

ARTICLE

Estimating fuel consumption over flight segments using gradient boosting methods

Muhammad Fazlur Rahman ^{*}, Aidana Tassanbi ^{*}, Zhan-Peng Yang ^{*}, Mohammed El Dor ^{*} and Bob van Dillen ^{*}

Faculty of Aerospace Engineering, Delft University of Technology, Delft, The Netherlands

^{*}Corresponding author: m.f.rahman@tudelft.nl; a.tassanbi@tudelft.nl; z.yang-10@tudelft.nl; m.eldor@tudelft.nl; b.vandillen@tudelft.nl

(Received: 30 Mar 2026; Revised: 23 Jul 2026; Accepted: 18 Oct 2026; Published: 26 Oct 2026)

(Editor: Enrico Spinielli; Reviewers: Richard Alligier and Quinten Goens)

Abstract

Accurately estimating aircraft fuel consumption from surveillance data is essential for assessing environmental impact and improving the efficiency of air traffic operations. However, many variables that determine fuel burn, such as aircraft mass, engine performance, and atmospheric conditions, are not directly observable from ADS-B. This study investigates the extent to which fuel consumption can be inferred statistically from ADS-B trajectories that are enriched with weather information, inferred mass, and physics-informed features. We construct a multi-stage preprocessing pipeline that includes ADS-B cleaning, interpolation, resampling, wind-data augmentation using ERA5, estimation of take-off weight, and derivation of fuel flow and mass profiles based on the OpenAP performance model. Several feature groups are engineered, including statistical embeddings of ADS-B and ACARS variables, detailed vertical-rate phase decomposition, mass-related descriptors, and meteorological corrections. Using these features, we train a LightGBM regression model to predict fuel flow, which is then integrated to obtain segment-level fuel consumption. Experimental results on the challenge ranking dataset show that mass estimation, especially when using a heuristic interpolation between take-off and landing weight, substantially improves predictive accuracy. Models trained on fuel flow consistently perform better than those trained directly on fuel consumption, because the fuel flow target exhibits lower skewness and better reflects the underlying physics. The final configuration achieves an RMSE below 200 on fuel flow prediction, demonstrating the value of combining physics-informed features, statistical embeddings, and robust mass estimation. These findings highlight the importance of high-quality preprocessing and feature engineering for large-scale fuel-burn inference from sparse operational data.

Keywords: Fuel consumption estimation; Fuel flow; ADS-B; OpenAP; LightGBM

Abbreviations: ADS-B: Automatic Dependent Surveillance–Broadcast, ACARS: Aircraft Communications Addressing and Reporting System, ATOW: Actual Take-Off Weight, TOW: Take-Off Weight, MTOW: Maximum Take-Off Weight, MLW: Maximum Landing Weight, OEW: Operational Empty Weight, TAS: True Airspeed, CAS: Calibrated Airspeed, QAR: Quick Access Recorder, ERA5: ECMWF Reanalysis Version 5, ECMWF: European Centre for Medium-Range Weather Forecasts, GBDT: Gradient Boosting Decision Trees, RMSE: Root Mean Squared Error

1. Introduction

Increasing aviation activities contributes significantly to global CO₂ emissions, which makes accurate fuel consumption estimation essential for assessing environmental impact and improving operational efficiency [1]. However, data sources regarding aviation fuel consumption is often inac-

cessible where airline operators rarely shares it without restrictive licenses and agreement, making open research on this topic prohibitive.

Although the mass of fuel burnt during a flight can be inferred from the associated trajectory to a certain extent, it varies significantly based on numerous other variables such as aircraft mass, engine characteristics, flight phases, weather conditions, and many more. Most open and accessible trajectory data sources lack records of such features, e.g., automatic dependent surveillance broadcast (ADS-B) database. This challenge in scarcity of open data motivates the central question of this study:

To what extent can aircraft fuel consumption be statistically inferred from ADS-B trajectory data, supplemented by open-access operational and environmental information where available?

Previous research has explored both physics-based and data-driven approaches. Physics-based models use aerodynamic principles and engine performance data. Basis of aircraft data (BADA) [2] utilizes proprietary data to build aircraft performance models (APMs) which includes fuel flow rate in its calculation. OpenAP [3] builds on this foundation and incorporates open data and statistical regression to formulate the key parameters in the APMs. Data-driven systems such as Acropole [4] combine QAR data, operational information, and machine learning methods to estimate fuel flow. Other studies, including neural-network based work by Jarry *et al.* [5], show that high-frequency flight data can reveal latent performance characteristics.

In the 2025 data challenge [6] organized by the Performance Review Commission (PRC) of EUROCONTROL, we investigate a statistical approach to fuel consumption estimation using ADS-B trajectory data supported with weather information, inferred mass, and physics-informed features. The ground truth data of fuel consumption are provided by the organizers who extracted the relevant information from aircraft communication addressing and reporting system (ACARS) messages. We evaluate how data preprocessing choices, feature design, and regression models influence the prediction accuracy.

This paper is organized as follows. Section 2 introduces the datasets. Section 3 describes the methodologies of this study, including data analysis, preprocessing pipeline, and feature engineering procedures. Section 4 outlines the experimental setup, section 5 discusses results of the experiments, and finally section 6 concludes the study.

2. Data Description

The training data provided for this challenge comprises flight list information, flight trajectory data, fuel interval records, and airport data. In addition, we incorporate open-source ECMWF ERA5 weather data [7] and the OpenAP aircraft performance database [3].

2.1 Challenge Data

Flight list. The flight list dataset contains aircraft type codes, departure and arrival times, and origin and destination airport identifiers. In total, 11,037 worldwide flights spanning 26 distinct aircraft types are provided, covering the period from April to September 2025.

Flight trajectories. The flight trajectories consist of ADS-B recordings enhanced with ACARS data. Each record contains the aircraft type code, position, track, vertical rate, and groundspeed derived from ADS-B, supplemented by occasional Mach number, TAS, or CAS values from ACARS, all provided per timestamp. The trajectory data are generally noisy and contain numerous outliers.

Fuel intervals. The fuel interval records are extracted from ACARS messages. Each entry specifies the start and end time of a fuel measurement interval and the corresponding fuel consumed within that flight segment. The dataset contains 131,530 fuel intervals, with durations ranging from 1.61 s to 1 hour and fuel consumption values between 0.453 and 32,205 kg.

Airport information. The airport dataset contains the geographic position and elevation of each airport.

2.2 Supplementary Open-Source Data

Wind data. As supplementary meteorological input, we use ECMWF ERA5 reanalysis data [7], sourced via the *fastmeteo* library [8]. ERA5 provides worldwide atmospheric fields on pressure levels ranging from 1 hPa to 1000 hPa at a horizontal grid resolution of 0.25°. In this study, we use wind information interpolated to the timestamp and position of each trajectory point.

Performance database. We additionally employ the OpenAP aircraft performance library [3], which supports 36 aircraft types. Specifically, we use OpenAP's phase detection module, which applies a type-agnostic fuzzy-logic classifier, and its fuel flow model, which draws on aircraft type-specific parameters. In addition, OpenAP provides convenient atmospheric modeling tools that are used to convert between true airspeed (TAS), calibrated airspeed (CAS), and Mach number.

3. Methods

This section describes the methodological framework used to process the raw flight data and develop the fuel prediction model. The workflow consists of four main steps: exploratory data analysis, data preprocessing and feature construction, aircraft mass estimation, and regression model selection.

First, an exploratory analysis is performed to investigate the spatial and temporal characteristics of the ADS-B and ACARS observations and to identify potential data limitations in Section 3.1. The insights obtained from this analysis guide the design of a preprocessing pipeline aimed at removing noise, filtering unrealistic observations, and preparing the dataset for model training. This is explained in Section 3.2, together with the construction of additional physical features. Subsequently, estimations of the aircraft mass along the trajectory are incorporated in Section 3.3. Finally, the processed dataset is used to train the selected regression model (Section 3.4).

3.1 Exploratory Data Analysis

Figure 1 shows the spatial distribution of data points for 1000 randomly sampled trajectories. As illustrated, the flights span a nearly global range, with the majority of observations originating from Europe and North America. The temporal sampling intervals are not consistent across the dataset. This variability is primarily due to differences in data transmission frequency and receiver coverage. ADS-B transmissions are automatic and frequent; however, their coverage is largely limited to populated land areas. In contrast, ACARS messages are transmitted less frequently but are also available over oceanic regions.

To further examine the characteristics of the dataset, we consider an example trajectory. Figure 2 presents several variables from the training dataset plotted against time. ACARS data are shown in red, while ADS-B observations are shown in blue. The take-off and landing points, obtained from airport coordinates, are indicated by green triangular markers. The shaded regions represent fuel segments with the known amount of fuel consumed within the segment. As can be seen, these

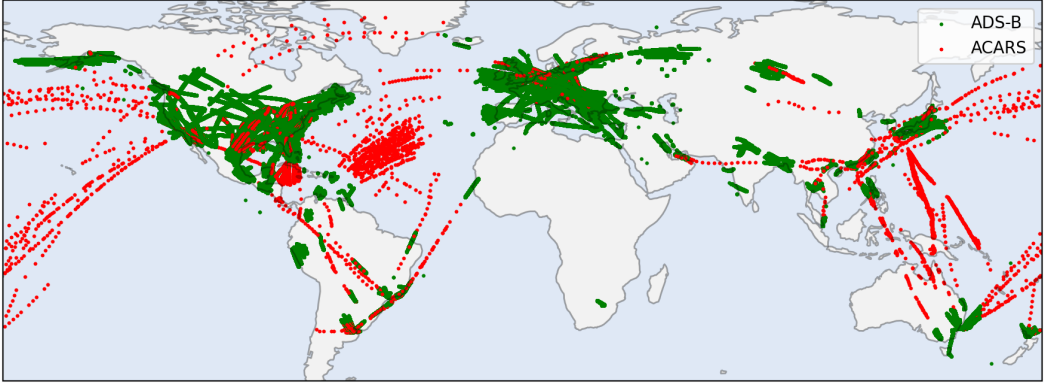


Figure 1. Scatter plot of ACARS and ADS-B data points for 1000 randomly sampled flights from the training set.

segments appear between ACARS message intervals because the fuel consumption information is derived from ACARS reports.

Several important observations can be drawn from this example. First, some fuel segments associated with the considered flight occur after the recorded landing time. Similarly, certain flights contain fuel segments that occur before the recorded take-off time. Consequently, some segments fall outside the actual flight duration and therefore lack corresponding ADS-B trajectory information, providing only segment duration and fuel consumption data for model training.

Another observation is that the duration of the fuel reporting windows varies considerably, as already shown in Section 2.1. As a result, the reported total fuel consumption per segment can vary significantly. When the reporting interval is large, a single fuel segment may span multiple flight phases (e.g., cruise and descent), as illustrated by the altitude profile in Figure 2. This is particularly relevant because fuel flow rates differ substantially between flight phases.

Furthermore, misalignment between ACARS and ADS-B recordings can be observed in several cases, which introduces additional uncertainty in the data. Inspection of the longitude and latitude profiles also reveals occasional outliers, such as isolated ADS-B points located far from the actual trajectory. In general, the trajectory data contain measurement noise and incomplete coverage; for instance, the beginning of the considered flight is missing in the example shown.

Figure 3a shows a log-scale histogram of total fuel consumed per fuel segment. As expected, the distribution is very wide due to the varying duration of the reporting windows. The minimum recorded value is below 0.5 kg, the median value is approximately 200 kg, and the maximum exceeds 30 000 kg. Such a wide range complicates subsequent data preparation and statistical model training.

To address this issue, an average fuel flow rate is assigned to each segment by dividing the total fuel consumed by the segment duration. The resulting distribution is shown in Figure 3b. Compared to total fuel consumption, the fuel flow rate distribution is considerably narrower, with a minimum value below 0.01 kg/s and a maximum value of approximately 16.5 kg/s. Both extreme values are physically unrealistic for commercial passenger aircraft and likely represent data artifacts or reporting errors.

The fuel flow rate representation also facilitates the identification and filtering of noisy observa-

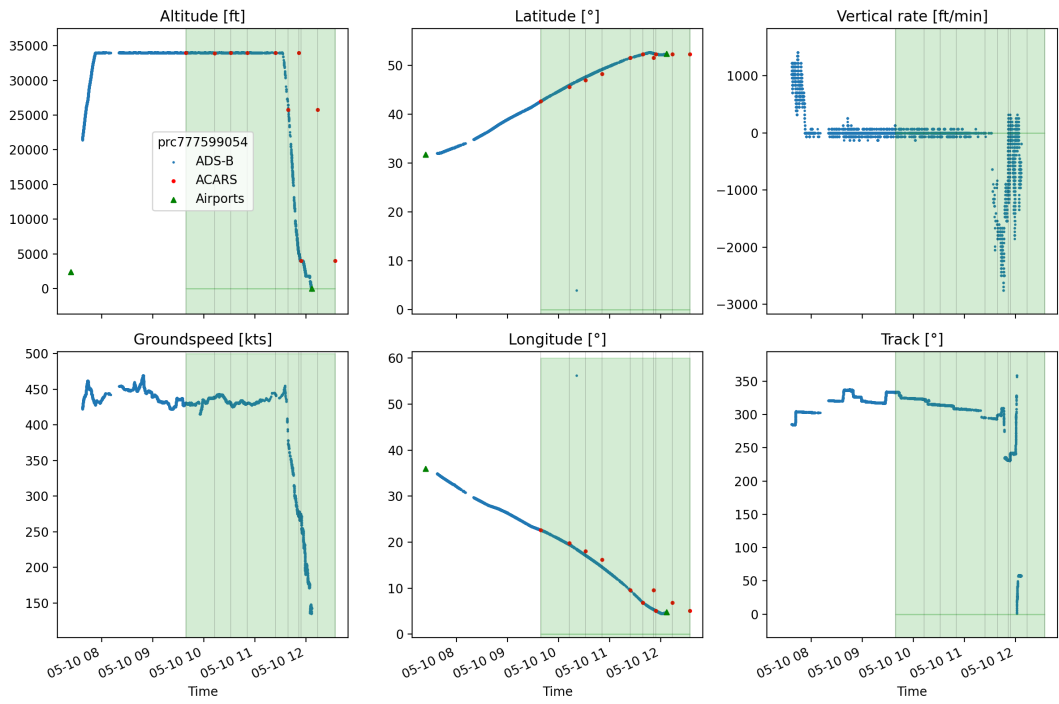


Figure 2. An example of time-series profiles of flight parameters including altitude, latitude, longitude, vertical rate, speed, and track. Green arrows mark airport arrival and departure points. Green shaded regions highlight the fuel segments.

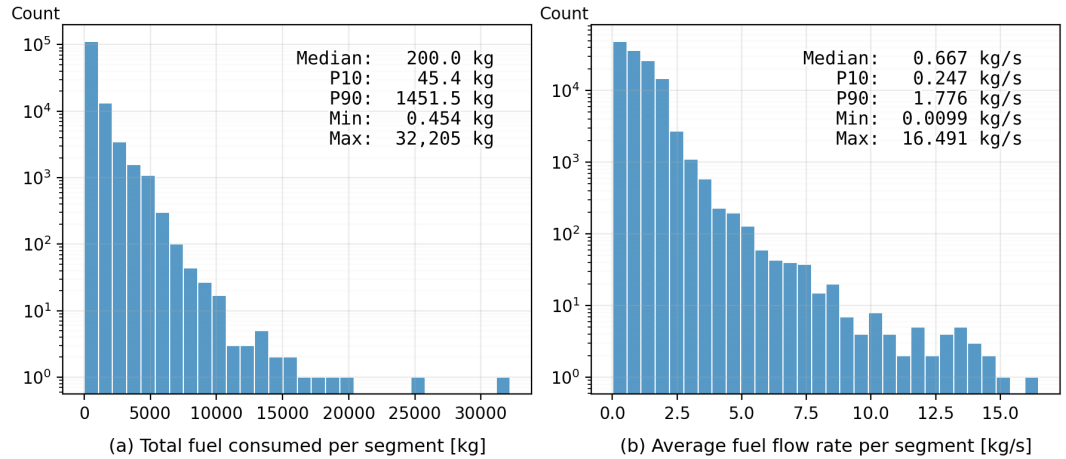


Figure 3. Log-scale histograms of (a) fuel consumption and (b) fuel flow rate in the training dataset.

tions by applying physically meaningful bounds. For example, Baklacioglu [9] reported minimum and maximum fuel flow rates for different flight phases based on Flight Data Recorder (FDR) measurements from B738 aircraft operating in Turkey in 2009. The reported bounds across all phases range between 0.04 kg/s and 1.84 kg/s (in the paper 146-6621 kg/h). In addition, the ICAO Aircraft Engine Emissions Databank [10] provides fuel flow rates measured across various flight phases for emissions certification. The reported fuel flow rates range from 0.023 kg/s to 4.69 kg/s per engine. These reference values are used as extreme-case bounds rather than literal thresholds. Baklacioglu's measurements are drawn from a single aircraft type (B738), and the ICAO Aircraft Engine Emissions Databank reports the largest per-engine fuel-flow values across the full range of engine types it covers. Moreover, the highest ICAO values are likely to originate from instantaneous fuel rates during the brief initial take-off thrust setting, which is not representative of the average fuel flow over a full segment as used in this study. Since neither source directly matches our fleet composition or our segment-averaged target, we treated their reported ranges as outer bounds on plausible fuel flow and selected filtering thresholds within that range that retain the largest share of usable training data. Segments with a fuel flow label outside these thresholds are removed from the training set. The upper bound of 6.5 kg/s was chosen on this basis as a reasonable intermediate value rather than a value taken directly from either reference.

Aircraft performance varies significantly across different aircraft types. Table 1 summarizes the number of flights and fuel segments available for each aircraft type in the dataset. Out of the 26 aircraft types present in the data, nine have fewer than ten available training flights, which may be insufficient for building reliable statistical models.

Table 1. Number of flights and fuel segments in the training dataset per flight

	Typecode	Number of flights	Number of fuel segments		Typecode	Number of flights	Number of fuel segments
1	A20N	3,220	40,488	14	B737	29	221
2	A21N	393	3,138	15	B738	551	7,457
3	A306	10	75	16	B739	28	180
4	A318	1	4	17	B744	90	703
5	A319	19	251	18	B748	3	17
6	A320	3,023	31,980	19	B752	3	28
7	A321	285	2,572	20	B763	2	20
8	A332	532	5,441	21	B772	103	858
9	A333	156	1,521	22	B77L	2	17
10	A359	1,571	23,942	23	B77W	183	1,762
11	A388	4	9	24	B788	583	7,469
12	B38M	20	195	25	B789	224	3,173
13	B39M	1	2	26	MD11	1	7

3.1.1 Data Limitations

Several limitations were identified in the dataset:

- **Limited trajectory coverage.** The spatial coverage of ADS-B and ACARS observations is incomplete, and portions of some trajectories are missing.
- **Noisy measurements.** ADS-B receivers are subject to encoding and transmission errors, which may introduce noise or spurious observations.
- **Unrealistic fuel consumption reports.** Some reported fuel values are physically implausible

and likely represent reporting errors.

- **Fuel segments outside flight duration.** Certain fuel segments occur before take-off or after landing, preventing the use of trajectory information for those intervals.
- **Limited data for some aircraft types.** For several aircraft types, the number of available flights is small, which may introduce bias or reduce model accuracy.
- **Additional reporting artifacts.** The dataset occasionally contains inconsistencies such as identical fuel consumption values reported over multiple segments with varying duration.

These limitations may affect the accuracy of full-flight trajectory reconstruction and take-off mass estimation. Furthermore, training statistical models on noisy or incomplete data may degrade model performance and introduce additional uncertainty into the results.

3.2 Data Processing and Physical Feature Construction

In the aforementioned Section 3.1, the raw data was intermittent, noisy, and not ready for direct integration with model training. We devised several preprocessing procedures to address these issues to enable the models to learn from meaningful features.

The preprocessing procedures are composed into different pipelines, made possible by utilizing the Traffic library and our self developed extension to the library to allow fast iterations on the feature engineering on the trajectory data. Each pipeline can be separated into four general steps shown in Figure 4.

The first step is filtering and resampling. We noticed there are duplicated records in the dataset as well as observations falling outside of the physically plausible range. These data points are removed using heuristics. Most of the data is irregularly sampled with noisy measurements. Therefore, a cascaded above-median filter with a sliding window is used for filtering different features incorporating feature-specific kernels. Its implementation is included with the Traffic library, however, we discovered that the default altitude filter kernel is too aggressive so the values were adjusted accordingly. A resampling strategy is then applied to the filtered values every one second. The positional features of a trajectory are first projected onto a local flat-Earth approximation then linearly interpolated. This approach helps to better preserve the great-circle nature of trajectories on long haul flights compared to computing on the latitude and longitude directly. For the other features, we used simple linear interpolations to resample the data points.

The second step is carried out with OpenAP-based phase identification. It utilizes a fuzzy-logic approach to identify the resampled trajectory points in one of the 6 phases: on-ground, climb, descend, level flight, cruise, and unlabeled. We noticed that all flight trajectories are already trimmed to not include on-ground segments, and many descend phases during the approach to destination airport are identified as 'unlabeled'. Nonetheless, these categorical features are used in the final model prediction. Additionally, the absolute date and time of the flight are converted to timestamps in seconds, with zero marking the first available data point in the trajectory instead of the takeoff time. A cumulative distance is also calculated based on the recorded latitude and longitude at each trajectory point. These features are mostly utilized for their semantic embedding in the trajectory to facilitate the model prediction.

The third step is augmenting the speed features in the trajectories. Fuel flow rate is affected by many parameters during a flight, but TAS is one of the factors that carries a great impact from our experiences analyzing the quick access recorder data. Although ACARS data provides the crucial CAS, TAS, and mach number, they are never available for all the trajectories. The airspeed features are converted to one another based on availability. If none of the features are present in the trajectories, groundspeed is used to substitute the TAS. The TAS feature recovery is also attempted from a

different direction incorporating the meteorological data. After pre-synchronizing ERA5 data using fastmeteo [11], the horizontal wind components (u , v) were interpolated for each ADS-B data point on the ERA5 meteorological data grid. The TAS was then analytically computed from the reported ground speed by correcting for the interpolated wind vector.

The last major step is inferring mass related features from the trajectory data. Mass is an important feature for fuel burn prediction, however, it is also the only feature that is not directly observed in any data source. This is the step which we believe is the deciding factor that allowed us to score consistently lower RMSE compared to the other teams. More details will be explained in the next section.



Figure 4. General data preprocessing pipeline.

3.3 Aircraft mass estimation

Mass estimation is required to compute fuel flow along the trajectory. At the base lies the estimation of the take-off weight (TOW), which can subsequently be used to estimate the mass along the flight’s trajectory. We explored two approaches: (i) a data-driven regression model with a recursive mass estimation, and (ii) a deterministic heuristic approximation.

3.3.1 Data-driven regression model with recursive mass

For the data-driven approach, a LightGBM regression model was trained using the PRC 2024 TOW Estimation Challenge dataset [12]. On internal train/test splits, the model achieves an RMSE of approximately 3000 kg. It should be noted that the training data consist of European flights from 2024 and do not cover all aircraft types considered in this work. Missing type coverage may introduce additional uncertainty in both TOW and subsequent fuel flow estimation.

The LightGBM model predicts only the initial TOW. Subsequent mass values along the trajectory are obtained via integration using fuel flow estimates from OpenAP [3]. It depends on the current mass, and the mass at each timestamp is updated by subtracting accumulated fuel consumption:

$$m_n = m_0 - \sum_{i=0}^n \dot{m}_i \cdot \Delta t_i, \quad \forall n \in \{0, 1, \dots, k\}, \quad (1)$$

where m_n is the aircraft mass at timestamp n , m_0 corresponds to TOW, $\dot{m}_i$ is the fuel flow in kg/s, and Δt_i is the time interval between consecutive timestamps. In practice, five iterations were found sufficient for convergence to stable mass values.

However, in certain flights, abnormal ADS-B data, particularly near take-off and landing, caused numerical instabilities in the initial mass estimates. In such cases, the recursive integration accumulated errors, occasionally leading to physically impossible negative mass values.

3.3.2 Heuristic Mass

To mitigate instability in the recursive approach, we implemented a deterministic heuristic mass model. Instead of relying on fuel flow integration, aircraft mass is approximated as a linear interpolation between assumed take-off and landing weights, which are based on maximum take-off weight (MTOW), maximum landing weight (MLW) and operational empty weight (OEW). Using aircraft properties from the OpenAP *Prop* module [3], we defined:

$$m_{\text{TO}} = 0.8 \cdot \text{MTOW}, \quad m_{\text{LD}} = 0.5 \cdot (\text{MLW} + \text{OEI}). \quad (2)$$

Let t_{TO} and t_{LD} denote the take-off and landing timestamps. The fraction of flight completed at timestamp t_n is:

$$p_n = \frac{t_n - t_{\text{TO}}}{t_{\text{LD}} - t_{\text{TO}}}, \quad (3)$$

clipped to $[0, 1]$ to avoid extrapolation. The mass estimate is then:

$$m_n = m_{\text{TO}} - p_n \cdot (m_{\text{TO}} - m_{\text{LD}}), \quad \forall n \in \{0, 1, \dots, k\}. \quad (4)$$

Although less physically detailed than the recursive approach, this method guarantees monotonic mass decrease and prevents numerical instability. While it provides a less accurate approximation of TOW, it avoids negative mass artifacts.

3.4 Model selection

While the primary emphasis of this work is on data processing and feature engineering, these need to be included into a model for predicting fuel consumption. The data structure follows a tabular structure and on such data tree-based models show good performance at a relatively low computational cost [13]. Specifically gradient-boosted decision trees can be utilized here, and this was implemented with LightGBM [14], which is a model also used in the previous edition of the PRC Data challenge [15]. It is well suited to our setting for three reasons. First, the computational speed of this model is high, due to the histogram-based split finding, making it easier to test with different feature sets. Second, also the computational memory needed is low. Third, the native handling of missing values (*NaN*), which is favorable given the irregular and partially missing nature of ADS-B-derived features.

In general, gradient boosting decision tree (GBDT) is a framework that learns an additive ensemble of regression trees by sequentially fitting each new tree to the negative gradient of the loss function with respect to the current model prediction. The LightGBM model is composed of two techniques, being Gradient-based One-Side Sampling and Exclusive Feature Bundling. The former enables the model to handle large datasets and the latter working with many features in an efficient and accurate way. [14] This makes LightGBM a fast gradient boosting method and therefore very useful for our case.

4. Experimental Setup

This section describes the experimental configuration used to evaluate the proposed fuel flow estimation pipeline. The evaluation follows a structured design organized into a preliminary filtering experiment and four feature-set blocks. Each block isolates a specific factor that influences prediction accuracy: the composition of kinematic and airspeed features (Block 1), the addition of mass-related features derived from recursive estimation (Block 2), the substitution of heuristic mass features (Block 3), and the choice of training target (Block 4). All experiments share a common preprocessing pipeline, data split, and evaluation metric to ensure that observed performance differences are attributable solely to the experimental variable under consideration. The hyperparameters, feature groups, block definitions, and evaluation criterion are detailed in the subsections below, and the outcomes of these experiments are reported and interpreted in Section 5.

4.1 Dataset Split

Model performance is reported on the official challenge ranking set, which serves as the test set. The remaining labelled data are used for training, with a portion held out as a validation split for early stopping and preliminary hyperparameter tuning. The split into training and validation subsets is performed at the level of individual flights rather than individual segments. Because segments belonging to the same flight are strongly correlated, a naive random split over segments would place segments of the same flight in both subsets, leaking information and yielding an optimistic estimate. To prevent this, the data are sorted first by flight and then by time, and the split is taken on this ordering 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.

4.2 Hyperparameters

Unless otherwise noted, all models use the baseline LightGBM hyperparameters listed in Table 2. These values were selected through preliminary tuning on a validation split and remain fixed across all experimental blocks to ensure that observed differences in performance are attributable to feature composition rather than model configuration.

Table 2. Baseline LightGBM parameters used across all experimental blocks unless stated otherwise.

Parameter	Value
n_estimators	5000
learning_rate	0.01
subsample	0.8
colsample_bytree	0.8
reg_lambda	0.5
reg_alpha	0.1
metric	rmse
random_state	46

4.3 Fuel Flow Filtering

As discussed in Section 3.1, the distribution of reported fuel consumption per segment spans several orders of magnitude, ranging from below 0.5 kg to over 30,000 kg. This spread reflects the heterogeneity of segment durations in the dataset and makes direct modelling of fuel consumption susceptible to the influence of extreme values. Converting the target to average fuel flow (kg/s), defined as the fuel consumption divided by the segment duration, produces a considerably narrower distribution (Figure 3), although extreme values persist. Inspection of the fuel flow histogram reveals observations below 0.01 kg/s and above 16.4 kg/s, both of which are physically implausible for conventional passenger aircraft and are likely attributable to data artifacts, timestamp irregularities, or reporting errors.

To mitigate the influence of these spurious values on the regression loss, a heuristic filter is applied that removes fuel flow labels outside a physically plausible operating range prior to model training. Specifically, any fuel segment whose fuel flow label y_i (kg/s) falls below a lower bound l or above an upper bound u is excluded from the training set:

$$\mathcal{D}_{\text{train}} = \{i : l \leq y_i \leq u\}, \quad (5)$$

where l and u are chosen to exclude values that fall outside the operational envelope of the aircraft types in the dataset while retaining the vast majority of valid observations. The selected bounds of $l = 0.05$ kg/s and $u = 6.5$ kg/s, correspond to the assumed minimum descent fuel flow of a small narrow-body aircraft and the assumed maximum climb fuel flow of a large wide-body aircraft, respectively.

To evaluate the sensitivity of model performance to this filtering strategy, three filter configurations are compared using a fixed feature set and fixed hyperparameters. The feature set used for this comparison is the benchmark configuration (ADS-B + ACARS + Vrate Split, 41 features as defined in Table 3), trained with the baseline hyperparameters in Table 2. This choice isolates the effect of the filter from any variation in feature composition or model tuning. The sensitivity of these filter configurations is reported in Section 5.1, and the resulting filter is applied consistently to all subsequent feature-set experiments.

4.4 Overview of Feature-Set Experiments

A central question in this study is which combination of data sources and derived features yields the most accurate fuel flow prediction. Because the engineered features originate from fundamentally different sources, including ADS-B trajectory data, ACARS messages, ERA5 meteorological reanalysis, and physics-based mass models, their individual and joint contributions are not known *a priori*. A structured set of experiments is therefore designed to systematically vary the included feature groups while holding all other factors constant, including the fuel flow filter established in Section 4.3.

The full set of candidate features is partitioned into eight non-overlapping groups, each corresponding to a distinct data source or derived quantity. The composition of each group is summarized in Table 3 and described below.

Table 3. Overview of feature groups used in the ablation study. Feature counts refer to the number of scalar input variables per segment after statistical embedding.

Group	Source	Count	Variables
ADS-B	Trajectory data	15	Segment duration and distance, flight duration and distance, flight phase, altitude mean, ground speed (mean, std), track angle (mean, std), vertical rate (min, max, mean, std), aircraft type
ACARS	ACARS data	6	True airspeed (mean, std), calibrated airspeed (mean, std), Mach number (mean, std)
Wind	ERA5 reanalysis	6	Wind-corrected TAS (mean, std), CAS (mean, std), Mach (mean, std)
Vrate Split	Trajectory data	20	Vertical-rate mean and std for each of 10 equal-duration sub-intervals per segment
MTOW + OEW	OpenAP database	2	Maximum take-off weight, operational empty weight
TOW Est	Calculated	1	Estimated take-off weight
FF Est	Derived (Sec. 3.3.1)	2	Estimated fuel flow along the trajectory (mean, std)
Mass Est	Derived (Sec. 3.3.2)	2	Estimated aircraft mass along the trajectory (mean, std)

The first group, *ADS-B*, comprises 15 features derived directly from the Automatic Dependent Surveillance - Broadcast trajectory data. These include segment-level context variables, such as segment duration, segment distance, flight phase (categorical, from the first data point in the segment), and

aircraft type (categorical), as well as full-flight context variables such as total flight duration and great-circle distance. The remaining ADS-B features are kinematic embeddings computed over the segment time window: mean altitude, mean and standard deviation of ground speed and track angle, and the minimum, maximum, mean, and standard deviation of vertical rate.

The second group, *ACARS*, contains six features: the mean and standard deviation of true airspeed (TAS), calibrated airspeed (CAS), and Mach number. Since ACARS availability is sparse and not all speed variables are consistently reported, missing values are inferred using OpenAP conversion utilities as described in Section 2.2.

The third group, *Wind*, consists of six features derived from ERA5 reanalysis data. For each trajectory point, horizontally interpolated wind components are used to compute a wind-corrected true airspeed from the reported ground speed and track angle. From this corrected TAS, consistent CAS and Mach values are derived using Standard Atmospheric model. The resulting time series are summarized as mean and standard deviation over the segment. These wind-augmented features provide an alternative airspeed characterization that does not depend on ACARS availability.

The fourth group, *Vrate Split*, provides a higher-resolution representation of vertical dynamics within each segment. Each fuel segment is divided into 10 equal-duration sub-intervals, and the mean and standard deviation of the vertical rate are computed for each sub-interval, yielding 20 features per segment. This decomposition captures transitions between flight phases (e.g., climb-to-cruise) that a single segment-level feature would obscure. The split is applied only to the vertical rate, as each additional variable would contribute another 20 features per segment. We therefore restrict this higher-resolution representation to the vertical rate, which is the most informative indicator of within-segment phase transitions. The number of sub-intervals is set to ten as a heuristic compromise between temporal resolution and feature count, and this value was not tuned systematically.

The remaining four groups capture mass-related information. *MTOW + OEW* includes the maximum take-off weight and operational empty weight obtained from the OpenAP database for each aircraft type. *TOW Est* is the estimated actual take-off weight, obtained from an external LightGBM model trained on the PRC 2024 TOW Estimation Challenge dataset [16]. *FF Est* and *Mass Est* contain the mean and standard deviation of the estimated fuel flow and mass time series, respectively, computed along the trajectory. For these last two groups, two estimation strategies are available, recursive calculation and heuristic interpolation, as described in Sections 3.3.1 and 3.3.2. When a given experimental configuration specifies one of these strategies, both the FF Est and Mass Est features are computed using that same approach.

4.5 Feature-Combination Design

Building on the feature groups defined above, the experiments are organized into four blocks, each designed to isolate a specific experimental variable. All configurations within and across blocks share the same preprocessing pipeline, data-filtering procedure (Section 4.3), and evaluation metric to ensure comparability. Unless otherwise noted, models use the baseline hyperparameters listed in Table 2, with `n_estimators` set to 5,000 and early stopping on a validation split.

Block 1: Baseline without mass features. The first block establishes baseline performance using only trajectory-derived and airspeed-related features, without any mass-related information. The benchmark model combines ADS-B kinematic features (15 variables), ACARS airspeed features (6 variables), and segment-wise vertical-rate split features (20 variables), totaling 41 input features trained on the filtered fuel flow target (kg/s). This benchmark coincides with the model used in the fuel flow filter evaluation (Section 4.3), thereby providing a consistent reference point across both sets of experiments.

Starting from this benchmark, the marginal contribution of individual feature groups is evaluated by systematically adding or removing them. Specifically, the experiments test whether wind-augmented airspeed variables provide complementary information beyond ACARS, and whether ACARS features, despite their sparsity, carry information not captured by wind-corrected alternatives. Additional configurations isolate the effect of the vertical-rate split by comparing models with and without it, and assess a minimal model using only ADS-B kinematic features.

Block 2: Mass features via recursive estimation. The second block extends the benchmark by incorporating mass-related features computed using the recursive mass-estimation approach (Section 3.3.1). The full configuration includes all eight feature groups: ADS-B (15), ACARS (6), Wind (6), Vrate Split (20), MTOW + OEW (2), TOW Est (1), FF Est (2), and Mass Est (2), totaling 54 input features.

To quantify the incremental contribution of each mass-related component, a top-down ablation is performed by successively removing components from the full set. The estimated mass statistics (Mass Est) are removed first, then both the estimated mass and fuel flow statistics (FF Est + Mass Est), and finally the estimated take-off weight as well (TOW Est + FF Est + Mass Est), reducing the mass-related inputs to only the aircraft-type constants MTOW and OEW. In addition, configurations that retain the full mass feature set but remove either ACARS or Wind features are tested, to assess potential redundancy between airspeed sources when mass information is available.

Block 3: Mass features via heuristic estimation. The third block replaces the recursive mass-estimation features with those derived from the heuristic interpolation approach (Section 3.3.2). This substitution directly tests whether the simpler, more numerically stable heuristic mass model produces features that are more useful for regression, despite being less physically detailed. The heuristic approach approximates aircraft mass as a linear interpolation between assumed take-off and landing weights, avoiding the error-accumulation issues inherent in recursive fuel flow integration.

Three configurations are evaluated, mirroring the most informative comparisons from Block 2: the full feature set with heuristic mass, and two reduced variants that remove either ACARS or Wind features. For this block, $n_{\text{estimators}}$ is increased to 7,500, as early stopping on validation RMSE indicated that additional boosting iterations improved convergence for this feature configuration.

Block 4: Training on fuel consumption. The fourth block examines the effect of the learning target. Instead of training on average fuel flow (kg/s), models are trained directly on segment-level fuel consumption (kg), while applying the same fuel flow filter (0.05–6.5 kg/s) to remove out-of-range segments prior to conversion to fuel mass. This block uses the recursive mass features and tests three configurations with different features availability.

These experiments evaluate whether the fuel flow representation, which normalizes for segment duration and produces a tighter and less skewed target distribution, yields a more effective training signal than raw fuel consumption. All three configurations in this block use the baseline hyperparameters with $n_{\text{estimators}}$ set to 5,000.

4.6 Evaluation Metric

Performance across all experiments is evaluated using the root mean squared error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}, \quad (6)$$

where N is the number of fuel segments, y_i the ground-truth label, and $\hat{y}_i$ the model prediction. For Blocks 1–3, both y_i and $\hat{y}_i$ are expressed in fuel flow units (kg/s); for Block 4, they are expressed in fuel consumption units (kg). In both cases, the final ranking RMSE is computed on the fuel-consumption scale (kg) to ensure comparability across blocks. RMSE penalizes large deviations more strongly than mean absolute error (MAE), which is desirable for accurately modeling high fuel flow regimes. Moreover, RMSE is the official evaluation metric of the challenge, ensuring consistency between training optimization and scoring.

Having established the experimental framework, including the filtering procedure, the four feature-set blocks, and the evaluation criterion, the following section reports the outcomes of each experiment and interprets their implications for fuel consumption estimation.

5. Results and Analysis

This section presents the results of the experiments defined in Section 4. The section follows the same structure as the Experimental Setup (Section 4): the fuel flow filter sensitivity is examined first (Section 5.1), followed by the four feature-set blocks (Section 5.2). For each block, the main findings are reported and then interpreted in light of the underlying physical and methodological considerations. A synthesis of the overall findings, together with a discussion of limitations and directions for future work, concludes the section. All RMSE values reported below are computed on the official ranking set, which serves as the primary evaluation criterion.

5.1 Fuel Flow Filter Sensitivity

Table 4. Fuel flow filter sensitivity on the benchmark feature set (ADS-B + ACARS + Vrate Split, 41 features).

Lower Limit	Upper Limit	RMSE
-1	1000	228.05
-1	6.5	205.00
0.05	6.5	205.85

As described in Section 4.3, three filter configurations are compared using the benchmark feature set and baseline hyperparameters. Table 4 reports the results. When no meaningful filtering is applied (bounds of -1 to 1000 kg/s), the RMSE increases substantially to 228.05 kg, compared with 205.85 kg for the selected filter (0.05 – 6.5 kg/s). Retaining only the upper bound while removing the lower limit results in a marginally lower RMSE of 205.00 kg.

These results indicate that the upper-bound filter provides the dominant improvement, as it removes the influence of extreme high fuel flow values that are physically implausible. The lower bound contributes a smaller, and slightly negative, marginal effect in this configuration, which suggests that very low fuel flow values introduce less distortion to the regression loss than their high-value counterparts. The filter-induced improvement of approximately 22 RMSE points over the unfiltered case confirms that target-space outlier removal is a necessary preprocessing step for this regression

Table 5. Model configuration and corresponding RMSE across all four experimental blocks. Checkmarks indicate the feature groups included in each configuration.

Description	ADS-B	ACARS	Wind	Vrate Split	MTOW + OEW	TOW Est	FF Est	Mass Est	RMSE
Block 1: Fuel flow [kg/s], no Mass									
Benchmark	✓	✓		✓					205.85
Benchmark + Wind	✓	✓	✓	✓					205.25
ADSB + Wind + Split	✓		✓	✓					212.80
ADSB only	✓								212.67
ADSB + Split	✓			✓					211.42
ADSB + ACARS	✓	✓							210.71
Block 2: Fuel flow [kg/s], with Recursive Mass									
Full features	✓	✓	✓	✓	✓	✓	✓	✓	203.75
No Mass Est	✓	✓	✓	✓	✓	✓	✓		206.22
No FF + Mass Est	✓	✓	✓	✓	✓	✓			206.06
No TOW/FF/Mass Est	✓	✓	✓	✓	✓				205.95
No ACARS	✓		✓	✓	✓	✓	✓	✓	206.48
No Wind	✓	✓		✓	✓	✓	✓	✓	203.09
Block 3: Fuel flow [kg/s], with Heuristic Mass									
Full features	✓	✓	✓	✓	✓	✓	✓	✓	201.04
No ACARS	✓		✓	✓	✓	✓	✓	✓	208.71
No Wind	✓	✓		✓	✓	✓	✓	✓	199.91
Block 4: Fuel estimate [kg], with Recursive Mass									
Full features	✓	✓	✓	✓	✓	✓	✓	✓	220.56
ADSB + ACARS	✓	✓							219.13
ADSB + ACARS + Split	✓	✓		✓					218.47

task. Nevertheless, the filter with both bounds is retained for all subsequent experiments, as it provides a physically motivated and conservative operating envelope that excludes values outside the plausible range of conventional passenger aircraft.

With this filter established, the following subsections report the results of the four feature-set blocks.

5.2 Feature-Set Experiments

Table 5 summarizes the RMSE for all feature-set configurations across the four experimental blocks defined in Section 4.5. The results are organized by block to facilitate direct comparison within each experimental group. The subsections below discuss the findings for each block individually.

5.2.1 Block 1: Baseline without Mass Features

The first block, as defined in Section 4.5, evaluates six configurations using only trajectory-derived and airspeed-related features. The benchmark model (ADS-B + ACARS + Vrate Split) achieves an RMSE of 205.85 kg. Adding wind-augmented airspeed features to this benchmark yields a marginal improvement to 205.25 kg. However, replacing ACARS with wind features (ADS-B + Wind + Split) leads to a notable increase in RMSE to 212.80 kg, a value comparable to the ADS-B-only configuration (212.67 kg). This result indicates that wind-corrected airspeed features do not serve as an effective

substitute for ACARS-derived variables.

Among the reduced configurations, models that include ACARS consistently outperform those that do not. The ADS-B + ACARS model (210.71 kg) achieves a lower RMSE than either ADS-B + Split (211.42 kg) or ADS-B only (212.67 kg), which suggests that ACARS airspeed features carry predictive information that is complementary to trajectory-derived kinematics. The vertical-rate split provides a further improvement when combined with ACARS, as evidenced by the reduction from 210.71 kg (ADS-B + ACARS) to 205.85 kg (Benchmark).

These findings suggest that the benchmark combination of ADS-B, ACARS, and Vrate Split represents the most effective baseline configuration for fuel flow prediction in the absence of mass-related features. ACARS data, despite their sparsity, provide airspeed measurements that are not fully recoverable from wind-corrected ground speed alone. One possible explanation is that the ERA5 wind fields, which are interpolated from a coarse reanalysis grid, do not capture local wind conditions with sufficient accuracy to replicate the information content of directly measured airspeed. In contrast, the vertical-rate split captures within-segment transitions between flight phases (e.g., climb-to-cruise), and this additional resolution translates into a measurable reduction in prediction error when combined with reliable airspeed data.

5.2.2 Block 2: Mass Features via Recursive Estimation

The second block extends the benchmark by incorporating mass-related features derived from the recursive mass-estimation approach described in Section 3.3.1. The full recursive-mass configuration (53 features) achieves an RMSE of 203.75 kg, representing an improvement of approximately 2 points over the best Block 1 result (205.25 kg).

The top-down ablation reveals a gradual degradation as mass-related components are removed. Removing Mass Est alone increases the RMSE to 206.22 kg, while removing both FF Est and Mass Est yields a similar value of 206.06 kg. Removing all estimated mass components (TOW Est, FF Est, and Mass Est) results in an RMSE of 205.95 kg, which is comparable to the Block 1 benchmark. This pattern indicates that the mass-related features contribute collectively. The differences between these configurations are, however, small enough that the ranking among the individual components should be interpreted with caution (Section 6).

An informative result emerges from the airspeed-source comparison within this block. The model that retains ACARS but removes Wind features achieves the lowest RMSE in this block (203.09 kg), marginally below the full 53-feature configuration (203.75 kg). In contrast, removing ACARS while retaining Wind increases the RMSE to 206.48 kg. Given the small size of the former difference, it may partly reflect stochastic variation (Section 6) rather than a robust effect. The consistent direction across blocks is nonetheless compatible with wind-augmented features introducing some noise when combined with mass-related information and ACARS. The result also reinforces the Block 1 observation that ACARS provides more informative airspeed characterization than wind-corrected alternatives.

Taken together, these results indicate that mass information, even when derived through the recursive approach, provides a meaningful improvement over kinematic and airspeed features alone. However, the benefit is moderate (approximately 2–3 RMSE points) and sensitive to the composition of the remaining feature set. As discussed in Section 3.3.1, the recursive mass approach is constrained by error accumulation along the trajectory, which may limit the quality of the derived features and therefore constrain the achievable improvement.

5.2.3 Block 3: Mass Features via Heuristic Estimation

The third block, as described in Section 4.5, replaces recursive mass features with those derived from the heuristic interpolation approach. The full heuristic-mass configuration achieves an RMSE of 201.04 kg, which is lower than all Block 2 configurations. The best result in this block, an RMSE of 199.91 kg, is obtained when wind features are removed while ACARS is retained.

The sensitivity to airspeed-source removal is more pronounced in this block than in Block 2. Removing ACARS increases the RMSE to 208.71 kg, a degradation of approximately 7.7 points relative to the full model. In contrast, removing wind features improves the RMSE by approximately 1.1 points. This pattern is consistent with the findings from Blocks 1 and 2, and further confirms that wind-augmented features introduce more noise than useful signal when mass information and ACARS are already present.

These results indicate that the heuristic mass-estimation approach produces features that are more useful for the regression model than those derived from the recursive method. This finding is notable because the heuristic approach is physically less detailed: it approximates aircraft mass as a smooth linear interpolation between assumed take-off and landing weights (Section 3.3.2), rather than recursively integrating fuel flow estimates along the trajectory. One possible explanation is that the heuristic method avoids the error-accumulation issues inherent in recursive fuel flow integration. Because the recursive approach derives mass at each timestamp by subtracting cumulative fuel burn (Section 3.3.1), any inaccuracy in early fuel flow estimates propagates to all subsequent mass values. Flights with missing or unreliable trajectory data near take-off are particularly susceptible to this effect. The heuristic method, in contrast, produces numerically stable and bounded mass profiles that do not depend on noisy intermediate calculations. The resulting features are more consistent across flights, which appears to translate into more effective inputs for the gradient-boosted model. This finding suggests that, for the purpose of feature construction, stability and consistency of the mass estimate are more valuable than physical fidelity.

5.2.4 Block 4: Training on Fuel Consumption

The fourth block, as defined in Section 4.5, compares models trained directly on segment-level fuel consumption (kg) with the fuel flow based models from the preceding blocks. All three configurations in this block produce substantially higher RMSE values than their fuel flow counterparts. The full recursive-mass model achieves an RMSE of 220.56 kg, compared with 203.75 kg for the equivalent fuel flow model in Block 2. Similarly, the reduced ADS-B + ACARS + Split configuration reaches 218.47, compared with 205.85 kg for the Block 1 benchmark.

Within this block, the reduced configurations slightly outperform the full model. ADS-B + ACARS + Split achieves 218.47 kg, ADS-B + ACARS reaches 219.13 kg, and the full model yields 220.56 kg. This ordering is the reverse of what is observed in the fuel flow blocks, where additional features consistently improve performance, although the differences within this block are small and may partly reflect stochastic variation (Section 5.3.3). One possible explanation is that the wider and more skewed distribution of fuel consumption targets amplifies the noise introduced by mass-related features, reducing their net benefit.

These results confirm that training on average fuel flow (kg/s) yields substantially better performance than training on fuel consumption (kg), even when using the same input features and the same data filter. The fuel flow representation normalizes for segment duration, which produces a tighter and less skewed target distribution. This normalization reduces the influence of long-duration segments, which can dominate the squared-error loss when the target is expressed in total fuel mass. The finding carries implications beyond the specific choice of training target. Predicting average fuel flow rather than total fuel consumption per segment aligns the model output more

closely with instantaneous fuel-burn rates. In principle, if each timestamp were treated as a minimal fuel segment, the model could serve as a pointwise fuel flow estimator along the trajectory, analogous to physics-based tools such as BADA [2], OpenAP [3], and Acropole [4]. This perspective suggests that fuel flow based modeling is not only more effective for the present regression task, but also more amenable to generalization toward higher-resolution fuel-burn estimation.

5.3 Discussion

The results presented across the four experimental blocks, taken together with the filter sensitivity analysis, reveal several patterns that inform the design of fuel flow estimation systems. This subsection synthesizes these observations and discusses their broader implications.

5.3.1 Role of Mass Estimation

The most consistent finding across Blocks 2 and 3 is that mass-related features improve prediction accuracy beyond what kinematic and airspeed variables alone can achieve. Without mass information (Block 1), the best RMSE obtained is 205.25. With recursive mass features (Block 2), the best RMSE decreases to 203.09, and with heuristic mass features (Block 3), it further decreases to 199.91. These improvements indicate that aircraft mass, a key determinant of required thrust and therefore fuel burn, provides information that the regression model cannot fully infer from observable kinematic or atmospheric variables.

The comparison between the two mass-estimation strategies is particularly instructive. The recursive approach derives mass at each timestamp through cumulative fuel flow integration, a process in which early estimation errors propagate to all subsequent values. The heuristic approach, in contrast, assigns mass through linear interpolation between assumed take-off and landing weights, producing smooth and bounded profiles regardless of trajectory data quality. The finding that the heuristic method yields consistently lower RMSE values suggests that, for the purpose of constructing regression features, numerical stability is more important than physical detail. This result should be interpreted in light of the fact that the current heuristic model relies on a fixed assumption of 80% of the maximum take-off weight (MTOW) as the initial mass, which does not account for flight-specific loading variation. Incorporating a more accurate take-off weight estimate into the heuristic framework could further improve performance by combining stability with flight-specific information.

5.3.2 Relative Value of ACARS and Wind-Corrected Features

A recurring pattern across all blocks is that ACARS-derived airspeed features provide more predictive value than wind-corrected alternatives. In Block 1, replacing ACARS with wind features increases the RMSE by approximately 7 points. In Blocks 2 and 3, removing ACARS consistently degrades performance more than removing wind features. In the heuristic-mass configuration (Block 3), removing wind features even improves the RMSE, suggesting that the wind-corrected variables introduce counterproductive noise in the presence of more reliable information sources.

Our hypothesis from these findings is that the ACARS airspeed information, which are obtained from the onboard sensors, is closer to the ground truth local wind condition than the ones obtained from weather data reanalysis, hence yields better prediction result. One possible explanation is that the spatial and temporal resolution of the ERA5 reanalysis grid is insufficient to capture the local wind conditions experienced by the aircraft. As a result, the wind-corrected airspeed features introduce interpolation artifacts that, rather than complementing ACARS, partially degrade the signal available to the model. This observation carries practical implications: when ACARS data are available, they should be preferred over wind-corrected alternatives. The inclusion of wind features appears justified only when ACARS data are absent, in which case wind-corrected airspeed provides a rea-

sonable, if less accurate, substitute. This is an approach we adopted through our empirical findings, and could be further studied in future works.

5.3.3 Implications and Limitations

The best-performing model in this study achieves an RMSE of 199.91 on the ranking set, using heuristic mass features combined with ADS-B, ACARS, and Vrate Split, but excluding wind features. This result demonstrates that a carefully engineered feature set combined with a gradient-boosted model can produce representative fuel flow estimates from operationally available data sources.

Several limitations should be noted. First, the heuristic mass model relies on a fixed assumption of 80% MTOW for take-off weight, which does not account for variation in actual loading across individual flights. Second, the evaluation is based on a single ranking set with models trained under a single random seed, and some of the observed RMSE differences between configurations are small enough that they may fall within the range of stochastic variation due to random seed effects or data sampling. Establishing which of the close-margin differences are statistically robust would require repeating each configuration under a more comprehensive evaluation protocol, which we outline in Section 5.4. Third, the dataset exhibits substantial imbalance across aircraft types (see Table 1), with some types represented by fewer than 10 flights. Model performance on underrepresented types may therefore not generalize reliably. Fourth, the ACARS features, while shown to be valuable, are not consistently available across all flights, which limits the applicability of the best-performing configuration to scenarios where ACARS data are present.

5.4 Future Work

The findings of this study point to several directions for future investigation. A first improvement to the mass-estimation pipeline is to replace the fixed 80% MTOW assumption in the heuristic model with the estimated take-off weight obtained from the TOW regression model. This substitution would retain the numerical stability of the heuristic approach while incorporating flight-specific mass information, and would allow a more controlled assessment of how take-off weight accuracy affects downstream fuel flow prediction.

A second direction concerns the development of a more accurate take-off weight estimator. The current TOW model, trained on the PRC 2024 dataset [16], achieves an RMSE of approximately 3,000 kg and does not cover all aircraft types in the present dataset. Expanding the training data and improving the model architecture could reduce this uncertainty and thereby improve the quality of mass-related features throughout the pipeline.

A third direction concerns extensions to the engineered feature set. Energy-rate descriptors could be tested directly, since thrust relates to the drag and the energy rate divided by airspeed, and the sub-interval decomposition could be extended from the vertical rate to airspeed and altitude to assess whether their within-segment profiles add predictive value. The number of sub-intervals, currently fixed at ten without tuning, could also be swept over a range of values (for example 5, 10, 15, and 20) to determine whether a coarser or finer decomposition is preferable.

A fourth direction is a more granular and statistically rigorous ablation study. A finer-grained analysis would clarify the role of individual features, including isolating the contribution of the flight-phase label and disentangling the value of ACARS-derived airspeed into its higher fidelity and its availability on segments where groundspeed is missing, for example by comparing NaN ratios across sources and by masking ACARS wherever groundspeed is absent. These comparisons would further benefit from a more rigorous statistical protocol. Because the results reported are based on a single seed and a single train/test split, several of the close-margin differences in Table 5 cannot be attributed to a configuration change with confidence. Training each configuration under

multiple random seeds and reporting the mean RMSE, or estimating performance through K-fold cross-validation or bootstrapped resampling of the training and test sets, would provide confidence intervals and allow the observed differences to be tested for significance.

Finally, the observation that fuel flow based modeling aligns the prediction target with instantaneous fuel-burn rates (Block 4) suggests that future work could explore pointwise fuel flow estimation at the timestamp level. Such an approach would require adapting the feature-engineering pipeline to operate at individual trajectory points rather than segment-level summaries, but could provide higher-resolution fuel-burn profiles with broader applicability for flight performance analysis and environmental assessment.

6. Conclusion

This study has investigated the extent to which aircraft fuel consumption can be inferred statistically from ADS-B trajectory data, supplemented by weather information, inferred mass, and physics-informed features. Developed in the context of the PRC Data Challenge 2025, the work addresses the fundamental difficulty that key determinants of fuel burn, such as aircraft mass and engine state, are not directly observable from surveillance data. To address this, we developed a multi-stage preprocessing and feature-engineering pipeline that incorporates ADS-B cleaning, ERA5 wind augmentation, OpenAP-based phase identification, and two alternative mass-estimation strategies, and trained a LightGBM regression model to predict average fuel flow, which is then integrated to obtain segment-level fuel consumption.

A structured ablation study across four experimental blocks yielded three principal findings. First, data preprocessing choices substantially affect prediction quality. Filtering physically implausible fuel flow values reduces the RMSE by approximately 22 points. Second, ACARS-derived airspeed features provide more predictive value than wind-corrected alternatives reconstructed from ERA5 reanalysis fields. This finding suggests that ACARS, despite its sparsity, captures local airspeed conditions more accurately than gridded meteorological data, which suffer from interpolation noise and scale mismatch with the trajectory. Third, mass-related features improve accuracy beyond what kinematic and airspeed variables alone can achieve, reducing the best RMSE from 205.25 without mass information to 199.91 with heuristic mass features. Notably, the heuristic mass-estimation approach, which linearly interpolates between assumed take-off and landing weights, outperforms the physically more detailed recursive method that integrates fuel flow estimates along the trajectory. This result indicates that numerical stability in feature construction is highly important for this regression task, as the recursive approach is susceptible to error accumulation from noisy or incomplete trajectory data near take-off.

The proposed approach demonstrates that a carefully engineered feature set, combining statistical embeddings of ADS-B and ACARS variables, vertical-rate phase decomposition into ten sub-intervals, and heuristic mass descriptors, can produce representative fuel flow estimates from operationally available data. The final configuration achieves an RMSE of 199.91 kg on the ranking set. An additional finding concerns the choice of training target: models trained on average fuel flow (kg/s) consistently outperform those trained directly on segment-level fuel consumption (kg), because the fuel flow representation normalizes for segment duration and produces a tighter, less skewed target distribution that reduces the influence of long-duration segments on the regression loss.

Several limitations should be noted. The heuristic mass model relies on a fixed assumption of 80% MTOW for take-off weight, which does not account for flight-specific loading variation. The evaluation is based on a single ranking set, and some of the observed RMSE differences between configurations are small enough that they may fall within the range of stochastic variation. The

dataset exhibits substantial imbalance across the 26 aircraft types, with nine types represented by fewer than ten flights, which may limit generalization for underrepresented types. Furthermore, the best-performing configuration requires ACARS data, which are not consistently available across all flights.

Future work should investigate several directions. First, replacing the fixed take-off weight assumption with flight-specific estimates from a dedicated TOW regression model would retain the stability of the heuristic approach while incorporating per-flight mass information. The current TOW model, trained on the PRC 2024 dataset, achieves an RMSE of approximately 3,000 kg and does not cover all aircraft types in the present dataset; expanding its training data and improving its architecture could reduce this uncertainty. Second, the observation that fuel flow based modeling aligns the prediction target with instantaneous fuel-burn rates suggests that pointwise fuel flow estimation at the timestamp level could provide higher-resolution fuel-burn profiles, analogous to physics-based tools such as OpenAP and BADA, with broader applicability for flight performance analysis and environmental assessment.

Author contributions

- Muhammad Fazlur Rahman: Conceptualization, Methodology, Software, Formal Analysis, Writing (Original Draft), Data Curation, Visualization
- Aidana Tassanbi: Conceptualization, Methodology, Software, Formal Analysis, Writing (Review and Editing), Data Curation, Visualization
- Zhan-Peng Yang: Conceptualization, Methodology, Software, Formal Analysis, Writing (Review and Editing), Data Curation, Visualization
- Mohammed El Dor: Conceptualization, Methodology, Software, Formal Analysis, Writing (Review and Editing), Data Curation, Visualization
- Bob van Dillen: Conceptualization, Methodology, Software, Formal Analysis, Writing (Review and Editing), Data Curation, Visualization

Open data statement

The dataset is available on Zenodo at <https://doi.org/10.5281/zenodo.19184661> under a Creative Commons Attribution 4.0 International license (CC BY 4.0). The dataset consists of Parquet files totaling approximately 5 GB provided by EUROCONTROL PRC.

Additional data related to mass estimation is available on <https://github.com/meldorr/PRC-2025/tree/main>

Meteorological data is obtained through the fastmeteo library [11] that incorporates open access ERA5 reanalysis data

Reproducibility statement

For the code, please refer to <https://github.com/meldorr/PRC-2025/tree/main>

Information on how to reproduce this research, including access to: 1) source code related to the research, 2) source code for the figures, 3) source code/data for the tables when applicable.

References

- [1] Roger Teoh, Zebediah Engberg, Ulrich Schumann, Christiane Voigt, Marc Shapiro, Susanne Rohs, and Marc E. J. Stettler. “Global aviation contrail climate effects from 2019 to 2021”. en. In: *Atmospheric Chemistry and Physics* 24.10 (May 2024), pp. 6071–6093. ISSN: 1680-7324. DOI: 10.5194/acp-24-6071-2024. URL: <https://acp.copernicus.org/articles/24/6071/2024/> (visited on 11/29/2025).
- [2] Angela Nuic, Damir Poles, and Vincent Mouillet. “BADA: An advanced aircraft performance model for present and future ATM systems”. In: *International journal of adaptive control and signal processing* 24.10 (2010), pp. 850–866.
- [3] Junzi Sun, Jacco Hoekstra, and Joost Ellerbroek. “OpenAP: An open-source aircraft performance model for air transportation studies and simulations”. In: *Aerospace* 7.8 (2020), p. 104. DOI: 10.3390/aerospace7080104. URL: <https://doi.org/10.3390/aerospace7080104>.
- [4] Gabriel Jarry, Daniel Delahaye, and Christophe Hurter. “Towards aircraft generic Quick Access Recorder fuel flow regression models for ADS-B data”. In: *International Conference on Research in Air Transportation*. 2024.
- [5] Gabriel Jarry, Ramon Dalmau, Philippe Very, and Junzi Sun. “On the generalization properties of deep learning for aircraft fuel flow estimation models”. In: *Transportation Research Part C: Emerging Technologies* 176 (2025), p. 105143. ISSN: 0968-090X. DOI: <https://doi.org/10.1016/j.trc.2025.105143>. URL: <https://www.sciencedirect.com/science/article/pii/S0968090X25001470>.
- [6] Junzi Sun, Enrico Spinielli, and Martin Strohmeier. “Aircraft Fuel Burn Estimation: The EUROCONTROL PRC 2025 Data Challenge”. In: *Journal of Open Aviation Science* 4.3 (2026).
- [7] Hans Hersbach, Bill Bell, Paul Berrisford, Gionata Biavati, András Horányi, Joaquín Muñoz Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Iryna Rozum, et al. “ERA5 hourly data on pressure levels from 1940 to present”. In: *Copernicus climate change service (c3s) climate data store (cds)* (2023). DOI: 10.24381/cds.bd0915c6.
- [8] Junzi Sun and Esther J. Roosenbrand. “Fast contrail estimation with OpenSky data”. In: *Journal of Open Aviation Science* (2024). DOI: <https://doi.org/10.59490/joas.2023.7264>.
- [9] 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 (2016), pp. 52–62.
- [10] International Civil Aviation Organization (ICAO). *ICAO Aircraft Engine Emissions Databank*. European Union Aviation Safety Agency (EASA), 2025. URL: <https://www.easa.europa.eu/en/domains/environment/icao-aircraft-engine-emissions-databank>.
- [11] Junzi Sun, Xavier Olive, Enrico Spinielli, Gabriel Jarry, and Benoit Figuet. *Fast Meteo: A super-fast Python package to obtain meteorological parameters for your flight trajectories*. URL: <https://github.com/open-aviation/fastmeteo>.
- [12] Enrico Spinielli, Junzi Sun, Martin Strohmeier, Xavier Olive, Quinten Goens, Rainer Koelle, Allan Tart, and John Fitzgerald. “Aircraft Takeoff Weight Estimation: The EUROCONTROL PRC 2024 Data Challenge”. In: *Journal of Open Aviation Science* 3.2 (July 2025). DOI: 10.59490/joas.2025.8252. URL: <https://journals.open.tudelft.nl/joas/article/view/8252>.
- [13] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. “Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?” In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS ’22. Red Hook, NY, USA: Curran Associates Inc., Nov. 28, 2022, pp. 507–520. ISBN: 978-1-7138-7108-8.
- [14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/Ke-17_paper.pdf.

://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

- [15] Mayara Condé Rocha Murça, Marcos Ricardo Omena De Albuquerque Maximo, João Paulo De Andrade Dantas, João Basílio Tarelho Szenczuk, Carolina Rutili De Lima, Lucas Orbolato Carvalho, and Gabriel Adriano De Melo. “Open Machine Learning Models for Actual Takeoff Weight Prediction”. In: *Journal of Open Aviation Science* 3.2 (Apr. 8, 2025). issn: 2773-1626. doi: 10.59490/joas.2025.7963.
- [16] Junzi Sun, Enrico Spinielli, Xavier Olive, and Martin Strohmeier. *Dataset for EUROCONTROL PRC 2024 Data Challenge on Aircraft Take-off Weight Estimation*. 2025. doi: 10.4121/8cb8484b-dbe7-4750-8b87-a5b1dbc621b4.v2.