

EDITORIAL

Reviews and Responses for Estimating fuel consumption over flight segments using gradient boosting methods

Authors: Muhammad Fazlur Rahman, Aidana Tassanbi, Zhan-Peng Yang, Mohammed El Dor, and Bob van Dillen

Reviewers: Richard Alligier and Quinten Goens

Editor: Enrico Spinielli

1. Original paper

The DOI for the original paper is <https://doi.org/10.59490/joas.2026.8759>

2. Review - round 1

2.1 Reviewer 1

This article presents the methodology used by the authors to obtain a machine learning model predicting the amount of fuel burnt at certain segment of flight. This work was done to compete in the Performance Review Commission (PRC) Data Challenge 2025 in which the authors ranked first.

The article describes the data and the method used to obtain good predictions. An analysis of different feature set and configuration is performed, showing which sets of features are important or not. The paper is well written and clear. I believe that the reliability of the table comparing the configurations could be strenghted by training more models, either using different seeds and/or different training/test sets; and some experiments could be done to better investigate the impact of the ACARS features (see “Detailed Remarks/Questions” below). However, I don’t believe that these possible improvements are “mandatory”, and I recommend an accept with minor revision.

2.1.1 Detailed Remarks/Questions

1. I wonder if the energy rate related features could be a good features? Basically, the thrust is equal to the drag plus energy rate divided by the airspeed. Did you try such features?
2. Segments are divided on 10 sub-intervals, where the mean and std of the vertical rate is considered as features. Did you consider also adding airspeed and/or altitude, computed on these sub-intervals, as features?
3. The paper states that there are some numerical instabilities when computing the recursive mass. One solution to this could be to clip the computed fuel flow just like the fuel flow that we want to predict was clipped. That being said, it does not change that we have incomplete trajectory data.
4. An other way to deal with the mass would be to use a good prediction of the TOW $\hat{m}_{TOW}$, a good prediction of the total fuel burnt during the whole flight $f_{whole-flight}$ (I do not know how to obtain it but I guess some formulas could exist like the Breguet formulas maybe??), and then perform a linear like in the Heuristic Mass but with $m_{TO} = \hat{m}_{TOW}$ and $m_{LD} = \hat{m}_{TOW} - f_{whole-flight}$.

5. It is said (on line 555) that ACARS measurements provides airspeed with higher fidelity. It might be the case but the better predictive results could also come from the fact that, on some segments, ACARS measurement is available whereas groundspeed is not available. Two hypothesis/ explanation here: (i) ACARS measurement provides airspeed with higher fidelity (as stated on line 555), and/or (ii) ACARS measurement might be available on some segments when groundspeed is not. To investigate on this, one could compute NaN ratio for the groundspeed, the ACARS speed, and their “intersection”. An other idea would be to test if ACARS speed, set to NaN when “groundspeed.isna()”, still improve the predictive performance.
6. The different configuration are compared with models trained with a given seed, on a given training set, and tested on a specific set of examples. Because of that, as stated on line 570, it could be difficult to draw conclusions from Table 5 when RMSE are close. Maybe, it could be nice to use several seeds to train several models, and display the average RMSE of these different models in Table 5. An other modification could be to test different training/test set, especially as the errors can be driven by some examples with a large error. Merging all the sets together and performing a K-fold validation could be a better way to estimate the performance of a configuration. An other way, if we want to keep the original train/test split, could be to perform train and test on several bootstrapped versions of the train and test set respectively.
7. Related to the previous remark, sometimes the difference in RMSE between configurations is too small to be sure that this difference actually exists. In such case, it is better to be cautious on the analysis, and even more on possible explanation. The authors are cautious in the article, but maybe they could be even more cautious on the possible explanation on lines 509-511.

2.1.2 Text Improvements

Clarifications

1. It could be nice to explain how you handle the split of the data set in Section 4.1, and why you did it that way. A simple random split was prone to data leakage here because the data are not iid draws. It is an interesting subject that could be useful to readers, even on completely different problems. It is classical “gotcha” but that can be easily “missed”.
2. “aircraft type (categorical)” appears on line 306 but not in ADS-B group on Table 3. As a side note, it would make it the 15th variables if I counted right. Whether the aircraft type is used as a feature or not should be more explicit.
3. The “flight phase” is a categorical feature, however points in the same segment could be tagged with different phases. I can guess that there is some kind of “majority vote” to tag the whole segment with a specific phase, but it should be more explicit in the text.

Typos

1. line 87: “Section ??”
2. line 101: “during within” → “within”
3. line 110: “which is also can be observed form Table 1” → “which can also be observed from Table 1”. As a note, Table 1 only shows number of flights and segments, not the duration of the segments, so maybe do not mention the Table 1 here.
4. line 165 “four(may)” → “four (may)”
5. line 165 “figure 4” → “Figure 4”
6. formula (1) (between line 216 and 217) and formula (4) “[0, 1, . . . , k]” → “{0, 1, . . . , k}”
7. both “fuel-flow” and “fuel flow” appears in the text, maybe make a choice on this
8. line 314 “Section ??”

2.2 Reviewer 2

2.2.1 Introduction

This document is a review of the submission Estimating fuel consumption over flight segments using gradient boosting methods by Rahman et al., submitted to the Journal of Open Aviation Science (JOAS) as part of the EUROCONTROL PRC 2025 Data Challenge. The paper investigates the extent to which aircraft fuel consumption can be inferred from ADS-B trajectory data, supplemented by ACARS messages, ERA5 weather data, and physics-informed mass features. The authors construct a multi-stage preprocessing pipeline, derive several feature groups, and train a LightGBM regression model to predict segment-level average fuel flow. A structured ablation study evaluates the marginal contribution of each feature group and of two alternative mass-estimation strategies.

The review is organised as follows. Section 2.2.2 summarises the strengths of the submission. Section 2.2.3 lists potential issues identified in the manuscript. Section 2.2.4 raises questions and suggestions for further research. Section 2.2.5 lists minor textual issues. Line and section references correspond to the reviewed pre-print (v1).

2.2.2 Strengths

- The description of the input data is complete and clearly written and the feature engineering section is well detailed. The rationale behind each feature group is convincingly substantiated.
- Section 5.1 (Fuel-Flow Filter Sensitivity) is a good example of how individual preprocessing decisions are justified through targeted experiments. Applying the proposed fuel-flow filter is shown to substantially reduce the model RMSE.
- Table 5, together with the analysis in Section 5, provides a clear quantitative basis for the inclusion of certain features and the exclusion of others (e.g. wind in the heuristic-mass configuration).
- The finding that fuel flow and the associated emissions depend strongly on aircraft mass (whilst not new) underscores the relevance of the PRC 2024 Data Challenge. The reuse of the winning PRC 2024 take-off weight model in the present feature-engineering pipeline is a commendable example of cumulative use of the open challenge results.
- The initially counter-intuitive finding that Block 3 (Heuristic Mass) outperforms Block 2 (Recursive Mass) is well explained in lines 471–473 by the accumulation of errors inherent to the recursive integration scheme.

2.2.3 Potential issues

- **OpenAP phase-identification module is aircraft-type agnostic** (lines 73–75). The authors state that they use “OpenAP’s phase detection module and fuel flow model, both of which draw on aircraft type-specific parameters included in the library”. This is correct for the fuel-flow model but not for the phase-detection module. The underlying fuzzy-logic approach uses universal membership functions for altitude, rate of climb, and speed that are explicitly designed to accommodate inter-type variability without per-type tuning (Sun, Ellerbroek, and Hoekstra [1]), and the corresponding implementation instantiates `FlightPhase()` with no aircraft-type argument and hard-coded membership parameters (TUDelft-CNS-ATM [2]). We suggest correcting this characterisation.
- **Source of the fuel-flow clipping bounds** (Section 4.2). The authors clip the fuelflow target to the interval $[0.05, 6.5]$ kg/s, with 0.05 kg/s described as the minimum descent fuel flow of a small narrow-body aircraft and 6.5 kg/s as the assumed maximum climb fuel flow of a large wide-body aircraft. Section 3.1 cites Baklacioglu (0.04–1.84 kg/s for B738) (Baklacioglu [3]) and the ICAO Aircraft Engine Emissions Databank (0.023–4.69 kg/s per engine) (International Civil Aviation Organization (ICAO) [4]) as reference bounds, but these references are not explicitly tied to the specific values selected in Section 4.2. In particular, the ICAO value is reported on a per-engine

basis, so its relation to the aggregate aircraft-level upper bound of 6.5 kg/s is not immediate. We suggest making the link between the cited references and the chosen clipping bounds explicit.

- **Variation in fuel-segment durations not visible in Table 1** (lines 109–110). The authors observe that “the duration of the fuel reporting windows varies considerably, which is also can be observed form Table 1”. Table 1 reports the number of flights and fuel segments per aircraft type but contains no information on segment durations. The intended supporting evidence should be added (for example a histogram or a summary statistic of segment durations) or the reference to Table 1 should be revised.

2.2.4 Questions and suggestions for further research

- **Treatment of implausible fuel-flow values** (Section 4.2). Implausible fuel-flow values are clipped to the nearest plausible boundary. The stated causes, data artefacts, timestamp irregularities, and reporting errors, suggest that these outliers are effectively random. Clipping them to a boundary value then introduces a systematic bias rather than preserving the underlying distribution. Alternative strategies, such as linear interpolation between neighbouring timestamps, substitution with a local mean, or outright exclusion of affected segments, could be investigated and may yield further improvements in model accuracy.
- **Choice of the number of vertical-rate sub-intervals** (Section 4.3). Each fuel segment is divided into ten equal-duration sub-intervals, for which the mean and standard deviation of the vertical rate are computed. The choice of exactly ten sub-intervals is not substantiated. Given that the importance of the vertical-rate split is highlighted in lines 445–448, a sensitivity analysis with respect to the number of sub-intervals would strengthen the result and could reveal whether a coarser or finer decomposition is preferable.

2.2.5 Typographical and editorial remarks

- Line 39: “performance review committee” should be “Performance Review Commission”.
- Line 87: “Section ??”, broken cross-reference.
- Lines 109–110: “which is also can be observed” should be “which can also be observed”.
- Line 110: “form” should be “from”.
- Lines 165–166: “(may change this number and figure)” appears to be an authoring note that has not been removed.
- Line 304: “Automatic Dependent Surveillance–Broadcast” overlaps with the line number in the rendered PDF.
- Line 314: “Section ??”, broken cross-reference.
- Figure 1 caption: “for randomly sampled 1000 flights” reads more naturally as “for 1000 randomly sampled flights”.
- Figure 2 caption: “fuel segment.” should be “fuel segments.” (multiple shaded regions are shown).
- Figure 2 caption: “yellow shaded regions” should be “green shaded regions” as it is green on the image.
- MLW and OEW used without local introduction (Equation 2). Equation 2 uses MLW and OEW without introducing them in the surrounding text.

3. Response - round 1

3.1 Response to reviewer 1

Response

We thank the reviewer for the positive assessment and for recommending acceptance with minor revision. We address each comment below.

3.1.1 Detailed Remarks/Questions

Comment 1

I wonder if the energy rate related features could be a good features? Basically, the thrust is equal to the drag plus energy rate divided by the airspeed. Did you try such features?

Response

We did not add energy-rate features separately, because the estimated fuel flow from OpenAP (the FF Est group) is already included as a feature in the training set. Since the OpenAP fuel-flow model accounts for the aircraft's energy state in its underlying thrust computation, this feature already conveys the information an explicit energy-rate term would provide, and we did not expect an additional gain from engineering it separately. We did also test energy-rate-related features during model development and found that they did not improve predictive performance; these runs were not retained, so we report this as a development observation rather than a documented result.

Comment 2

Segments are divided on 10 sub-intervals, where the mean and std of the vertical rate is considered as features. Did you consider also adding airspeed and/or altitude, computed on these sub-intervals, as features?

Response

The sub-interval split is intended to capture within-segment phase transitions, which the vertical rate signals most directly (e.g., climb-to-cruise). As we note in the paper, **“This decomposition captures transitions between flight phases (e.g., climb-to-cruise) that a single segment-level feature would obscure.”** Both altitude and airspeed are already retained as segment-level features, with mean altitude in the ADS-B group and airspeed from ACARS. We restricted the higher-resolution split to the vertical rate because each additional variable would contribute another 20 features per segment. Applying the mean and standard deviation of both altitude and airspeed across the 10 sub-intervals would add 40 features on top of the 40 already used, and we prefer to keep only the most informative features rather than double the input dimensionality. We have clarified this rationale in Section 4.4.

Comment 3

The paper states that there are some numerical instabilities when computing the recursive mass. One solution to this could be to clip the computed fuel flow just like the fuel flow that we want to predict was clipped. That being said, it does not change that we have incomplete trajectory data.

Response

This is a good suggestion, and we agree that clipping the computed fuel flow would mitigate some of the instability described in Section 3.3.1 — though, as the reviewer anticipates, it would bound rather than remove the effect of incomplete or noisy trajectory data near take-off and landing. In our case, however, the heuristic mass model already provides a stable alternative that avoids the error-accumulation problem by construction, so applying this fix to the recursive approach is no longer necessary for the

results presented here. Should this line of work be taken further, we would favour a formulation that combines the heuristic and recursive approaches — for instance, using the heuristic profile to constrain or initialise the recursive integration — rather than pursuing improvements to the recursive mass model in isolation.

Comment 4

An other way to deal with the mass would be to use a good prediction of the TOW $\hat{m}_{TOW}$, a good prediction of the total fuel burnt during the whole flight $f_{whole-flight}$ (I do not know how to obtain it but I guess some formulas could exist like the Breguet formulas maybe??), and then perform a linear like in the Heuristic Mass but with $m_{TO} = \hat{m}_{TOW}$ and $m_{LD} = \hat{m}_{TOW} - f_{whole-flight}$.

Response

We agree that improving the mass estimate with a more accurate take-off weight is a direction we already note as future work (Sections 5.4 and 6). We did not pursue the Breguet approach specifically, because it requires several parameters (e.g., specific fuel consumption, lift-to-drag ratio, cruise speed and altitude) that would each have to be estimated per flight and per aircraft type. Chaining these estimates accumulates uncertainty, adding noise that we expect would outweigh the benefit of the more physically grounded fuel-burn estimate over our current heuristic. The Breguet equation also assumes cruise-phase equilibrium, whereas the dataset covers all different flight phases.

Comment 5

It is said (on line 555) that ACARS measurements provides airspeed with higher fidelity. It might be the case but the better predictive results could also come from the fact that, on some segments, ACARS measurement is available whereas groundspeed is not available. Two hypothesis/explanation here: (i) ACARS measurement provides airspeed with higher fidelity (as stated on line 555), and/or (ii) ACARS measurement might be available on some segments when groundspeed is not. To investigate on this, one could compute NaN ratio for the groundspeed, the ACARS speed, and their “intersection”. An other idea would be to test if ACARS speed, set to NaN when “groundspeed.isna()”, still improve the predictive performance.

Response

We agree that the “higher fidelity” is an ambiguous word that needs to be revised to better illustrate our findings and intents. For clarification, on the raw dataset, when ACARS airspeed is present, the ADS-B ground speed is never exactly available. Since the raw ACARS and ADS-B messages do not coincide on the timestamp, the resampling process interpolates the states from different sources to place them on the same 1-second grid. From our observation of the dataset, there could be several scenarios regarding air speed and ground speed availability: i) ACARS messages are embedded in dense ADS-B messages; ii) ACARS messages are isolated from ADS-B messages, e.g. segments with no ADS-B coverage;

Our hypothesis is that: In case i), this is the ideal segment where the interpolation of data from both ADS-B and ACARS provides strong prior to form prediction. However, there still could be the scenarios where ACARS messages carry no airspeed information. The only data we could rely on are those from ADS-B sources. In case ii), the ADS-B data are interpolated between very long horizons, or even extrapolated. If ACARS airspeeds are available, they provide the only strong prior for prediction.

Our earlier trials included airspeed reanalysis from ERA5 weather data and ground speed. However, that version performed worse than including the direct ACARS airspeed data. The rest of the paragraph from line 555 explains our hypothesis: The ACARS airspeed data better captures the local physical and weather condition than the reanalysis could provide. It is “higher fidelity” in a sense that ACARS

airspeed, and its interpolation, is closer to the local ground truth, but we agree that we should phrase this in a more rigorous way.

The two candidate airspeed sources (ACARS and weather reanalysis) differ in reporting mechanism, sparsity, and downstream processing, we regard an isolated attribution of the observed RMSE gap to measurement fidelity alone as difficult to establish. We will accordingly revise the text to present the ACARS preference as an empirical result rather than a validated fidelity claim.

Comment 6

The different configuration are compared with models trained with a given seed, on a given training set, and tested on a specific set of examples. Because of that, as stated on line 570, it could be difficult to draw conclusions from Table 5 when RMSE are close. Maybe, it could be nice to use several seeds to train several models, and display the average RMSE of these different models in Table 5. An other modification could be to test different training/test set, especially as the errors can be driven by some examples with a large error. Merging all the sets together and performing a K-fold validation could be a better way to estimate the performance of a configuration. An other way, if we want to keep the original train/test split, could be to perform train and test on several bootstrapped versions of the train and test set respectively.

Response

We agree. As noted in our limitations (Section 5.3.3), the evaluation uses a single ranking set and a single seed, and some of the observed RMSE differences are small enough to fall within the range of stochastic variation. We report single-run results on the challenge ranking set to stay consistent with the competition's fixed evaluation protocol, and a full multi-seed, K-fold, or bootstrap study across all configurations was beyond the scope of the submission because of the training cost of repeating every configuration. Following the reviewer's suggestion, we have added this evaluation protocol as an explicit direction for future work (Section 5.4), and we have toned down the wording where conclusions rested on close margins, so that no claim now depends on a difference that could stem from randomization.

Comment 6

Related to the previous remark, sometimes the difference in RMSE between configurations is too small to be sure that this difference actually exists. In such case, it is better to be cautious on the analysis, and even more on possible explanation. The authors are cautious in the article, but maybe they could be even more cautious on the possible explanation on lines 509-511.

Response

We have revised the affected passages in Section 5 so that the explanations of close-margin results are now explicitly explained, noting that the small differences may partly reflect stochastic variation rather than a specific effect. This addresses the caution requested here and complements the future-work evaluation protocol added in response to the previous remark.

3.1.2 Text Improvement

Clarifications

It could be nice to explain how you handle the split of the data set in Section 4.1, and why you did it that way. A simple random split was prone to data leakage here because the data are not iid draws. It is an interesting subject that could be useful to readers, even on completely different problems. It is

classical “gotcha” but that can be easily “missed”.

Response

The data would indeed result in leakage if split randomly, since segments from the same flight are strongly correlated. A naive random split over segments would place segments of the same flight in both subsets and leak information. To avoid this, we split at the level of individual flights rather than segments, sorting the data first by flight and then by time so that all segments of a given flight remain within a single subset. The only exception is the single flight that straddles the split boundary, so at most one flight out of roughly 10^4 can contribute to leakage, which has a negligible effect on the reported results. We have added a description of this procedure and its rationale as a new Section 4.1.

“aircraft type (categorical)” appears on line 306 but not in ADS-B group on Table 3. As a side note, it would make it the 15th variables if I counted right. Whether the aircraft type is used as a feature or not should be more explicit.

Response

Aircraft type is indeed used as a categorical feature within the ADS-B group but was omitted from Table 3. It has now been added explicitly, which brings the ADS-B group count from 14 to 15.

The “flight phase” is a categorical feature, however points in the same segment could be tagged with different phases. I can guess that there is some kind of “majority vote” to tag the whole segment with a specific phase, but it should be more explicit in the text.

Response

The segment-level “flight phase” label is taken from the phase of the first trajectory point within the segment, not a majority vote, and we have made this explicit in Section 4.4 where the ADS-B feature group is described. We include it as a simple, low-cost categorical feature that gives the model a rough indication of the segment’s starting regime, which is useful information given how strongly fuel flow depends on the flight phase. The finer within-segment behaviour, including phase transitions occurring later in the segment, is then captured by the vertical-rate split over the 10 sub-intervals. We did not isolate the individual contribution of the flight-phase feature in this study, and we have added such an analysis to the future-work directions.

Typos

1. line 87: “Section ??”
2. line 101: “during within” → “within”
3. line 110: “which is also can be observed form Table 1” → “which can also be observed from Table 1”. As a note, Table 1 only shows number of flights and segments, not the duration of the segments, so maybe do not mention the Table 1 here.
4. line 165 “four(may” → “four (may”
5. line 165 “figure 4” → “Figure 4”
6. formula (1) (between line 216 and 217) and formula (4) “[0, 1, . . . , k]” → “0, 1, . . . , k”
7. both “fuel-flow” and “fuel flow” appears in the text, maybe make a choice on this
8. line 314 “Section ??”

Response

We have addressed all the items in the revised version.

3.2 Response to reviewer 2

Response

We thank the reviewer for the thorough and detailed review, and for highlighting both the strengths of the submission and several issues that improve the accuracy and clarity of the manuscript.

3.2.1 Potential issues

Comment 1

OpenAP phase-identification module is aircraft-type agnostic (lines 73–75). The authors state that they use “OpenAP’s phase detection module and fuel flow model, both of which draw on aircraft type-specific parameters included in the library”. This is correct for the fuel-flow model but not for the phase-detection module. The underlying fuzzy-logic approach uses universal membership functions for altitude, rate of climb, and speed that are explicitly designed to accommodate inter-type variability without per-type tuning (Sun, Ellerbroek, and Hoekstra [1]), and the corresponding implementation instantiates `FlightPhase()` with no aircraft-type argument and hard-coded membership parameters (TUDelft-CNS-ATM [2]). We suggest correcting this characterisation.

Response

We have made corresponding changes in Section 2.2, lines 74–76.

Comment 2

Source of the fuel-flow clipping bounds (Section 4.2). The authors clip the fuelflow target to the interval $[0.05, 6.5]$ kg/s, with 0.05 kg/s described as the minimum descent fuel flow of a small narrow-body aircraft and 6.5 kg/s as the assumed maximum climb fuel flow of a large wide-body aircraft. Section 3.1 cites Baklacioglu (0.04–1.84 kg/s for B738) (Baklacioglu [3]) and the ICAO Aircraft Engine Emissions Databank (0.023–4.69 kg/s per engine) (International Civil Aviation Organization (ICAO) [4]) as reference bounds, but these references are not explicitly tied to the specific values selected in Section 4.2. In particular, the ICAO value is reported on a per-engine basis, so its relation to the aggregate aircraft-level upper bound of 6.5 kg/s is not immediate. We suggest making the link between the cited references and the chosen clipping bounds explicit.

Response

We agree the link should be explicit. The cited bounds are not meant to be applied literally: Baklacioglu’s figures come from a single aircraft type (B738) [3], and the ICAO databank’s per-engine values represent the extremes across whatever engine types ICAO considered [4], which need not match the aircraft types in our dataset. In addition, the largest ICAO values are likely to reflect instantaneous fuel rates at the short initial take-off thrust phase rather than a sustained operating condition, so they overstate what a segment-averaged fuel flow should look like. We therefore treated both sources as outer reference bounds rather than precise limits, and chose filtering thresholds within that broader range aimed at retaining as much valid training data as possible. 6.5 kg/s was selected on this basis as a reasonable middle-ground upper bound rather than a value read directly off either citation. We have added this reasoning explicitly to Section 4.3 (previously Section 4.2).

Comment 3

Variation in fuel-segment durations not visible in Table 1 (lines 109–110). The authors observe that “the duration of the fuel reporting windows varies considerably, which is also can be observed from Table 1”. Table 1 reports the number of flights and fuel segments per aircraft type but contains

no information on segment durations. The intended supporting evidence should be added (for example a histogram or a summary statistic of segment durations) or the reference to Table 1 should be revised.

Response

We have removed the reference to Table 1 in that sentence and instead point to Section 2.1, where the range of fuel-interval durations (1.61 s to 1 hour) is already reported.

3.2.2 Questions and suggestions for further research

Comment 4

Treatment of implausible fuel-flow values (Section 4.2). Implausible fuel-flow values are clipped to the nearest plausible boundary. The stated causes, data artefacts, timestamp irregularities, and reporting errors, suggest that these outliers are effectively random. Clipping them to a boundary value then introduces a systematic bias rather than preserving the underlying distribution. Alternative strategies, such as linear interpolation between neighbouring timestamps, substitution with a local mean, or outright exclusion of affected segments, could be investigated and may yield further improvements in model accuracy.

Response

We thank the reviewer, whose comment prompted us to correct an imprecision in Section 4.3 (previously Section 4.2). In our implementation, segments whose fuel flow label falls outside the plausible range are actually *removed* from the training set, not clamped to the boundary. The previous description (and Eq. 5) characterised this as clipping, which was inaccurate. We have revised Section 4.3 so that it correctly describes the procedure as outright exclusion of the affected segments, precisely the alternative the reviewer identifies as preferable. Because these points are removed rather than relabelled, they cannot be biased at the boundary values by construction.

For the remaining suggestions, linear interpolation between neighbouring timestamps does not map onto our setting, since the target is a segment-level aggregate rather than a timestamp-level series; and substituting a local mean would first require defining which segments count as local to one another. Since the segments are independent aggregates that can have completely different feature profiles, there is no well-defined local mean to compute, so any definition would be an additional heuristic choice with its own parameters.

Comment 5

Choice of the number of vertical-rate sub-intervals (Section 4.3). Each fuel segment is divided into ten equal-duration sub-intervals, for which the mean and standard deviation of the vertical rate are computed. The choice of exactly ten sub-intervals is not substantiated. Given that the importance of the vertical-rate split is highlighted in lines 445–448, a sensitivity analysis with respect to the number of sub-intervals would strengthen the result and could reveal whether a coarser or finer decomposition is preferable.

Response

We agree that the choice of ten sub-intervals is heuristic and not substantiated in the current manuscript. A sweep over several values (for example 5, 10, 15, and 20 sub-intervals) would clarify whether ten is close to optimal and whether a coarser or finer decomposition is preferable. We did not run this sweep because it requires repeating a full experimental block for each setting, which was beyond the scope of the submission. We have noted the heuristic nature of this choice in Section 4.4 (previously Section 4.3)

and added the sensitivity analysis as a direction for future work in Section 5.4.

3.2.3 Typographical and editorial remarks

1. Line 39: “performance review committee” should be “Performance Review Commission”.
2. Line 87: “Section ??”, broken cross-reference.
3. Lines 109–110: “which is also can be observed” should be “which can also be observed”.
4. Line 110: “form” should be “from”.
5. Lines 165–166: “(may change this number and figure)” appears to be an authoring note that has not been removed.
6. Line 304: “Automatic Dependent Surveillance–Broadcast” overlaps with the line number in the rendered PDF.
7. Line 314: “Section ??”, broken cross-reference.
8. Figure 1 caption: “for randomly sampled 1000 flights” reads more naturally as “for 1000 randomly sampled flights”.
9. Figure 2 caption: “fuel segment.” should be “fuel segments.” (multiple shaded regions are shown).
10. Figure 2 caption: “yellow shaded regions” should be “green shaded regions” as it is green on the image.
11. MLW and OEW used without local introduction (Equation 2). Equation 2 uses MLW and OEW without introducing them in the surrounding text.

Response

All items have been addressed in the revised manuscript.

References

- [1] Junzi Sun, Joost Ellerbroek, and Jacco Hoekstra. “Large-Scale Flight Phase Identification from ADS-B Data Using Machine Learning Methods”. In: *7th International Conference on Research in Air Transportation*. Philadelphia, USA, 2016.
- [2] TUDelft-CNS-ATM. *Openap*: *openap/phase.py*. URL: <https://github.com/TUDelft-CNS-ATM/openap/blob/master/openap/phase.py>.
- [3] Tolga Baklacioglu. “Modeling the Fuel Flow-Rate of Transport Aircraft during Flight Phases Using Genetic Algorithm-Optimized Neural Networks”. In: *Aerospace Science and Technology* 49 (Feb. 1, 2016), pp. 52–62. ISSN: 1270-9638. DOI: [10.1016/j.ast.2015.11.031](https://doi.org/10.1016/j.ast.2015.11.031).
- [4] International Civil Aviation Organization (ICAO). *ICAO Aircraft Engine Emissions Databank*. 2025. URL: <https://www.easa.europa.eu/en/domains/environment/icao-aircraft-engine-emissions-databank>.