMSE (RMSE at maximum noise level − RMSE under clean conditions) for each model–dataset combination across four noise types, with the resulting sensitivity ranking.

Model	**Dataset**	ΔRMSE Gaussian	ΔRMSE Drift	ΔRMSE Spike	ΔRMSE Missing	**Ranking**
FC-AE	Zurich	0.063497	0.022653	0.003446	0.148815	M > G > D > S
FC-AE	Hangzhou	0.003786	0.004500	0.000180	0.003128	D > G > M > S
FC-AE	Guangzhou	0.003165	0.004599	0.000129	0.002638	D > G > M > S
FC-AE	Harbin	0.002506	0.004696	0.000094	0.005076	M > G > D > S
LSTM-AE	Zurich	0.005429	0.003193	0.000283	0.009640	M > G > D > S
LSTM-AE	Hangzhou	0.002713	0.005091	0.000088	0.002178	D > G > M > S
LSTM-AE	Guangzhou	0.001999	0.004554	0.000064	0.001731	D > G > M > S
LSTM-AE	Harbin	0.002347	0.005376	0.000078	0.004166	D > M > G > S
GRU-AE	Zurich	0.069575	0.008906	0.006115	0.047384	G > M > D > S
GRU-AE	Hangzhou	0.002330	0.004720	0.000079	0.002277	D > G > M > S
GRU-AE	Guangzhou	0.002311	0.004546	0.000077	0.001945	D > G > M > S
GRU-AE	Harbin	0.002880	0.005099	0.000099	0.005125	M > D > G > S

All pairwise differences between noise types are statistically significant ($p < 0.05$) across all 12 combinations. Full pairwise test results, including W-statistics and p-values, are provided in Appendix X of the revised manuscript.

The original boxplots (Figure 8) have been replaced with median ± IQR line plots, which more clearly convey the trend of reconstruction error as noise intensity increases across all four noise types, three architectures, and four datasets.

We have also clarified in Section 4.2 that all trajectory coordinates are normalised to [0, 1] using min-

max scaling per coordinate dimension, and that RMSE is computed in this normalised space. This is now stated explicitly in both the main text and figure captions.

Medium Issue 2. Several key experimental parameters lack justification. The autoencoder uses a 32-dimensional latent representation for a 200-dimensional input, but no rationale is provided for this compression ratio. More critically, the noise injection parameters (e.g., Gaussian sigma values, drift rates, spike probabilities in Figure 8) are not justified relative to realistic ADS-B error magnitudes reported in the literature.

Response

We thank the reviewer for raising this point. In the revised manuscript, the choice of a 32-dimensional latent space for a 200-dimensional input is now explicitly justified in Section 4.2, following established practice in trajectory autoencoder literature, specifically, Olive & Basora (2020) and Krauth (2023) adopt comparable compression ratios for flight trajectory reconstruction tasks. Regarding noise parameters, as detailed in our response to Major Issue 3, all noise injection parameters are now calibrated against real ADS-B error magnitudes documented in the literature. A mapping between normalised parameter values and their real-world equivalents is provided in Section 4.3.

Medium Issue 3. The experimental scope is narrow. Using data from a single airport (Zurich) over a two-month period and selecting only 100 trajectories as the baseline raises questions about representativeness. It is unclear whether these trajectories adequately capture the diversity of aircraft types, flight phases, and operational scenarios discussed in Section 2.

Response

We acknowledge this limitation. In the revised manuscript, we have substantially expanded the experimental scope by including three additional airport datasets, namely Harbin, Guangzhou, and a second European dataset, alongside the original Zurich data. The experiments now cover four airports across different geographic regions and operational environments, each with its own baseline trajectory set. All analyses (three architectures $\times$ four noise types) are repeated across all four datasets, allowing us to assess whether findings generalise beyond a single airport context. The expanded results are presented in Sections 4.4 and 4.5. We note that while the baseline size per dataset remains modest, the consistency or divergence of results across datasets provides stronger evidence than a single-airport study alone. In particular, we find that noise sensitivity rankings are largely dataset-dependent rather than architecture-dependent, which itself speaks to the importance of multi-airport evaluation.

Medium Issue 4. The paper correctly identifies a lack of transparency in the literature regarding data-cleaning procedures, yet its own experimental methodology is incompletely specified in the main text. Critical implementation details for the autoencoder and noise injection are deferred to a GitHub repository. Given the paper's emphasis on methodological clarity and reproducibility, these details should be fully documented within the manuscript.

Response

We accept this criticism. The revised manuscript now documents all implementation details inline: autoencoder architectures, hyperparameters, and training procedures in Section 4.2 (Table 2); noise

injection parameters and their literature-calibrated ranges in Section 4.3; and the preprocessing pipeline in Section 4.1. The GitHub repository remains available for full code reproducibility but is no longer required to understand or replicate the experimental design.

Minor Issue 1. The paper states that results are averaged over 10 repetitions, yet Figure 8 boxplots suggest a larger number of samples. The figure caption should clarify whether the distributions represent variation across trajectories, repetitions, or both.

Response

We thank the reviewer for pointing out this ambiguity. In the original manuscript, the boxplots in Figure 8 depicted the distribution of per-trajectory RMSE values across the baseline trajectory set within a single experimental run, rather than variation across repetitions. The 10 repetitions mentioned in the text referred to the noise injection process, where each trajectory was corrupted independently with random noise, and the reported metrics were averaged over these repetitions to reduce stochastic variability. The figure caption did not make this distinction clear.

In the revised manuscript, we have replaced the boxplots with median ± IQR line plots, where the median and interquartile range are computed across trajectories at each noise level, after averaging over repetitions. This is now stated explicitly in the figure captions and in Section 4.2, so that the source of variation is unambiguous.

Minor Issue 2. Section 3.3 provides a thoughtful discussion of the limitations of cleaning methods, but these points could be strengthened by explicitly linking them to specific studies cited earlier in the literature review.

Response

We appreciate this suggestion. The limitations discussed in Section 3.3 were originally framed as general methodological observations rather than conclusions drawn from specific empirical findings in the cited literature. In the revised manuscript, we have restructured Section 3.3 to more clearly connect these points to the experimental evidence presented in Sections 4.4 and 4.5. For instance, the observation that noise impact varies depending on data context is now directly supported by our own multi-airport results, which show that sensitivity rankings differ across datasets. Where possible, we have also added references to relevant studies that discuss similar methodological considerations in the ADS-B processing literature. We recognise that a more systematic survey linking specific cleaning method limitations to documented empirical cases remains a valuable direction, and we intend to pursue this in future work.

Minor Issue 3. In Table 1, entries in the "Role of ADS-B Data" column such as "Core data source" are vague. More specific descriptions of how data quality requirements differ across application categories would better motivate why cleaning strategies may need to vary.

Response

We agree that the original descriptions in the "Role of ADS-B Data" column were too general to convey meaningful distinctions across application categories. In the revised manuscript, we have added a new column to Table 1, "Key Data Quality Requirements," which identifies the specific data quality

dimensions most relevant to each application category. For example, trajectory prediction applications are most sensitive to positional accuracy and temporal continuity, whereas operational safety applications prioritise inter-aircraft relative accuracy and update rate. These distinctions help motivate why a uniform cleaning strategy is unlikely to suit all downstream tasks, which is a central argument of the paper.

3.2 Response to reviewer 2

Response

We sincerely thank the reviewer for the careful and thorough reading of our manuscript. The detailed comments on citation accuracy, acronym usage, figure presentation, and table content have been very helpful in improving the clarity and professionalism of the paper. We have addressed each point as follows.

Citation corrections. We have carefully reviewed all references flagged by the reviewer. The mismatched citations in the introduction (Lines 11–14), including the incorrect association of Schlosser et al. with flight phase identification and Roosenbrand et al. with airport operations optimization, have been corrected to accurately reflect each cited work. The description of Schultz et al. (Lines 150–152) has also been revised to match the actual content of the referenced paper. For the citation at Line 41 regarding ADS-B transmission via on-board equipment, we have added a more specific reference following the reviewer's suggestion.

Literature search criteria (Line 90). We have revised "Paper title and keywords" to "Paper title and authors" as suggested.

Barometric vs. geometric altitude (Line 219). We agree with the reviewer that this is a structural difference arising from different measurement definitions rather than an accuracy issue. In the revised manuscript, this example has been moved from the "Accuracy" dimension to "Consistency," with an explicit note that the discrepancy reflects different physical measurement bases rather than instrument error.

Figure 2 ordering. The pie chart sectors have been reordered in decreasing order from 12 o'clock as suggested. The figure caption has also been simplified, with abbreviated category labels now referencing Table 1 for full definitions.

Figure 3 positioning. We appreciate this suggestion. The vertical layout was intentionally chosen to allow side-by-side placement with Figure 2, reducing overall vertical space. We have, however, adjusted the float placement to position Figure 3 closer to its first citation in the text.

Acronym definitions. We have conducted a thorough pass to ensure all acronyms are defined at first use in the running text. Acronyms appearing only within Table 1 (DAA, DTW, KPI, MILP, SDR, TimeGAN, GMM) are now defined in a table footnote. Unnecessary abbreviations such as 4D and UTC have been replaced with their full forms where clearer.

Table 1 specificity. We agree that the original "Role of ADS-B Data" column was too generic. A new column, "Key Data Quality Requirements," has been added to Table 1, identifying the specific data quality dimensions most relevant to each application category.

Formatting and typographical corrections. All spacing issues, redundant acronym redefinitions, and other typographical errors identified by the reviewer have been corrected in the revised manuscript.

We are grateful for the reviewer's meticulous attention to detail, which has substantially improved the presentation quality of our paper.