

ARTICLE

A Physics-Guided Gradient Boosting Framework for Open-Source Aviation Fuel Estimation

Yiannis Grigoriou ^{*,1} Eftychios Eftychiou ² Christian Vitale ¹ Giorgos Petteimeridis ³ Nicolas Souli ¹ and Panayiotis Kolios ³

¹Department of Electrical and Computer Engineering and the KIOS Research and Innovation Center of Excellence (KIOS CoE), University of Cyprus, Nicosia, Cyprus.

²Ministry of Transport, Communications and Works, Nicosia, Cyprus.

³Department of Computer Science and the KIOS Research and Innovation Center of Excellence (KIOS CoE), University of Cyprus, Nicosia, Cyprus.

*Corresponding author: grigoriou.yiannis@ucy.ac.cy

(Received: 7 Apr 2026; Revised: 17 Jul 2026; Accepted: 22 Jul 2026; Published: 23 Jul 2026)

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

Abstract

Accurate segment-level fuel consumption estimation is essential for aviation emission accounting and cost management, yet existing models rely on proprietary flight recorder data and fail to generalize across heterogeneous open trajectory sources. This paper presents a physics-guided gradient boosting framework for direct fuel consumption prediction, solving the EUROCONTROL PRC 2025 Data Challenge. The pipeline integrates raw ADS-B trajectories with METAR meteorological observations, airport infrastructure data, aircraft performance coefficients, and regional passenger load factors, including segment boundary injection, linear interpolation, and timestamp correction. Physics-informed features are derived by running the OpenAP FuelFlow model with dynamic mass tracking over complete flight trajectories, producing physically consistent fuel burn estimates for all 36 aircraft types in the dataset. To address the narrowbody/widebody class imbalance, synthetic widebody training samples are generated via Gaussian noise perturbation with kinematic bounding, increasing the widebody training share from 18.7% to 34.1%. Sequential Forward Selection reduces 133 candidate features to a compact subset, and Bayesian hyperparameter optimization via Optuna identifies the final XGBoost configuration. On the held-out validation set, the proposed model achieves a mean absolute error (MAE) of 85.7 kg, RMSE of 204.6 kg, and $R^2 = 0.961$ over segments with an average duration of 8.87 minutes. On the official hidden competition test sets, which have an average segment duration of 7.24 minutes, the model achieves an RMSE of 214.92 kg, outperforming standard machine learning baselines such as Random Forest and LightGBM. Comparison against linear models such as Ridge Regression (MAE 262.2 kg, $R^2 = 0.355$) confirms the high nonlinearity of the task. These results demonstrate that competitive fuel consumption estimation accuracy is achievable from publicly available data alone, without reliance on proprietary airline records.

Keywords: fuel consumption prediction; gradient boosting; ADS-B trajectory; physics-informed features; open aviation data; machine learning

Abbreviations: ADS-B: Automatic Dependent Surveillance-Broadcast, RMSE: Root Mean Squared Error, MAE: Mean Absolute Error, METAR: Meteorological Aerodrome Report, XGBoost: Extreme Gradient Boosting, OpenAP: Open Aircraft Performance, PRC: Performance Review Commission, AGL: Above Ground Level, SFS: Sequential Forward Selection

1. Introduction

Aviation is a significant contributor to global gas emissions, accounting for approximately 2% to 3% of worldwide CO₂ output [1]. This environmental impact is projected to increase as air traffic volumes continue to grow, highlighting the need for transparent and independently verifiable fuel consumption models that can be applied across heterogeneous flight data sources. At the same time, fuel costs represent the largest single variable expense for airlines, typically comprising between 30% and 40% of total operational costs [1]. As a result, even marginal improvements in fuel prediction accuracy can provide substantial economic benefits and support the industry's transition toward low-carbon operations. Therefore, the airline industry's environmental and economic considerations highlight the necessity for robust, data-driven methodologies in modeling fuel consumption.

Accurate fuel-flow estimation is challenging because it depends on aircraft performance, operating procedures, aircraft condition, and uncertain environmental factors. Traditional methods employ parametric models, such as the Base of Aircraft Data (BADA) [2], which provide useful physics-based fuel-flow baselines under nominal conditions [3, 4]; however, these techniques can not fully capture aircraft-specific variability, aging effects, or deviations from real-world operations [5]. To address these limitations, recent data-driven works have enhanced physics-based approaches by learning from operational or high-fidelity simulator data while using BADA as a reference baseline [3, 4].

More recently, machine-learning methods, including ensemble models trained on operational recorder data, have shown a greater potential for aviation fuel prediction in multiple flight phases and aircraft types [6]. However, many high-performing approaches are still based on proprietary Quick Access Recorder (QAR) or flight-recorder datasets from specific airlines, aircraft types, or operating conditions, which limits their reproducibility and transferability to heterogeneous open-data settings. This leads to the need for an open-data framework that can combine noisy ADS-B trajectories with external information, such as meteorological observations, aircraft characteristics, airport metadata, and passenger load factors.

The EUROCONTROL Performance Review Commission (PRC) Data Challenge series has emerged as an important community benchmark for open, reproducible aviation data science. The 2024 edition [7] focused on predicting actual takeoff weight (ATOW), a key intermediate variable for fuel burn estimation, using open trajectory data from the OpenSky Network. Building directly on this foundation, the PRC 2025 Data Challenge [8] shifted the target to direct fuel consumption prediction for individual flight segments, using trajectory data with all the noise and incompleteness of operational sources. This paper presents our solution to that challenge, a modular, end-to-end machine learning pipeline that integrates METAR meteorological data, airport characteristics, and regional passenger load factor estimates with physics-informed feature engineering derived from aircraft performance models, and trains an XGBoost ensemble optimized via Optuna hyperparameter search [9].

Accordingly, this work focuses on developing and evaluating a reproducible, physics-guided machine-learning framework for direct segment-level fuel-consumption estimation using heterogeneous open aviation data. The main contributions of this work are summarized as follows.

- A modular, open-data pipeline for segment-level fuel consumption prediction that integrates heterogeneous sources including ADS-B trajectories, surface weather observations, airport infrastructure data, and regional passenger load factors, enabling direct prediction from publicly available inputs without reliance on proprietary airline records.
- A physics-informed feature engineering framework that runs the OpenAP FuelFlow model over complete flight trajectories with flight-level dynamic mass tracking, ensuring that the aircraft mass state at each segment is physically consistent rather than independently estimated, and combining the resulting outputs with trajectory aggregate statistics, flight phase durations, and aircraft

performance parameters.

- A targeted synthetic data augmentation strategy that generates 25,000 perturbed widebody training samples with emphasis on long-duration segments, addressing the class imbalance between narrowbody and widebody aircraft and improving generalization to long-haul operations.
- A systematic model development workflow combining Sequential Forward Selection for dimensionality reduction with Bayesian hyperparameter optimization over a large configuration space, with all preprocessing steps strictly fitted on the training fold to prevent data leakage, validated against the PRC 2025 Data Challenge benchmark.

2. Related Work

Accurate aircraft fuel estimation has been studied from three main perspectives: (i) physics-based and phase-specific modeling, (ii) data-driven learning from proprietary operational records, and (iii) open-data approaches that estimate aircraft mass, fuel burn, or emissions from trajectory observations. A smaller but relevant body of work also considers transferability to new routes and adjacent transport domains. The following related work is organized around these macro-level themes to reflect the main research directions in the literature. This thematic structure also helps clarify, in the subsequent discussion, how the proposed approach differs from existing studies in terms of data sources, modeling strategy, and prediction scope.

2.1 Physics-Based and Phase-Specific Fuel Modeling

Early studies on aviation fuel estimation often used semi-empirical or phase-specific models based on aircraft performance physics. Chati and Balakrishnan combined physically informed variables with Gaussian process regression to estimate fuel flow during low-altitude climb-out and approach, and reported better results than classical baselines such as BADA and ICAO-BFFM2 [10]. Turgut et al. studied how descent flight-path angle affects fuel burn during continuous descent operations and showed that descent geometry can strongly influence fuel use [11]. In a similar way, Oruç et al. proposed a particle-swarm-based model for descent fuel-flow prediction using mainly altitude and true airspeed, while Atasoy et al. used Gaussian process regression to predict climb-out fuel consumption from gross weight, speed, temperature, wind, and trajectory variables [12, 13]. Collectively, these studies show that phase-specific models can achieve high accuracy and remain interpretable when they are applied to tightly constrained operating conditions and supported by high-quality state information.

However, their predictive value is often limited outside these controlled settings. Most of these studies rely on small datasets, specific aircraft types, narrow flight envelopes, or inputs that are difficult to obtain in open-data environments, such as known aircraft weight, true airspeed, or clean recorder measurements [10, 11, 12, 13]. Related work also shows that sensor noise, turbulence, and measurement uncertainty can reduce the reliability of fuel-consumption analysis and make precise prediction more difficult, even in focused scenarios with limited operating variability [14]. In addition, Zou et al. combined takeoff mass estimation with BADA-based fuel and emissions modeling to improve downstream calculations when true mass is unavailable, but the framework still depends on limited aircraft and engine assumptions [15]. Overall, this line of work shows that physics-aware and phase-specific models can perform well, but their accuracy often depends on controlled conditions and high-quality information that are not always available in open trajectory data.

2.2 Data-Driven Fuel Prediction from Proprietary Flight Data

When high-quality QAR or Flight Data Recorder (FDR) data are available, data-driven models usually report the best fuel-prediction accuracy. Baumann et al. compared neural networks with tree-based

methods for both full-flight missions and individual flight phases. Their results showed that neural networks performed very well for complete missions, while tree-based models often generalized better in more uniform settings such as cruise [16]. Zhao et al. used an RBF neural network with phase-segmented QAR data and showed that nonlinear models can predict fuel consumption effectively during takeoff/climb and cruise, although prediction becomes more difficult in descent and approach because of frequent control actions and rapidly changing conditions [17]. Wahid et al. also found that stacked Long Short-Term Memory (LSTM) models outperform linear, tree-based, and Support Vector Regression (SVR) baselines across taxi, cruise, and approach phases, but at the cost of very high training time and difficult manual tuning [18]. Tang et al. extended this direction with a hybrid Wavelet Packet Decomposition (WPD)-Improved Arctic Puffin Optimization (IAPO)-BiLSTM-KAN model for range fuel prediction, showing that deep hybrid architectures can achieve very high accuracy when rich proprietary data and extensive hyperparameter optimization are available [19].

Taken together, these studies confirm that nonlinear data-driven methods are highly effective for fuel prediction, especially when trained on detailed operational records. However, they also reveal an important limitation: the best results are usually obtained on proprietary airline datasets, specific aircraft fleets, or narrowly defined flight phases [16, 17, 18, 19]. This means that their predictive performance may not transfer well to more open and heterogeneous settings. The difficulty becomes even clearer when ADS-B data are used instead of recorder data. Huang showed this for general aviation, where Classification and Regression Tree (CART) and neural-network models trained on FDR data transferred less effectively to ADS-B observations, mainly because ADS-B signals are noisier, less complete, and more affected by missing or degraded measurements [20]. Overall, this body of work shows that data-driven models can provide very accurate fuel predictions, but their performance often depends on clean and detailed input data that are not always available in open-data settings.

2.3 Open-Data Trajectories, Takeoff Mass, and Emissions Estimation

Open-data approaches address the accessibility problem more directly. The EUROCONTROL PRC 2024 Data Challenge identified ATOW estimation from open trajectory data as an important step for later fuel-consumption and emissions modeling [7]. Earlier large-scale studies often relied on heuristic assumptions about aircraft mass or on approximate payload reconstruction. For example, some studies bounded the initial mass as a fraction of maximum takeoff weight (MTOW), while others estimated it by combining empty weight, reserve fuel, expected cruise fuel, and payload approximations based on fleet load factors [21, 22]. More recent work has shown that these assumptions can be replaced by learning-based methods. Murça et al. reported strong ATOW prediction accuracy using features derived from open trajectories, and other studies have also explored direct fuel-flow prediction from trajectory data using neural networks and tree-based models [23, 24].

Another important challenge is the quality and completeness of the trajectory source itself. The OpenSky Network provides strong coverage over many continental areas, but coverage is still limited over oceans and remote regions. This has motivated studies that combine ground-based, satellite-based, and contract-based surveillance data to reconstruct more realistic long-haul trajectories and improve global emissions estimates [25, 26, 27, 21]. This line of work is highly relevant to fuel-consumption prediction because missing trajectory segments, uncertain aircraft mass, and limited environmental information can all increase prediction error. Huang's study on general aviation using ADS-B data, as well as the ADS-B-based framework of Zou et al. that integrates takeoff mass estimation into fuel and emissions calculations, both show the potential of open-data modeling while also highlighting its sensitivity to noisy measurements, aircraft diversity, and incomplete operational information [20, 15]. Overall, this literature shows that open-data methods improve transparency

and scalability, but accurate fuel prediction remains difficult when the available trajectories are noisy, incomplete, or missing important operational variables.

2.4 Transferability, New Routes, and Cross-Domain Extensions

A smaller body of related work provides useful motivation for the modeling choices adopted in our proposed approach. Rather than focusing on commercial aviation fuel prediction, these studies show broader patterns that help justify the use of contextual features, route-aware representations, and flexible nonlinear models.

Within aviation, Yanto et al. first clustered flights according to their operational characteristics and then trained regression models within each group, showing improved fuel estimation performance on new routes [28]. Although the proposed method does not use clustering, this result still underlines the importance of accounting for route structure and operational heterogeneity in fuel-prediction models. In a related aviation setting, Ceylan et al. reported strong performance from tree-ensemble methods such as XGBoost for fuel estimation in jet unmanned aerial vehicles, while Fang et al. reviewed artificial-intelligence applications in civil aviation and highlighted the growing use of ensemble, neural, and hybrid models across operational tasks [29, 30]. Although these studies address different applications, they support the idea that flexible machine-learning models can adapt well to complex operational data.

Similar observations appear outside aviation. In smart mobility, Varshini et al. showed that combining richer contextual information with nonlinear learning can improve predictions of fuel use, cost, and routing outcomes [31]. In maritime transport, Xie et al. showed that strong predictive performance can be achieved by combining careful data preprocessing with gradient-boosted models [32]. In public transit, Ayman et al. demonstrated the importance of integrating telemetry with elevation, weather, and traffic information for energy prediction [33]. While these studies are not directly comparable to segment-level fuel estimation for commercial aircraft, they reinforce two design principles that are relevant here: prediction quality benefits from contextual data fusion, and flexible nonlinear models are often better suited than simple linear baselines for noisy operational environments.

2.5 Positioning of the Proposed Approach

The reviewed literature reveals a clear methodological gap. Models based on proprietary operational data can achieve very high fuel-prediction accuracy, but they are often difficult to reproduce and are usually developed for specific aircraft fleets, flight phases, or operating conditions. In contrast, open-data studies improve transparency and scalability, but many of them focus on intermediate tasks, such as takeoff mass estimation, or on broader emissions-calculation pipelines rather than direct segment-level fuel prediction.

The proposed approach is designed to address this gap. It combines heterogeneous open data sources with OpenAP-derived fuel and mass-state features, aircraft performance parameters, and physical constraints applied during mass estimation and data augmentation, which together guide a standard gradient-boosting model for direct segment-level fuel estimation. In this way, it aims to preserve the robustness and flexibility of tabular machine-learning methods while avoiding dependence on proprietary airline records. The approach therefore differs from much of the existing literature by targeting accurate fuel prediction in a fully open-data setting using a feature-engineering pipeline that remains closely linked to aircraft performance physics.

3. Methodology

3.1 System Overview

The predictive framework developed in this study comprises the complete end-to-end methodology and is implemented as a modular, three-phase pipeline designed to handle the high-dimensional, noisy, and heterogeneous nature of the aviation data. The system transitions from raw data fusion to physics-informed feature refinement, concluding in a gradient-boosted ensemble for high-fidelity fuel consumption estimation.

As illustrated in Fig. 1, the pipeline spans three phases. Phase 1 (Data Ingestion) integrates PRC 2025 Challenge [34] flight records with sub-hourly METAR observations, regional airport metadata, and aircraft performance coefficients to ensure each flight segment is described by its trajectory, atmospheric conditions, and operational context. Phase 2 handles data alignment through a timestamp correction algorithm that synchronizes ADS-B trajectory points with gate-to-gate boundaries, followed by missing value imputation and feature engineering that transforms raw inputs into a unified vector representation including kinematic aggregates and aerodynamic interaction terms. Phase 3 applies domain constraints and dimensionality reduction, where the OpenAP FuelFlow model provides a physical performance baseline, dynamic mass tracking ensures realistic aircraft state progression across flight segments, and Sequential Forward Selection (SFS) identifies a compact, high-value feature subset that is subsequently tuned via Bayesian hyperparameter optimization.

Finally, the core inference engine is an XGBoost model based on gradient-boosted decision trees. Each tree is trained sequentially to correct the residuals of the preceding tree, capturing non-linear relationships between atmospheric conditions, aircraft mass, and fuel consumption. The final prediction is summed across the ensemble and rescaled to produce final fuel consumption predictions in kg.

3.2 Phase 1 - Data Ingestion

Figure 2 illustrates the global route coverage of the PRC 2025 dataset, visualizing all flight trajectories as great-circle arcs colored by the number of flights per origin-destination pair. The dataset encompasses a geographically diverse set of European and intercontinental routes, with the highest traffic density concentrated over Western Europe and the North Atlantic corridor. Route frequency ranges from fewer than 100 to over 5,000 flights per pair, reflecting the significant operational imbalance between high-frequency short-haul routes and long-haul widebody operations.

3.2.1 Data Sources

The dataset used in this study is composed of core competition records provided by the EUROCONTROL PRC 2025 Data Challenge and supplementary external datasets integrated to enrich the feature space.

The primary competition data consists of three splits (train, rank, and final) and includes four main components, namely Flightlist, containing metadata such as origin/destination ICAO codes and aircraft types, Fuel Intervals, defining the target segments and ground-truth fuel consumption for training, Trajectories, providing high-frequency ADS-B state vectors (latitude, longitude, altitude, and groundspeed), and Airport Metadata, providing coordinates and elevations for all airports.

Four additional external datasets are integrated at the feature engineering stage to account for physical constraints and environmental conditions. The airport dataset provided by the competition is enriched by scraping SkyVector [35] to add runway infrastructure, including headings and lengths, which provides geometric context for flight phase detection. The aircraft performance dataset [36] provides manufacturer-reported and OpenAP-derived parameters such as Operating Empty Weight (OEW) and aerodynamic coefficients. METAR meteorological reports provide surface weather observations (wind and temperature) for origin and destination airports, decoded into numerical features

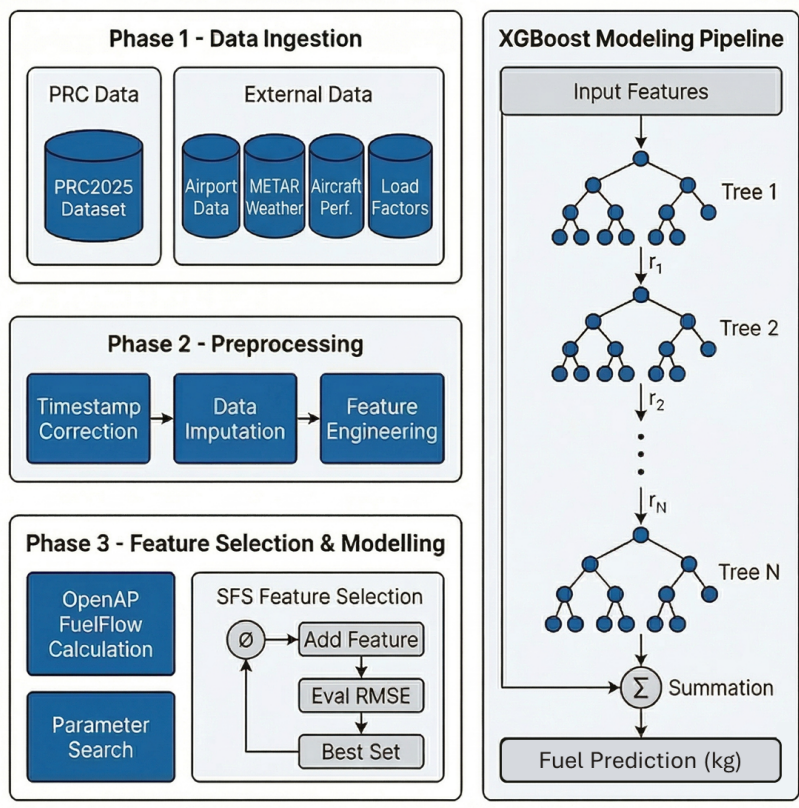


Figure 1. Overview of the proposed physics-guided fuel-estimation pipeline, from data ingestion and preprocessing to OpenAP feature generation, feature selection, hyperparameter tuning, and XGBoost prediction.

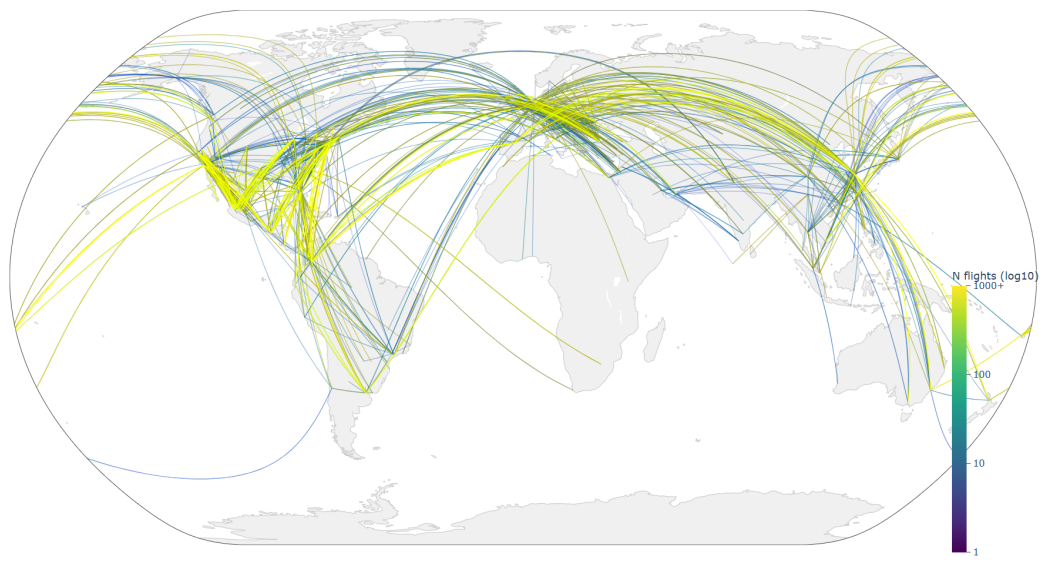


Figure 2. Global route coverage of the PRC 2025 dataset, showing origin-destination arcs colored by route frequency.

to capture atmospheric effects during climb and descent. Finally, regional passenger load factors based on IATA [37, 38] historical data are used to initialize a more realistic aircraft mass state for each flight.

Table 1 summarizes the heterogeneous data sources integrated into the pipeline, their respective join keys, and their contribution to the final featured dataset.

Table 1. Integrated data sources used in the fuel-prediction pipeline, including join keys, extracted features, and functional roles.

Source Dataset	Join Key	Key Features Extracted	Functional Role
Flightlist	flight_id	Origin/Dest ICAO, A/C type, nominal times	Base metadata & primary key
Fuel Intervals	flight_id, idx	Start/end timestamps, actual fuel (y)	Target variable & segments
Trajectories (ADS-B)	flight_id, time	Latitude, Longitude, Alt, GS, VR, Mach, CAS/TAS	Kinematic & aerodynamic profiles
Airport Metadata	ICAO Code	Runway length/heading, elevation	Infrastructure context
Aircraft Performance	aircraft_type	OEI, MTOW, wing MAC, dynamic mass, fuel burn (OpenAP) drag coefficients	Physical & performance constraints
METAR Reports	ICAO, Timestamp	Wind (u, v), temperature, visibility	Local atmospheric conditions
Load Factors	Region Pair	Regional cabin occupancy estimates	Initial mass state estimation

3.2.2 Physics-Informed Feature Augmentation with OpenAP

To inject physics-based knowledge that cannot be recovered from trajectory statistics alone, the OpenAP FuelFlow model [36, 39] is run over the complete trajectory of each flight with dynamic mass tracking from the moment of takeoff. In contrast to a static or segment-local mass estimate, this flight-level mass tracking subtracts fuel burned in all preceding intervals before computing the fuel flow for the current interval, producing a continuously decreasing mass profile consistent with thermodynamic principles. For each flight interval, the model outputs a predicted fuel consumption, the starting mass at interval entry, and segment-level summary statistics including mean altitude and groundspeed, mean vertical speed, interval duration, and altitude change. This process is applied with synonym lookup enabled to cover all 36 aircraft types in the dataset, with unsupported types assigned to the closest synonym. The resulting augmented dataset constitutes a physics-informed prediction baseline and a source of mass state features used in downstream model training.

3.2.3 Meteorological Data Integration

Environmental factors, specifically wind and temperature at origin and destination airports, significantly affect fuel consumption during climb and descent phases. To capture these effects, highly detailed Meteorological Aerodrome Reports (METARs) sourced from the Iowa Environmental Mesonet (IEM) weather archive [40] were integrated into the pipeline.

The integration process involves mapping origin and destination ICAO codes from the flight list to the nearest active METAR station. A vectorized Haversine distance algorithm calculates the great-circle distance between airports and available weather stations, associating non-reporting airports with the nearest station within a 100 nautical mile (NM) radius.

3.2.4 Regional Passenger Load Factor and Airport Imputation

To improve takeoff mass estimation, regional passenger load factors were computed from historical IATA data [37] covering the April-August 2025 period. Each airport pair in the dataset is assigned a load factor based on its origin and destination regions, providing a more resembling estimate than global averages. The airport database is further enriched by recursively scraping SkyVector to obtain runway infrastructure and elevation data. For airports with missing coordinates or elevations, values are imputed using regional medians to ensure geometric consistency. The pygeomag World Magnetic Model is independently applied to convert magnetic runway headings to true tracks, aligning the runway data with the true tracks recorded in the flight dataset for accurate flight phase determination and runway assignment.

3.3 Phase 2 - Preprocessing

3.3.1 Filtering

Raw ADS-B trajectory files contain erroneous or noisy records from transponder artifacts and network effects. Each trajectory file is filtered using the traffic library [41] with a composite filter chain comprising constant-position detection, constant-speed detection, derivative-based spike rejection, and isolated-point removal. Erroneous data points are masked (set to NaN) during this process, ensuring that noisy measurements are excluded without altering the time index. No interpolation is applied during this filtering stage to preserve the raw state of the data before further processing. Masking is also propagated to co-located meteorological fields (u-component of wind, v-component of wind, and temperature) whenever the corresponding position or altitude record is flagged as erroneous.

3.3.2 Segment Boundary Injection and Interpolation

Following filtering, the segment start and end timestamps defined in the fuel interval files are injected into each trajectory. These injected timestamps serve as alignment markers to synchronize the trajectory with the fuel measurement intervals. Since these markers do not have associated sensor readings, they are supplemented by linear interpolation from the nearest valid (non-masked) ADS-B observations for altitude, groundspeed, and lateral position. This step ensures that trajectory segments can be extracted precisely at the interval boundaries even when those boundaries do not align with recorded ADS-B transmissions.

3.3.3 Flight Phase Detection

Flight phase labels are assigned to each trajectory point using the OpenAP FlightPhase model [36, 42, 43], which is conditioned on altitude, groundspeed, and vertical rate time series. The resulting phase sequence covers ground (GND), climb (CL), cruise (CR), descent (DE), and level flight (LVL). It must be noted that a distinction is made between cruise and level flight. While both represent segments with a zero vertical rate, cruise (CR) is reserved for steady-state flight at high altitudes (typically above 10,000 ft), whereas level (LVL) captures intermediate altitude holds during climb and descent. These phase labels are used both in the physics-informed fuel augmentation step and in the feature engineering stage.

Phase-specific fuel consumption features are calculated by multiplying OpenAP phase durations with aircraft-specific fuel burn rates from the imputed aircraft performance database. Specifically, the time spent in each identified phase, including climb, cruise, and descent, is multiplied by its corresponding fuel flow rate. These contributions are combined to produce a baseline fuel estimate for the segment. The segment's average burn rate is computed as a weighted average across its flight phases, producing a measure of fuel consumption.

In addition to the OpenAP FlightPhase model, a hierarchical rule based flight phase detection model

was implemented as shown in Table 2, which uses the trajectory data along with the airport characteristics such as runway headings and elevations to assign each trajectory point one of nine flight phases. Following the classification of each trajectory point into operational flight phases, the phase fraction features are computed to quantify the proportional time spent in each phase during a given flight segment.

Table 2. Hierarchical rule-based criteria used to classify trajectory points into operational flight phases.

Phase	Classification Criteria
Parked	Groundspeed < 5 kts, altitude AGL $\leq$ 25 ft, within 5 km of origin or destination airport, outside flight time window
Takeoff	Within 5 km of origin, speed > 30 kts, altitude AGL < 1,000 ft, aligned with origin runway
Taxi-Out	Altitude AGL $\leq$ 25 ft, speed 5-30 kts, within 5 km of origin
Climb	Vertical rate > 500 ft/min, beyond 5 km from origin
Cruise	All remaining points above ground level
Approach	Vertical rate < -500 ft/min, altitude AGL < 10,000 ft
Descent	Vertical rate < -500 ft/min
Landing	Within 20 NM of destination, altitude AGL < 3,000 ft, speed 100-160 kts, aligned with destination runway ($\pm 15^\circ$)
Taxi-In	Altitude AGL $\leq$ 25 ft, speed 5-100 kts, within 5 km of destination

3.3.4 Timestamp Correction

Scheduled takeoff and landing timestamps recorded in the flightlist frequently deviate from actual gate-out and gate-in events due to airline reporting conventions. Corrected timestamps are derived by identifying, within the interpolated trajectory, the point nearest to the origin or destination airport at the lowest altitude, then adjusting the time based on the estimated vertical distance to the runway threshold using a standard climb or descent rate [44]. This correction improves the accuracy of segment-level elapsed time features and aligns flight duration estimates more closely with actual airborne time.

3.3.5 Takeoff Mass Estimation

The takeoff mass estimation combines OEW with estimated payload and fuel masses. Payload is computed using route-specific passenger load factors (defaulting to 85% when unavailable) and an average passenger weight of 100 kg including baggage. Fuel mass follows aviation standard planning: trip fuel (flight duration $\times$ average burn rate), 5% contingency, and 30-minute final reserve. The resulting mass estimate is constrained to not exceed the aircraft's certified MTOW.

3.3.6 Synthetic Widebody Data Augmentation

Widebody aircraft account for 18.7% of raw training segments, yet their high per-segment fuel burns mean that prediction errors on these segments contribute disproportionately to overall RMSE. To improve generalization on long-haul operations, additional synthetic widebody training samples are generated using empirical distribution sampling with Gaussian noise perturbation, bringing the widebody share of the augmented dataset to 34.1% (Table 3).

The choice of 24,996 synthetic widebody samples was determined through validation sweeps during the development phase. Increasing the synthetic dataset size beyond this threshold did not provide statistically significant improvements in validation RMSE, but it increased training time and memory footprint during hyperparameter tuning. This size was selected as the optimal trade-off between model generalization and computational cost.

Table 3. Training set composition before and after synthetic widebody augmentation, reported by aircraft category.

Aircraft Category	Original	Synthetic	Total
Narrowbody (A320, B737, etc.)	86,615	0	86,615
Widebody (A350, B777, etc.)	19,919	24,996	44,915
Total	106,534	24,996	131,530

For each widebody type, let X_w denote the set of real training observations, where each $x_i \in X_w$ is a vector of M numerical features. To ensure the model effectively captures long-haul operations, 25% of synthetic samples are generated by specifically sampling baseline segments ($x_b \in X_w$) from the top 25% of segment durations for each widebody type, while the remaining 75% are drawn from the general widebody population. All M numerical features of the selected baseline segment are then perturbed as:

$$\tilde{x}^{(j)} = x_b^{(j)} + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, (\alpha \sigma_j)^2) \quad (1)$$

where σ_j is the empirical standard deviation of feature j and $\alpha \sim \mathcal{U}[0.05, 0.15]$ is a percentage (representing 5% to 15% of the standard deviation σ_j) drawn independently per feature. This scale-invariant ratio prevents larger features from dominating the perturbation. Varying α dynamically during training acts as a regularizer to improve model robustness. A physical bounding operator $\mathcal{B}$ is finally applied to enforce kinematic constraints (e.g., altitude $z \in [0, 45,000]$ ft, groundspeed $v \in [0, 600]$ kts, non-negative mass):

$$\tilde{x}_{\text{final}}^{(j)} = \mathcal{B}_j(\tilde{x}^{(j)}) \quad (2)$$

3.4 Phase 3 - Feature Selection & Modeling

3.4.1 Model Development and Optimization

The XGBoost model development process proceeds through three sequential stages. It begins with target transformation and preprocessing, followed by feature selection via SFS, and concludes with Bayesian hyperparameter optimization.

3.4.2 Target Transformation and Preprocessing

Fuel consumption targets span two orders of magnitude across the fleet, from short descent segments on narrowbody aircraft to multi-hour cruise legs on widebodies. Applying a log-transformation to the target variable stabilizes variance and shifts the training objective towards relative rather than absolute errors, preventing large-fuel widebody segments from dominating the gradient updates of early trees. At inference time, predictions are exponentiated to recover fuel values in kg. Categorical features (aircraft type, origin, destination) are converted into integer labels, with aircraft-type labels assigned alphabetically rather than according to a physical ordering such as increasing MTOW. All numerical features are standardized using a scaler fitted exclusively on the training fold to prevent data leakage.

3.4.3 Feature Selection and Importance

Starting from 133 candidate features, Sequential Forward Selection (SFS) [45] is applied to identify the most informative subset of 74 variables. At each step, the feature whose addition produces the largest reduction in five-fold cross-validated mean squared error is added to the current set, and the process continues until adding any remaining feature fails to improve the score. The search is evaluated using an XGBoost model with fixed hyperparameters so that the feature ranking reflects the

predictive signal rather than tuning artifacts. To ensure computational efficiency, feature selection is performed once offline rather than nested inside the hyperparameter optimization loop, as nesting SFS within the cross-validation partitions would require training over 35,000 XGBoost models on a dataset of over 130,000 rows.

The retained set spans six functional groups covering mass-state features, physics-informed OpenAP outputs, kinematic aggregates, phase-fraction features, aircraft performance parameters, and fuel budget variables. Together, these capture the primary energetic, aerodynamic, and trajectory-level signals needed to predict fuel consumption per segment. Although several aircraft performance parameters are correlated with aircraft type and size and may therefore partly act as proxies for the aircraft category, they describe distinct physical and operational characteristics. Retaining these variables alongside the aircraft-type identifier allows the feature-selection procedure to evaluate them individually and select the aircraft properties that are most relevant to the prediction task. Features eliminated by SFS included redundant altitude endpoints, correlated runway heading columns, and aircraft geometry fields whose information was already subsumed by the retained aerodynamic coefficients. The complete list of 74 selected features and the SFS execution logs are available in the project's GitHub repository.

The feature correlation analysis in Fig. 3a, which focuses on the 15 variables most strongly correlated with the target from the SFS set, shows that interval duration, aircraft mass, and fuselage height are the strongest linear predictors of fuel consumption, with Pearson correlation coefficients of 0.79, 0.57, and 0.54, respectively. Wing MAC and great circle distance also exhibit strong positive correlations of 0.53 and 0.52, while drag C_{d0} contributes a moderate positive correlation of 0.50. Flaps Sf/S shows a moderate negative correlation (-0.46), and drag efficiency shows a negative correlation of -0.39 , consistent with the aerodynamic expectation that higher lift-to-drag ratios reduce thrust requirements and fuel burn. Flight-phase and operational variables such as average groundspeed and fuel coefficient show weaker associations of 0.26 and 0.23, while altitude-related features and vertical speed contribute minimally, with coefficients of 0.18, 0.07, and 0.06, indicating a secondary but noteworthy contribution from route and operational geometry.

In contrast to the linear correlations, the XGBoost gain-based feature importance analysis in Fig. 3b highlights the dominant role of aircraft physical characteristics and predefined performance metrics in the model's non-linear decision-making. The gain-based importance score is retrieved using the XGBoost `get_score` method with the importance type set to gain, which measures the average fractional contribution of each feature to the reduction of the mean squared error (MSE) across all splits in the ensemble of trees. Fuselage height accounts for the largest share of information gain at approximately 52.9%, acting as a hard discriminator between aircraft size categories since narrowbody and widebody types occupy distinct height ranges that directly correspond to differences in gross weight, frontal drag area, and thrust requirements. Flap geometry parameters (Flaps Sf/S) contribute roughly 11.0% of the total gain. While higher ratios are often associated with complex high-lift systems, in the dataset the negative correlation highlights that this feature serves as an indicator for aircraft scale because widebody aircraft generally possess lower flap-to-wing area ratios than narrowbodies, allowing the model to leverage this geometric signature to differentiate heavier, high-consumption airframes. Pre-calculated fuel consumption estimates (9.7%) are also very important, with the baseline estimate derived from known aircraft-specific fuel rates multiplied by the interval duration and the physics-based OpenAP estimate (1.4%) contributing predictive value. Operating specifications such as final reserve fuel (5.4%) and operating empty weight limits (Limits OEI) at 2.7% also contribute notable predictive value. Segment-specific kinematic variables such as interval duration (3.3%) and segment duration (2.3%) rank low in gain, yet this does not imply a weak physical relationship with fuel consumption because the Pearson correlation of 0.79 for interval duration confirms their relevance. Their low gain is instead a sign of redundancy, as the

physics-based fuel estimate and mass-state features already capture their contribution, leaving little additional variance for kinematic variables to explain.

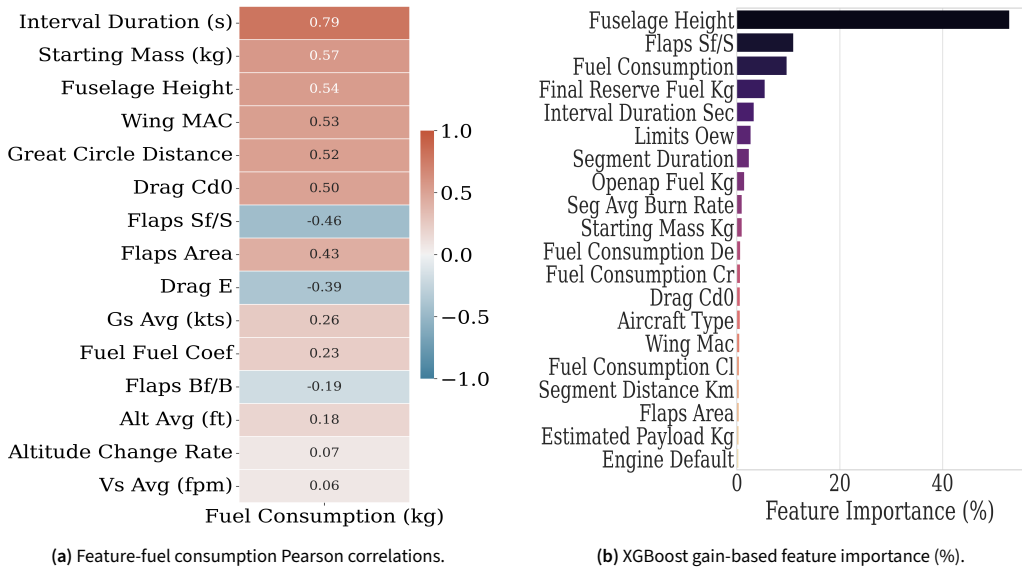


Figure 3. Feature analysis for the selected fuel-prediction variables: (a) Pearson correlations with fuel consumption and (b) XGBoost gain-based feature importance.

3.4.4 Hyperparameter Tuning and Final Ensemble

The model's representational power is refined through Bayesian optimization using the Optuna framework, which efficiently explores the high-dimensional configuration space defined in Table 4. This approach utilizes a Tree-structured Parzen Estimator (TPE) [46] sampler to iteratively focus the search on promising parameter regions, minimizing the cross-validation RMSE. Parameters such as tree depth, learning rate, and tree-level regularization terms like γ and λ are optimized. Stochasticity is introduced through row and column subsampling to improve the robustness of the ensemble. The final architecture consists of a vertically stacked set of gradient-boosted trees in which each subsequent tree minimizes the residual error of the preceding ensemble. During inference, the predictions are combined to generate the final segment-level fuel-consumption estimate.

To prevent overfitting to the cross-validation folds during hyperparameter optimization, a two-stage tuning strategy was implemented. First, Optuna was executed using 3-fold cross-validation on an 80% training split of the data to identify configuration candidates. Second, the top 10 hyperparameter configurations from this search were retrained on the full 80% training split and evaluated on the remaining unseen 20% validation holdout split. The single parameter configuration achieving the lowest validation RMSE on this holdout set was selected for the final model. This two-stage workflow ensures that the final parameters generalize well to unseen data by mitigating optimization leakage that can arise from evaluating many trials on the same cross-validation folds.

Design decisions are guided by a 20% held-out validation set of 31,306 segments. For test evaluation, all ML models are retrained on the full available training data.

Table 4. Hyperparameter search space and selected optimal values for the final XGBoost model.

Parameter Group	Search Range	Optimal Value
Tree Structure	max_depth [3, 12]	9
	n_estimators [500, 3000]	1455
	min_child_weight [1, 10]	4
Learning Dynamics	learning_rate [0.001, 0.1] (log)	0.029
	early_stopping_rounds	50
Regularization	gamma [10^{-8} , 1.0] (log)	6.2×10^{-8}
	lambda (L2) [10^{-8} , 10.0] (log)	2.4×10^{-8}
	alpha (L1) [10^{-8} , 10.0] (log)	4.9×10^{-3}
Sampling	subsample [0.6, 1.0]	1.00
	colsample_bytree [0.6, 1.0]	0.67

4. Performance Evaluation

Final generalization is assessed on the hidden test set, which consists of the rank and final datasets provided with the challenge and contains 61,745 segments. The OpenAP physics-based model, which leverages open-source aircraft performance parameters [36] to calculate fuel flow from kinematic state variables, serves as the baseline. However, its coverage is limited because OpenAP only supports a subset of the aircraft types present in its dataset. Specifically, the physics baseline covers approximately 54.5% of validation and 73% of test segments, whereas the proposed machine learning model achieves full coverage.

4.1 Overall Comparison with Baselines

Table 5 compares the OpenAP physics model, three standard ML baselines, and the proposed XGBoost model on both the held-out validation set (20% of the training set) and the hidden test set. The XGBoost reference uses the same hyperparameters as the proposed model but without SFS, isolating the effect of feature selection.

On the validation set, to account for statistical variations from the random seed split, metrics are reported as the mean and standard deviation across 5 random seeds (42, 43, 44, 45, and 46). Ridge Regression achieves a validation RMSE of 903.6 ± 81.1 kg and $R^2 = 0.2393 \pm 0.1485$, confirming the strongly non-linear nature of the task. Random Forest and LightGBM reduce validation RMSE to 229.5 ± 12.6 kg and 235.6 ± 14.6 kg, respectively, with R^2 values of 0.9513 ± 0.0044 and 0.9486 ± 0.0054 . The XGBoost reference reaches an RMSE of 200.3 ± 13.4 kg and $R^2 = 0.9628 \pm 0.0043$, demonstrating that the gradient boosting inductive bias is well suited to this tabular regression task. The proposed model, which incorporates SFS feature selection, synthetic widebody augmentation, and physics-informed features, achieves a validation RMSE of 203.1 ± 14.1 kg and $R^2 = 0.9618 \pm 0.0046$. While the full-feature reference model achieves a marginally lower mean RMSE on validation, the proposed SFS-tuned version demonstrates significantly better generalization on the hidden test set (RMSE 214.9 kg versus 223.2 kg), confirming that the compact feature subset effectively mitigates overfitting.

On the hidden test set, the proposed model achieves an RMSE of 214.92 kg and $R^2 = 0.920$, the lowest RMSE among all evaluated models, while LightGBM achieves 217.0 kg. OpenAP achieves a test RMSE of 193.6 kg on the approximately 73% of segments it covers and is therefore not directly comparable to the full-coverage ML models.

To provide a fair and direct comparison against the physics-based baseline, Table 6 evaluates the proposed XGBoost model, the model with the physics baseline feature removed, and the OpenAP

FuelFlow model exclusively on the flight segments where OpenAP successfully computed a prediction ($N = 17,073$ for validation and $N = 45,324$ for test). On these subsets, the proposed model achieves an RMSE of 117.62 kg and 126.18 kg, respectively, while the model trained without the OpenAP feature achieves 115.30 kg and 126.80 kg, both outperforming the OpenAP baseline estimates of 218.90 kg and 193.58 kg. Although the model without the physics feature achieves a slightly lower validation error, including OpenAP as an input feature improves prediction accuracy on the unseen test dataset. Overall, these results demonstrate that even without direct access to the physics estimate, the data-driven framework recovers the physical relations and outperforms the baseline by a large margin.

Table 5. Overall model performance on the validation set and hidden PRC 2025 test set, reported using MAE, RMSE, and R^2 .

Model	Validation (Mean $\pm$ Std)			Test (hidden, rank + final)		
	MAE	RMSE	R^2	MAE	RMSE	R^2
Ridge Regression	263.2 $\pm$ 3.8	903.6 $\pm$ 81.1	0.2393 $\pm$ 0.1485	162.1	660.7	0.268
Random Forest	90.5 $\pm$ 0.8	229.5 $\pm$ 12.6	0.9513 $\pm$ 0.0044	71.7	229.6	0.912
LightGBM	94.5 $\pm$ 1.1	235.6 $\pm$ 14.6	0.9486 $\pm$ 0.0054	71.1	217.0	0.921
XGBoost (reference, no SFS)	85.0 $\pm$ 0.8	200.3 $\pm$ 13.4	0.9628 $\pm$ 0.0043	71.0	223.2	0.917
XGBoost (proposed)	85.6 $\pm$ 1.0	203.1 $\pm$ 14.1	0.9618 $\pm$ 0.0046	70.6	214.92	0.920

Table 6. Performance comparison on the exact subset of validation ($N = 17,073$) and test ($N = 45,324$) segments with OpenAP coverage.

Model	Validation	Test
	RMSE	RMSE
OpenAP FuelFlow (physics)	218.90	193.58
XGBoost (proposed)	117.62	126.18
XGBoost (without openap_fuel_kg)	115.30	126.80

4.2 Performance by Flight Phase

The distribution of fuel consumption rates is highly dependent on the operational flight phase and aircraft category. As shown in the phase-wise analysis in Fig. 4, the climb phase exhibits the highest fuel flow rates with significant variance, reflecting the high-thrust requirements. Conversely, the cruise phase demonstrates a concentrated distribution at lower flow rates.

Table 7 reports per-phase performance alongside sample sizes on the validation and hidden test sets. The hidden test set is heavily dominated by cruise segments ($N = 44,100$), which typically span the majority of long-haul flights, compared to descent ($N = 9,656$) and climb ($N = 7,946$). On the validation set, cruise segments are the most accurately predicted (RMSE 204.4 kg, $R^2 = 0.964$). Climb consistently carries the highest absolute error on both the validation (RMSE 212.0 kg) and test sets (RMSE 281.9 kg), despite maintaining a sample size of nearly 8,000 test segments. This consistent pattern confirms that fuel-burn variability during climb is driven by complex physical dynamics. Conversely, descent achieves the lowest MAE on the test set (50.2 kg), as engines during this phase are frequently operating near idle thrust with highly predictable, low-variance fuel flows.

4.3 Performance by Aircraft Type

Table 8 reports per-type metrics for aircraft types with at least 100 test segments. Narrowbody types show consistently low errors on both validation and test. The A20N achieves a test MAE of 34.1 kg

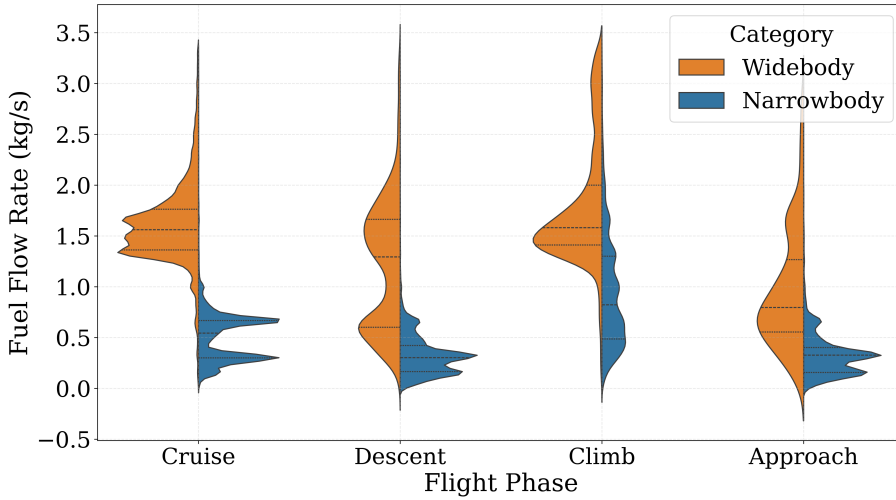


Figure 4. Phase-wise fuel-flow distribution (kg/s), showing differences in fuel-flow rates across climb, cruise, and descent segments.

Table 7. Per-phase performance of the proposed model on the validation and hidden test sets, reported for climb, cruise, and descent segments.

Phase	Validation				Test (hidden)			
	<i>N</i>	MAE	RMSE	R^2	<i>N</i>	MAE	RMSE	R^2
Climb	3,234	97.6	212.0	0.957	7,946	99.8	281.9	0.879
Cruise	24,006	86.1	204.4	0.964	44,100	69.7	216.2	0.928
Descent	3,917	73.0	180.7	0.927	9,656	50.2	154.4	0.900

and the A320 reaches 40.1 kg. Widebody types carry larger absolute errors owing to greater fuel burn magnitude, with A359 at 328.1 kg test RMSE and B789 at 303.1 kg. The A332 shows the lowest test R^2 at 0.736, likely due to limited training data relative to its fuel burn variability. For most types, validation and test metrics are broadly consistent, which confirms that the model generalizes across the fleet. The B77W is a notable exception, with validation RMSE of 444.5 kg and test RMSE of 871.5 kg, indicating distribution shift between the training and test populations for that type. Fig. 5 shows the mean predicted and actual fuel burn per type on the validation set.

4.4 Performance by Flight Haul Category

To understand where prediction error concentrates as a function of mission length, segments are grouped into haul categories based on total flight duration: short-haul (≤ 3 h), medium-haul (3-6 h), and long-haul (> 6 h). Table 9 reports performance on the held-out validation set and the hidden test set for each group.

On the validation set, short-haul segments show a relatively high RMSE of 161.0 kg and lower R^2 of 0.880, reflecting the large relative variability of fuel burn for short, mixed-phase intervals. Medium-haul segments are the most accurately predicted in absolute terms (RMSE 82.0 kg, $R^2 = 0.930$), benefiting from the concentration of narrowbody cruise operations in this range. Long-haul segments show the highest absolute errors (RMSE 212.9 kg) but the highest R^2 (0.967), consistent with the larger and more predictable fuel burns of widebody cruise legs. On the hidden test set, short-haul RMSE improves to 126.6 kg while long-haul error increases to 375.3 kg.

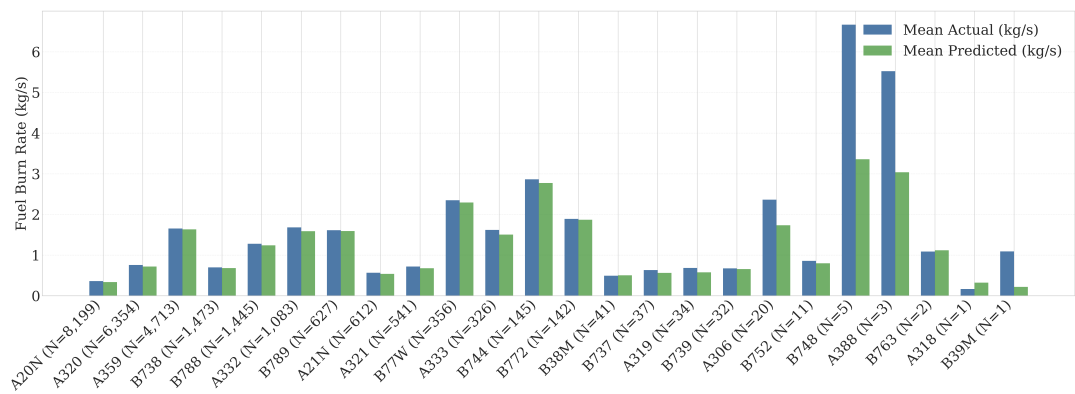


Figure 5. Mean predicted vs. actual fuel burn rate (kg/s) per aircraft type

Table 8. Per-type results for types with $N_{\text{test}} \geq 100$.

Validation					Test (hidden)			
Type	N	MAE	RMSE	R ²	N	MAE	RMSE	R ²
A20N	8,199	34.8	68.8	0.885	29,084	34.1	73.0	0.854
A320	6,354	43.4	101.9	0.914	13,035	40.1	67.8	0.958
A359	7,488	130.2	224.8	0.966	9,170	135.1	328.1	0.921
A332	1,680	174.4	408.5	0.902	2,201	160.2	378.1	0.736
B738	1,473	43.1	73.4	0.903	1,904	53.8	124.1	0.833
B789	969	124.2	200.4	0.976	1,437	140.0	303.1	0.936
B77W	547	249.0	444.5	0.939	1,031	403.7	871.5	0.796
A21N	612	44.4	86.8	0.782	954	38.8	90.6	0.807
B788	2,309	105.4	162.8	0.952	737	100.1	174.8	0.945
B744	212	263.9	558.3	0.908	582	399.3	800.0	0.790
A333	492	181.6	518.5	0.833	527	129.3	201.0	0.899
A321	541	61.4	158.3	0.720	307	122.1	249.7	0.719
B772	237	190.1	286.9	0.973	220	139.0	216.9	0.949

Table 9. Performance by haul category on validation and hidden test sets.

Validation				Test (hidden)		
Haul	N	RMSE	R ²	N	RMSE	R ²
Short (≤ 3 h)	4,490	161.0	0.880	7,048	126.6	0.857
Medium (3–6 h)	12,313	82.0	0.930	36,606	87.4	0.902
Long (> 6 h)	9,396	212.9	0.967	18,083	375.3	0.901

4.5 Contribution Ablation Study

Table 10 reports the ablation results. Augmentation reduces overall RMSE by 39.4 kg (16.1%) and MAE by 10.3 kg (10.7%), with the majority of gains concentrated in widebody segments (RMSE -61.8 kg, MAE -23.0 kg). Narrowbody metrics are essentially unchanged, confirming that the synthetic samples do not distort the narrowbody feature distribution.

To assess the specific impact of the Gaussian perturbation used during data augmentation, an ablation test was conducted comparing the proposed perturbed augmentation against a simple sample duplication scheme with 0% noise perturbation. Training the model with duplicate synthetic widebody samples instead of perturbed ones increases the overall validation MAE from 66.03 kg to 66.34 kg and the RMSE from 165.16 kg to 165.38 kg. For widebody segments, using simple duplication increases the MAE from 115.64 kg to 116.35 kg. For narrowbody segments, the MAE increases from 40.57 kg to 40.68 kg, and the RMSE increases from 86.58 kg to 87.52 kg. While the performance difference is small, adding stochastic perturbation to the augmented samples provides a regularizing effect that helps the model generalize beyond the exact observed trajectories.

Table 10. Validation-set ablation showing the effect of synthetic widebody augmentation on overall, narrowbody, and widebody performance.

Model	Split	N	MAE	RMSE	R^2
No Augmentation	Overall	31,306	96.0	244.0	0.945
	Narrowbody	17,356	40.7	96.8	0.888
	Widebody	13,950	164.8	349.2	0.926
With Augmentation	Overall	31,306	85.7	204.6	0.961
	Narrowbody	17,356	40.5	95.5	0.891
	Widebody	13,950	141.8	287.4	0.950

To quantify the contribution of each design component, an ablation study is conducted around the proposed configuration, which combines SFS feature selection, synthetic widebody augmentation, and METAR weather features. Each condition either removes one element or substitutes a single feature group while all other settings remain fixed. Design choices are motivated on the validation set and consistency is verified on the hidden test set. Specifically, the baseline models were evaluated against five variations, removing both synthetic widebody instances and METAR observations to isolate the SFS-only performance, omitting the load-factor feature group ($-C2$) to assess the impact of regional payload adjustments, replacing the dynamically tracked starting mass with the flight's static MTOW ($-C3$), adding the actual elapsed flight-time ($+C4$) which had been intentionally excluded by the algorithm, and removing the dynamic physics-based `openap_fuel_kg` feature ($-C5$).

Table 11 presents the validation and test set performance of the proposed model against five ablated conditions, isolating the contribution of synthetic data augmentation with METAR features, load-factor variables, dynamic mass initialization, corrected elapsed time, and physics-informed baseline features. Removing both synthetic augmentation and METAR features raises validation MAE by 10.3 kg (12.0%) and RMSE by 39.4 kg (16.1%), confirming these two components together are the dominant contributors to overall accuracy. Removing the load-factor feature group ($-C2$) worsens validation MAE by 1.4 kg (1.6%), confirming that regional payload estimates carry a meaningful signal. Replacing dynamic mass initialization with static MTOW ($-C3$) raises validation MAE by 11.9 kg (13.9%), showing that accurate mass tracking at segment start is important. Adding corrected elapsed flight-time features ($+C4$) produces no further improvement, suggesting the SFS-selected features already capture timing information adequately. On the hidden test set, differences between all conditions fall within a MAE of 0.3 kg, consistent with the larger test set size reducing sensitivity

to individual feature choices. Removing the physics-guided `openap_fuel_kg` feature (–C5) decreases validation MAE to 85.5 kg and RMSE to 202.6 kg on validation, but increases test RMSE to 222.2 kg and MAE to 73.4 kg on the final test set, showing its contribution to generalization on unseen test flights.

Table 11. Component ablation results showing the contribution of augmentation, METAR features, load factors, dynamic mass tracking, and elapsed-time features.

Condition	Validation			Test (hidden)		
	MAE	RMSE	R^2	MAE	RMSE	R^2
XGBoost (proposed)	85.7	204.6	0.961	70.6	214.92	0.920
No Synthetic augmentation, no METAR (SFS only)	96.0	244.0	0.945	70.6	216.6	0.921
Without load-factor features (–C2)	87.1	205.8	0.961	70.3	220.1	0.919
Static MTOW instead of dynamic mass (–C3)	97.6	238.3	0.948	70.4	218.7	0.920
With corrected elapsed time (+C4)	85.7	203.1	0.962	70.3	217.9	0.920
Without <code>openap_fuel_kg</code> (–C5)	85.5	202.6	0.962	73.4	222.2	0.917

4.6 Learning Dynamics

Fig. 6a shows training and validation RMSE versus boosting round. The training RMSE continues to decline while the validation RMSE plateaus after approximately 1,000 rounds. Although a gap remains between the two curves, the stable validation error indicates that the model is not overfitting to the training data. This generalization is further supported by the model's robust performance on the unseen hidden test sets. Fig. 6b shows the 5-fold cross-validated RMSE as training set size grows from 10% to 100%. The curve continues to decline at full scale, suggesting that additional labeled data could further improve accuracy. The narrow Interquartile Range (IQR) band confirms stable results across folds.

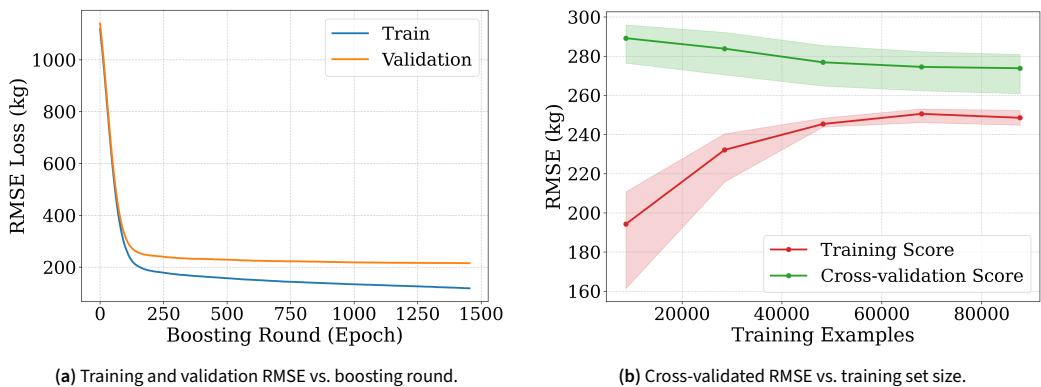


Figure 6. Learning dynamics of the final XGBoost model, showing RMSE convergence across boosting rounds and cross-validated RMSE as training-set size increases.

4.7 Challenge Results

Table 12 summarizes the RMSE achieved on these blind test sets [34]. The model demonstrates consistent performance across both challenge stages, with a combined RMSE of 214.92 kg over 61,745 flight segments.

Table 12. Model performance on the PRC 2025 challenge datasets

Dataset	Segments	RMSE (kg)
Rank Dataset	24,289	214.11
Final Dataset	37,456	215.44
Combined	61,745	214.92

5. Conclusion

This paper presented a physics-guided gradient boosting framework for segment-level aviation fuel consumption estimation, developed as our solution to the EUROCONTROL PRC 2025 Data Challenge. The proposed pipeline addresses three core challenges of open-data aviation modeling: noisy and incomplete ADS-B trajectories, heterogeneous aircraft types with no access to proprietary flight records, and a severe class imbalance between narrowbody and widebody operations.

The key contributions are as follows. A modular preprocessing pipeline filters and interpolates raw ADS-B trajectory data, corrects gate-to-gate timestamps, and injects METAR meteorological observations and airport infrastructure features into a unified per-segment feature vector. A physics-informed augmentation step runs the OpenAP FuelFlow model with dynamic mass tracking across the full trajectory of each flight, producing physically consistent fuel burn estimates and mass state features that serve as a strong learned prior. A targeted synthetic widebody augmentation strategy generates 24,996 perturbed samples to address the imbalance between narrowbody and widebody operations, reducing overall RMSE by 16.1% and widebody RMSE by 17.7%. SFS reduces the candidate feature space from 133 to a compact, interpretable subset, and Bayesian hyperparameter optimization via Optuna identifies an XGBoost configuration that minimizes five-fold cross-validated RMSE.

On the validation set of 31,306 segments, the proposed model achieves an MAE of 85.7 kg, RMSE of 204.6 kg, and $R^2 = 0.961$. When evaluated on the official blind test sets of the PRC 2025 Data Challenge, the framework demonstrates robust generalization, achieving an RMSE of 214.11 kg on the Rank dataset and 215.44 kg on the Final dataset, with a combined performance of 214.92 kg. These results validate the efficacy of the physics-informed data augmentation and the targeted widebody synthetic sampling strategy.

Finally, these results demonstrate that competitive fuel consumption estimation accuracy is achievable entirely from publicly available data sources, without reliance on proprietary airline records, by combining trajectory preprocessing, physics-informed feature engineering, and a well-regularized gradient boosted ensemble. The gap between internal validation (204.6 kg RMSE) and blind testing (214.92 kg RMSE) highlights the inherent difficulty of predicting fuel consumption on entirely unseen out-of-sample flights and days, yet the model maintains a high level of predictive precision across both scenarios.

6. Future Work

The first direction concerns the development of phase-specific sub-models. The current single-model approach is trained across all flight phases simultaneously, which requires the ensemble to learn phase-specific aerodynamics implicitly from the phase-fraction features. A hierarchical architecture that first assigns each segment to its dominant flight phase and then applies a dedicated regressor would allow each sub-model to specialize in the aerodynamic regime of that phase, reducing the inherent trade-off between generalization and phase-level precision. This is particularly relevant for long-haul widebody segments, which carry the highest validation RMSE (212.9 kg) and test RMSE

(375.3 kg) due to the high proportion of mixed-phase intervals and distribution shift between training and test populations, and for short-haul segments where the high relative variability of fuel burn across mixed-phase intervals contributes elevated error (161.0 kg validation RMSE) relative to their short mission duration.

The second direction concerns aircraft-specific fine-tuning. The per-type evaluation reveals non-trivial variance in predictive accuracy, with R^2 ranging from 0.719 on the A321 to 0.958 on the A320. This spread reflects the unequal distribution of training examples across the 36 aircraft types in the dataset. Aircraft types with a large number of records generalize well, while rarer types are underrepresented. A natural extension would be to train dedicated models stratified by aircraft category, for example a separate model for narrowbody aircraft and another for widebody aircraft, or finer-grained models targeting individual aircraft types. This approach would allow each model to specialize on the aerodynamic and operational characteristics of its category, potentially reducing the noise introduced by pooling structurally dissimilar aircraft into a single model.

A further area of improvement is using more rigorous cross-validation during training. Implementing multiple randomized splits or flight-level resampling would create a more robust model that generalizes better to unseen data. Additionally, future work should expand the data augmentation strategy. Applying targeted synthetic augmentation to severely underrepresented aircraft types (such as the B77W, which struggled during testing) could significantly improve performance across the entire fleet.

Acknowledgement

This work was supported by the European Union's Horizon Europe research and innovation program under grant agreement No 101187121 (EUSOME). It was also partially supported from the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

Author contributions

- Yiannis Grigoriou: Conceptualization, Methodology, Software (machine learning model, OpenAP integration, data augmentation, trajectory interpolation), Analysis, Visualization, Writing (Original Draft), Writing (Review and Editing)
- Eftychios Eftychiou: Software (timestamp correction, METAR integration, trajectory preprocessing, trajectory interpolation, data augmentation, dataset construction), Data Curation, Writing (Review and Editing)
- Christian Vitale: Writing (Original Draft, Related Work synthesis), Supervision, Writing (Review and Editing)
- Giorgos Petteimerides: Writing (Original Draft, Related Work), Writing (Review and Editing)
- Nicolas Souli: Writing (Original Draft, Related Work), Writing (Review and Editing)
- Panayiotis Kolios: Supervision, Writing (Review and Editing)

Open data statement

The primary flight trajectory and fuel consumption data used in this study were provided by the EUROCONTROL Performance Review Commission (PRC) 2025 Data Challenge, available at <https://zenodo.org/records/19184662>. Physical aircraft performance coefficients were sourced from the open-access OpenAP library [36]. Meteorological observations (METAR) for departure and arrival airports were retrieved from the Iowa Environmental Mesonet (IEM) ASOS/AWOS archives [40].

Airport infrastructure metadata, including elevations and runway configurations, were enriched by scraping aeronautical information from the SkyVector [35] aeronautical chart service. Finally, the processed feature datasets, validation metrics, and ablation results generated during this study are available in our project repository at <https://github.com/eeftychiou/PRCXGBoost>, with the official competition submission hosted at <https://github.com/PRC-Data-Challenge-2025/bright-lobster>.

Reproducibility statement

The full source code for the data preprocessing pipeline, feature engineering, model training, hyperparameter tuning, evaluation, and figure generation is available at <https://github.com/eeftychiou/PRCXGBoost>. The official competition submission repository, as submitted to the PRC 2025 Data Challenge, is available at <https://github.com/PRC-Data-Challenge-2025/bright-lobster>. The first repository includes all scripts necessary to reproduce the results reported in this paper, along with the processed outputs used to generate the tables and figures.

References

- [1] Gabriel Jarry, Ramon Dalmau, Philippe Very, and Junzi Sun. “Aircraft fuel flow modelling with ageing effects: From parametric corrections to neural networks”. In: *arXiv preprint arXiv:2509.15736* (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. doi: <https://doi.org/10.1002/acs.1176>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/acs.1176>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/acs.1176>.
- [3] Yashovardhan S Chati and Hamsa Balakrishnan. “Statistical modeling of aircraft engine fuel flow rate”. In: *30th congress of the international council of the aeronautical science (ICAS)*. 2016, pp. 1–10.
- [4] Navinda Kithmal Wickramasinghe, Daichi Toratani, Sachiko Fukushima, and Hiroko Hirabayashi. “A study on modeling techniques for fuel burn estimation based on flight simulator experiment data”. In: *Proceedings of the 2017 Winter Simulation Conference*. WSC ’17. Las Vegas, Nevada: IEEE Press, 2017. ISBN: 9781538634271.
- [5] Manuel de Jesús Gurrola Arrieta, Ruxandra M. Botez, and Axel Lasne. “An Engine Deterioration Model for Predicting Fuel Consumption Impact in a Regional Aircraft”. In: *Aerospace* 11.6 (2024). ISSN: 2226-4310. URL: <https://www.mdpi.com/2226-4310/11/6/426>.
- [6] Ziming Wang, Dabin Xue, Lingxiao Wu, and Ran Yan. “A reliable predict-then-optimize approach for minimizing aircraft fuel consumption”. In: *Transportation Research Part D: Transport and Environment* 142 (2025), p. 104693.
- [7] 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>.
- [8] 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 (Mar. 2026). doi: 10.59490/joas.2026.8750. URL: <https://journals.open.tudelft.nl/joas/article/view/8750>.
- [9] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. “Optuna: A Next-generation Hyperparameter Optimization Framework”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD ’19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2623–2631. ISBN: 9781450362016. doi: 10.1145/3292500.3330701. URL: <https://doi.org/10.1145/3292500.3330701>.

- [10] Yashovardhan S Chati and Hamsa Balakrishnan. "Data-driven modeling of aircraft engine fuel burn in climb out and approach". In: *Transportation Research Record* 2672.29 (2018), pp. 1–11.
- [11] Enis T Turgut, Oznur Usanmaz, Mustafa Cavcar, Tuncay Dogeroglu, and Kadir Armutlu. "Effects of descent flight-path angle on fuel consumption of commercial aircraft". In: *Journal of Aircraft* 56.1 (2019), pp. 313–323.
- [12] Ridvan Oruc and Tolga Baklacioglu. "Modeling of fuel flow-rate of commercial aircraft for the descent flight using particle swarm optimization". In: *Aircraft Engineering and Aerospace Technology* 93.2 (2021), pp. 319–326.
- [13] Vehbi Emrah Atasoy. "Fuel estimation of commercial aircraft for the climb-out phase using gaussian process regression model". In: *International Journal of Energy Studies* 7.2 (2022), pp. 179–194.
- [14] S Baumann, T Neidhardt, and U Klingauf. "Evaluation of the aircraft fuel economy using advanced statistics and machine learning". In: *CEAS Aeronautical Journal* 12.3 (2021), pp. 669–681.
- [15] Runyuan Zou, Bing Wang, Kun Wang, Wen-Long Shang, Dabin Xue, and Washington Yotto Ochieng. "A pathway to sustainable aviation: Modeling aircraft takeoff mass for precise fuel consumption and aircraft emission calculations". In: *Energy* 319 (2025), p. 135074.
- [16] Klingauf Baumann S et al. "Modeling of aircraft fuel consumption using machine learning algorithms". In: *CEAS Aeronautical Journal* 11.1 (2020), pp. 277–287.
- [17] Wang Zhao Yuandi et al. "Data driven fuel consumption prediction model for green aviation using radial basis function neural network". In: *Scientific Reports* 15.1 (2025), p. 26275.
- [18] Bukhari Wahid Mirza Anas et al. "Parametric estimation scheme for aircraft fuel consumption using machine learning". In: *Neural Computing and Applications* 35.35 (2023), pp. 24925–24946.
- [19] Dai Tang Weizhen et al. "Aircraft range fuel prediction study based on WPD with IAPO optimized BiLSTM-KAN model". In: *Scientific Reports* 15.1 (2025), p. 12231.
- [20] Cheng Huang Chenyu et al. "Estimation of aircraft fuel consumption by modeling flight data from avionics systems". In: *Journal of Air Transport Management* 99 (2022), p. 102181.
- [21] Junzi Sun, Aidana Tassanbi, Piotrek Obojski, and Philip Plantholt. "Evaluating transatlantic flight emissions and inefficiencies using space-based ADS-B data". In: *Proceedings of the 13th SESAR innovation days, sevilla, spain* (2023), pp. 27–30.
- [22] Flávio DA Quadros, Mirjam Snellen, Junzi Sun, and Irene C Dedoussi. "Global civil aviation emissions estimates for 2017–2020 using ADS-B data". In: *Journal of Aircraft* 59.6 (2022), pp. 1394–1405.
- [23] 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 (2025).
- [24] Linfeng Zhang, Lei Bian, Changmin Jiang, and Lingxiao Wu. "A comprehensive framework for estimating aircraft fuel consumption based on flight trajectories". In: *Transportation Research Part E: Logistics and Transportation Review* 203 (2025), p. 104339. ISSN: 1366-5545. DOI: <https://doi.org/10.1016/j.tre.2025.104339>. URL: <https://www.sciencedirect.com/science/article/pii/S1366554525003801>.
- [25] Junzi Sun, Xavier Olive, Martin Strohmeier, and Vincent Lenders. "OpenSky Report 2025: Improving Crowdsourced Flight Trajectories with ADS-C Data". In: *2025 Integrated Communications, Navigation and Surveillance Conference (ICNS)*. IEEE. 2025, pp. 1–8.
- [26] Junzi Sun, Luis Basora, Xavier Olive, Martin Strohmeier, Matthias Schäfer, Ivan Martinovic, and Vincent Lenders. "OpenSky Report 2022: Evaluating aviation emissions using crowdsourced open flight data". In: *2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC)*. IEEE. 2022, pp. 1–8.

- [27] Junzi Sun and Irene Dedoussi. “Evaluation of aviation emissions and environmental costs in Europe using OpenSky and OpenAP”. In: *Engineering Proceedings* 13.1 (2021), p. 5.
- [28] Jefry Yanto and Rhea P Liem. “Cluster-based aircraft fuel estimation model for effective and efficient fuel budgeting on new routes”. In: *Aerospace* 9.10 (2022), p. 624.
- [29] Gizem Ceylan and Metin Turan. “Modeling Fuel Consumption Curves in Jet UAVs with Machine Learning: Jet Motorlu İHA’larda Yakıt Tüketim Eğrilerinin Makine Öğrenmesi ile Modellenmesi”. In: *Journal of Aeronautics and Space Technologies* (2026), pp. 46–66.
- [30] Yingying Fang, Yunqi Ji, and Renshu Gu. “Artificial Intelligence and Large Foundation Models for Intelligent Operation Control in Civil Aviation: A Survey”. In: *Proceedings of the 4th International Conference on Artificial Intelligence and Intelligent Information Processing*. 2025, pp. 258–263.
- [31] Sahithi Varshini A Teja et al. “Performance evaluation of machine learning and deep learning models for fuel consumption prediction”. In: *2025 International Conference on Sustainable Communication Networks and Application (ICSCN)*. IEEE. 2025, pp. 1904–1909.
- [32] Xianwei Xie, Baozhi Sun, Xiaohe Li, Tobias Olsson, Neda Maleki, and Fredrik Ahlgren. “Fuel consumption prediction models based on machine learning and mathematical methods”. In: *Journal of Marine Science and Engineering* 11.4 (2023), p. 738.
- [33] Afiya Ayman, Michael Wilbur, Amutheezan Sivagnanam, Philip Pugliese, Abhishek Dubey, and Aron Laszka. “Data-driven prediction of route-level energy use for mixed-vehicle transit fleets”. In: *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE. 2020, pp. 41–48.
- [34] Junzi Sun, Enrico Spinielli, and Martin Strohmeier. *Aircraft Fuel Burn Estimation Data Challenge*. Zenodo, Mar. 2026. DOI: 10.5281/zenodo.19184662. URL: <https://doi.org/10.5281/zenodo.19184662>.
- [35] SkyVector. *SkyVector: Flight Planning / Aeronautical Charts*. <https://skyvector.com>. 2026.
- [36] Junzi Sun, Jacco M. Hoekstra, and Joost Ellerbroek. “OpenAP: An Open-Source Aircraft Performance Model for Air Transportation Studies and Simulations”. In: *Aerospace* 7.8 (2020). ISSN: 2226-4310. DOI: 10.3390/aerospace7080104. URL: <https://www.mdpi.com/2226-4310/7/8/104>.
- [37] IATA Sustainability & Economics. *Air Passenger Market Analysis – August 2025*. Tech. rep. Published 30 September 2025. International Air Transport Association (IATA), Sept. 2025. URL: <https://www.iata.org/en/iata-repository/publications/economic-reports/air-passenger-market-analysis-august-2025/> (visited on 03/13/2026).
- [38] International Air Transport Association. *IATA Annual Review 2024: Record Load Factors and Industry Performance*. 2024. URL: <https://www.iata.org/en/publications/annual-review/>.
- [39] Junzi Sun. *Fuel and Emission — The OpenAP Handbook*. https://openap.dev/fuel_emission.html. Accessed: 2025. 2025.
- [40] Iowa Environmental Mesonet. *ASOS/AWOS/METAR Data Archive*. <https://mesonet.agron.iastate.edu/request/download.phtml>. Accessed: 2025-10. 2025.
- [41] Xavier Olive. “traffic, a toolbox for processing and analysing air traffic data”. In: *Journal of Open Source Software* 4.39 (2019), p. 1518. DOI: 10.21105/joss.01518. URL: <https://doi.org/10.21105/joss.01518>.
- [42] Junzi Sun. *Flight Phase Identification — The OpenAP Handbook*. https://openap.dev/flight_phases.html. Accessed: 2025. 2025.
- [43] 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 (ICRAT)*. Ed. by D. Lovell and H. Fricke. Philadelphia, USA, 2016. URL: <https://repository.tudelft.nl/record/uuid:af67a6bd-d812-474d-a304-7a594991390b>.
- [44] Team Likable Jelly. *PRC Data Challenge 2024 Submission: Timestamp Correction and Weight Estimation*. https://github.com/PRC-Data-Challenge-2024/team_likable_jelly. Accessed: 2025-10. 2024.

- [45] David W. Aha and Richard L. Bankert. “A Comparative Evaluation of Sequential Feature Selection Algorithms”. In: *Pre-proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*. Ed. by Doug Fisher and Hans-Joachim Lenz. Vol. R0. Proceedings of Machine Learning Research. Reissued by PMLR on 01 May 2022. PMLR, Apr. 1995, pp. 1–7. URL: <https://proceedings.mlr.press/r0/aha95a.html>.
- [46] Shuhei Watanabe. “Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance”. In: *arXiv preprint arXiv:2304.11127* (2023).