Contingency Fuel Prediction with Explainable Machine Learning for Airline Operations

Phillipe Lothaller, Marta Ribeiro, Junzi Sun
Operations & Environment, Aerospace Faculty
Delft University of Technology
Delft, The Netherlands

Jasper de Wilde, Alexander Piva
ATM Strategy and Tactical Planning
Koninklijke Luchtvaart Maatschappij N.V.
Amsterdam, The Netherlands

***Abstract*—Aircraft carry additional fuel reserves, referred to as contingency fuel, used to account for unforeseen events during a flight. Previous research has attempted to quantify the magnitude of such events, most notably the probability of adverse weather or ATFM regulation, yet their inherent unpredictability introduces uncertainty and frequently results in the overestimation of contingency fuel requirements. Recent studies use data-driven fuel-burn predictions to better estimate contingency fuel sizing; however, most are confined to specific routes or regions, limiting generalizability. To address this, we utilise real operational airline data covering both regional and intercontinental flights, and develop a quantile regression framework for predicting contingency fuel requirements, capable of adapting to more diverse set of flight characteristics. Our framework integrates flight-plan data, TAF weather forecasts, and proxy congestion features to predict required contingency fuel at varying quantile levels, enabling trade-offs between efficiency and safety. Unlike the current Statistical Contingency Fuel process, which applies different coverage levels by risk category, this evaluation uses a single fixed quantile for all flights when generating predictions. In a four-month out-of-sample evaluation, a single fixed quantile matched the safety performance of the Statistical Contingency Fuel process while reducing excess fuel carriage by up to 235,364 kg (≈11%). A more conservative quantile configuration yielded smaller savings but reduced abnormal flight-phase events by 22.2%. The key drivers of the final predictions are evaluated, offering pilots and dispatchers transparent explanations that can build trust and reduce reliance on discretionary fuel loading.**

***Keywords*—Airline Operations, Contingency Fuel, Explainable Machine Learning, Gradient Boosting, Tree Based Learning Algorithms, Quantile Regression**

I. Introduction

The aviation industry has set ambitious climate goals, aiming to reduce CO_2 emissions by 50% by 2050 compared to 2005 levels, and to achieve net-zero emissions for all flights within and departing from the European Union by the same year [1]. One of the main areas identified by the study airline to achieve these targets is aircraft operations, with measures such as route optimisation, fuel-efficient technologies, and aircraft weight reduction, collectively aiming for a 2–4% reduction in fuel consumption by 2030 [2]. Fuel loading practices, particularly the allocation of contingency and discretionary fuel, have been identified as an area for potential fuel savings in commercial aviation. The former, *contingency fuel* refers to the fuel carried on board to ensure safe flight operations under unforeseen circumstances, such as air traffic delays, deviations from the planned route, or unexpected weather conditions. *Discretionary fuel*, by contrast, is fuel added at the pilot's discretion to account for specific operational uncertainties or perceived risks [3–6].

Unforeseen circumstances are naturally difficult to predict, leading airlines to have a conservative approach to the allocation of contingency fuel. Ryerson et al. [3] found that up to 4.48% of an average flight's fuel consumption results from carrying excess fuel, with unnecessary contingency fuel alone accounting for as much as 1.04%. Kang et al. [4] estimated that in 2012, six major U.S. airlines incurred an additional $1.46 billion in costs from carrying unused fuel. Hao et al. [7] aimed to investigate the relationship, estimating that carrying 6.12 minutes of contingency fuel costs an additional 21.93 kg of fuel per flight. These weight penalties translate to annual costs of $120.55–$452.43 million for the studied U.S. domestic airlines.

Reducing unnecessary fuel loading offers significant monetary and emissions savings, yet relatively few studies have addressed the prediction of contingency fuel requirements. Traditional methods, such as the simple statistical techniques currently employed by the study airline, remain the industry norm. However, recent research has explored more advanced approaches, including ensemble learning and deep neural networks, to improve the accuracy of fuel burn and delay predictions, showing promise for more effective fuel planning.

Despite these advances, existing work remains limited in scope, often focusing on specific routes/regions and providing little transparency in model outputs. Two key gaps emerge: the need for approaches that generalize across an airline's full operational fleet, and the need for explainable models that pilots and dispatchers can trust. This paper addresses these gaps by applying machine learning to contingency fuel prediction in a fleet-wide context, using a simple, interpretable tree-based LightGBM ensemble model to deliver accurate and transparent results.

The research paper is structured as follows: Section II introduces key concepts and definitions related to current industry fuelling practices, with a focus on the study airline. Section III reviews prior research on contingency and discretionary fuel loading, while Section IV outlines the methods used in this study. Section V highlights the steps taken to

apply the methods using the provided operational data. The results are presented in Section VI, followed by discussion and recommendations in Section VII. The study is concluded in Section VIII.

II. Problem Definition

This section provides more information on fuel loading practices. Subsection II-A categorizes all the fuel components in a flight, while Subsection II-B defines the concept of abnormal phase, a crucial safety measure in fuel prediction as defined by the study airline.

A. Fuel Loading Categories

A flight fuel plan consists of several components, grouped into **mission fuel** and **safety fuels**:

- **Mission fuel** covers the fuel needed to complete a flight, including **taxi fuel** for ground operations and **trip fuel** for climb, cruise, descent, approach, and landing. These quantities are calculated by the Flight Planning System (FPS) based on route, weather, aircraft performance, and payload.
- **Safety fuels** ensure that flights can manage unforeseen events. These include the **final reserve fuel** (30 minutes of holding at 1,500 ft), **alternate fuel** to divert and land at an alternate airport, and **contingency fuel**, typically 5% of trip fuel or 5 minutes of holding under EASA rules [8, 9]. The study airline applies approved EASA variations, which include reducing contingency to 3% when en-route alternates are available, or using statistical models [8].
- Dispatchers may add **additional fuel** for expected deviations (e.g., route changes, adverse weather, congestion, or technical faults). Finally, pilots may load **discretionary fuel** based on experience. An internal analysis found that 11.6% of flights carried discretionary fuel, though only 4.5% was typically used [10]. Pilot surveys confirm that weather and ATC conditions are the main drivers of this practice [11].

B. Abnormal Phase

We define the abnormal phase (ABNPH) as the point in a flight at which the aircraft begins to consume its alternate fuel. This situation typically arises when the allocated contingency fuel is insufficient to mitigate unforeseen operational circumstances. A flight enters ABNPH when:

$$\text{ABNPH} = \begin{cases} 1 & \text{if } F_{\text{rem}} < F_{\text{safety}}, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

where F_{rem} denotes the remaining fuel upon landing, and $F_{\text{safety}} = F_{\text{reserve}} + F_{\text{alternate}}$ the combined final reserve fuel and the alternate fuel.

III. Related Work

There is a limited amount of studies that directly address the calculation and optimisation of contingency fuel and discretionary fuel. This is partly due to the confidential nature of the operational data required for such analyses. Furthermore, any changes in fuel loading procedures must comply with strict regulatory requirements, and implementing such changes typically requires significant time and resources, further discouraging innovation in this area.

Several studies have applied statistical methods to model trip-fuel deviations and optimise contingency levels based on historical data [12–14]. While these approaches can reduce contingency fuel relative to regulatory minima, results are mixed. Kuts and Kovalenko [13] found that such models did not consistently outperform a Reduced Contingency Fuel policy, highlighting the trade-off between data-driven safety margins and economical benefits. Schneider et al. [15], by contrast, introduced a dynamic smoothing approach that reduced variability in fuel deviations and demonstrated substantial potential cost savings at scale.

Promising advancements in the field stem from Machine Learning approaches, particularly the work of Kang and Hansen in [5] and [6], in determining the fuel deviations. Utilizing data from a major U.S. airline, they employed ensemble methods and quantile regression techniques to achieve more accurate overall fuel burn predictions. In [5], Kang and Hansen demonstrated that leveraging more accurate fuel burn predictions to propose optimised fuel loads could result in annual savings of $61.5 million and 428 million kilograms of CO_2, all while maintaining the same safety performance as current practices, as measured by final reserve fuel usage risks. Subsequently, their work in [6] extended this analysis using quantile regression-based ML algorithms to improve Statistical Contingency Fuel (SCF) models. They estimated that adopting these enhanced models could lead to an annual reduction of approximately $64.8 million in fuel costs and 451 million kilograms of CO_2 emissions for the study airline. The data used in these studies is limited to U.S. domestic routes.

Zhu and Li developed a Spatial Weighted Recurrent Neural Network (SWRNN) with the goal to predict flight times with more accuracy, permitting reductions in fuel loading. The model integrates spatial and temporal dependencies within air traffic networks by utilizing Automatic Dependent Surveillance–Broadcast (ADS-B) data, weather reports (extracted from Meteorological Aerodrome Report (METAR) data), and airline records. Evaluations using data from a large airline operating from Hong Kong demonstrated significant improvements over traditional methods, with prediction errors of flight times reduced by up to 9 minutes. The model performed especially well on flights experiencing extreme delays. The SWRNN model enabled tailored fuel loading strategies with a pro-safety strategy resulted in fuel savings of up to 1.025% across the fleet, while a more pro-efficiency strategy saved 1.967% of fuel, amounting to 184.6 million kg of CO_2 reductions annually [16]. The scope of the paper was geographically limited, focusing only on flights between mainland China and Hong Kong. It addressed only en-route effects, overlooking potential ground delays that also affect fuel planning. Finally, the study did not address the inherent interpretability challenges of neural networks, which may affect the practical adoption of such models in operational settings.

A. Research Gaps and Contribution of this Paper

The present study covers the following gaps in the literature:

- Existing research on contingency fuel estimations has been largely restricted to specific regions or routes. As a result, prior models lack generalizability across more diverse domestic and intercontinental networks. In contrast, this study makes uses of a dataset spanning the airline's regional and intercontinental operations, enabling a broader evaluation of machine learning's potential in this domain.
- Another key barrier is pilot and dispatchers scepticism toward prediction models, as documented in prior studies [3, 6]. Current internal models only indirectly incorporate factors that flight crews consider during contingency loading, such as weather and congestion, leading to additional discretionary fuel loading. Explainable machine learning (XAI) offers a promising but under-explored solution by enhancing the transparency of model outputs. By highlighting the features considered and their relative influence, XAI can improve interpretability and strengthen user confidence in data-driven decision support systems. This study examines what methods can make fuel prediction models more transparent and trustworthy.

IV. Methodology

In the following section, we outline the methodology adopted for this study. We begin by introducing the underlying machine learning algorithm in Subsection IV-A. We then describe the construction of the target variable used for model training and inference, followed by a discussion of the feature engineering process in Subsection IV-C.

A. Model Selection and Development

Due to the complexity of flight operations and the variability in operational parameters, we employ quantile regression methods [17, 18] to estimate conditional quantiles of fuel requirements, $Q_\tau(Y \mid X)$. Unlike mean regression, this approach characterizes the full conditional distribution, which allows for the construction of prediction intervals that support risk-aware flight planning. Quantile regression is also more robust to heteroskedasticity and outliers.

In this work, quantile regression is implemented using Gradient Boosting Machines (GBMs) [19]. These models iteratively combine decision trees into an ensemble, each step reducing the quantile-specific loss. Tree-based learners are particularly attractive here because they capture complex feature interactions while maintaining interpretability through rule-based structures [20–22], an important property in safety-critical tasks such as fuel prediction.

For efficiency and scalability, we adopt LightGBM [23]. Its leaf-wise tree growth, histogram-based split finding, gradient-based one-side sampling (GOSS), and exclusive feature bundling (EFB) reduce training time and memory usage while maintaining a high degree accuracy. Additional strengths include native handling of categorical features, effective treatment of missing values, and direct support for quantile loss functions.

1) Hyperparameter Optimization: LightGBM hyperparameters were tuned using Optuna [24] with its default Tree-structured Parzen Estimator (TPE) search algorithm. Candidate settings were sampled from predefined ranges and evaluated on a validation set using the pinball loss. The tuning was limited to one model quantile ($\tau = 98.8$), and the search was run for 1000 trials to identify configurations that minimized the validation loss resulting in hyperparameters presented in Table I.

TABLE I. LightGBM Hyperparameters

Hyperparameter	Value
n_estimators	604
learning_rate	0.09
num_leaves	165
max_depth	4
min_child_samples	3
subsample	0.89
colsample_bytree	0.49
reg_alpha	2.0
reg_lambda	0.0
objective	Quantile

B. Target Variable

We define the **Required Contingency Fuel** (F_{rcf}), based on the landing fuel reserve margin, the difference between fuel remaining at landing (F_{rem}) and the safety components (F_{safety}):

$$F_{\text{extra}} = F_{\text{rem}} - F_{\text{safety}}, \tag{2}$$

where $F_{\text{extra}} > 0$ indicates over-provisioning and $F_{\text{extra}} < 0$ indicates under-provisioning. The required contingency fuel, F_{rcf}, is obtained by adjusting the variable fuel component for the excess fuel carried and accounting for its carriage cost (CTC_{extra}). Following Ryerson et al.[3], the cost-to-carry (CTC) factors are estimated separately for each aircraft type, using flight-level characteristics such as block fuel, aircraft mass, and flight distance. A detailed description of the estimation procedure is provided in Ryerson et al.[3].

$$F_{\text{rcf}} = \underbrace{(F_{\text{cont}} + F_{\text{disc}})}_{\text{variable fuel components}} - F_{\text{extra}} + CTC_{\text{extra}} \tag{3}$$

Equation (3) is limited in that it adjusts only contingency and discretionary fuel, whereas dispatcher-added uplifts for weather or traffic often serve a similar role. Excluding these means the model has not learned to explicitly account for them, so if dispatcher practices change, F_{rcf} may systematically underpredict the required contingency fuel.

The updated target definition differs from the company's current model, which proxies contingency fuel by the trip–fuel deviation. This measure reflects only en-route effects and often diverges from the true contingency requirement, particularly when atypical taxi times or additional uplifts are involved.

C. Feature Engineering

The features used for model training are grouped into three categories:

- The planning-stage flight-plan data (prepared approximately 90 minutes prior to departure), which provide baseline routing and operational details.
- For each city-pair, the following historical flight data elements are computed: mean and standard deviation of actual airborne time, and the trip time deviation (actual minus planned airborne time). These features are intended to reflect persistent route effects (e.g., ATC procedures, typical routings, and flow-management constraints) that are not fully explained by weather or distance alone. This approach follows the example of Kang & Hansen [6].
- TAF weather data is processed with METAFORA [25]. The forecast best matching the scheduled arrival time of the flight is used.

A complete overview of features used for training and inference is provided in Table II.

TABLE II. MODEL FEATURES

Feature	Unit	Type	Example
Flight Plan Features			
City pair	[-]	Cat.	XXX–XXX
Aircraft type	[-]	Cat.	777
Aircraft registration	[-]	Cat.	XXXXX
Season	[-]	Cat.	W24
Cost index	[-]	Num.	200
Takeoff weight	[kg]	Num.	315,000
Block fuel weight	[kg]	Num.	109,250
Performance factor	[-]	Num.	1.030
Great-circle distance	[NM]	Num.	5232
Avg wind component	[kt]	Num.	21
Arrival month	[-]	Num.	11
Arrival day	[-]	Num.	4
Historical Airborne Features			
Mean of historical actual trip time	[min]	Num.	650
Standard deviation of historical actual trip time	[min]	Num.	12
Mean of trip time deviation	[min]	Num.	-6
Standard deviation of trip time deviation	[min]	Num.	8
Weather Features			
CAVOK	[-]	Cat. (bool)	True
Wind dir. (compass)	[-]	Cat.	NW
Precipitation	[-]	Cat. (bool)	False
Obscuration	[-]	Cat. (bool)	False
Thunderstorms	[-]	Cat. (bool)	False
Freezing	[-]	Cat. (bool)	False
Showers	[-]	Cat. (bool)	True
Snow	[-]	Cat. (bool)	False
Ice	[-]	Cat. (bool)	False
Hail	[-]	Cat. (bool)	False
Clouds (present)	[-]	Cat. (bool)	True
Visibility distance	[m]	Num.	9000
Cloud base height	[m]	Num.	3,048
Cloud cover (amount)	[-]	Num.	6
Wind speed	$[m/s]$	Num.	18
Wind gust	$[m/s]$	Num.	12.86

V. CASE STUDY

To evaluate the proposed machine learning approach, this study makes use of large-scale operational data. Subsection V-A outlines the data sources and collection process, including the steps taken to ensure quality and consistency. We then introduce the baseline Statistical Contingency Fuel model in Subsection V-B, which represents the current airline implementation and serves as the operational benchmark against which the new model is assessed.

A. Data Collection and Processing

This study employs operational data from a major European airline, spanning January 2023–April 2025 and covering 537,057 flights across short-haul European and long-haul international routes. For model development, flights from January 2023–December 2024 (462,459) formed the training/validation set, while those from January–April 2025 (74,598) were reserved as a holdout test set. The temporal splitting mimics real-world application, using the most recent history for prediction while enabling evaluation under potential seasonal or distributional shifts. Meteorological inputs were derived from destination TAFs available during the flight planning stage.

The following data processing techniques were used to improve data quality:

1) Data Filtering Criteria: Flights are excluded if they have (i) missing key inputs (trip-fuel deviation, weighting factor, flight plan, or TAF), (ii) non-commercial flights (e,g. maintenance), (iii) invalid routing (identical origin–destination or mismatched arrival), (iv) unaccounted fuel loads, largely in tail-end ferry flights, or (v) extreme taxi-in/out times indicating inaccurate departure or arrival timestamps. After filtering, 432,113 flights remained for the training dataset, with most removals due to missing TAFs and unaccounted fuel. Future work will explore whether more selective filtering strategies could further improve model reliability.

2) Outlier Removal: Outliers were addressed using a weighted z-score method (threshold 4.0) applied within aircraft type and route groups. A temporal weighting factor highlighting recent operations aimed at capturing gradual shifts in the data (e.g., seasonality, traffic patterns):

$$w_i = 2 \cdot \frac{D_i + 1}{D_{total} + 2}, \tag{4}$$

where D_i is the day index of flight i, and D_{total} is the dataset span. The lower threshold was chosen to retain a larger portion of atypical yet operationally plausible flights, critical for modelling abnormal contingency fuel consumption, while still filtering out implausible extremes.

3) Missing Data Handling: Operational datasets often contain incomplete features, both during training and at inference. A key advantage of the chosen Light Gradient Boosting Machine (LGBM) model [23] is its native support for missing

values. Consequently, no imputation or feature augmentation is applied, ensuring consistency between training, validation, and prediction phases when flight plan data may be unavailable.

B. Baseline Comparison - Statistical Contingency Fuel

The study airline determines contingency fuel using a statistical model developed in 1986, which will serve as baseline comparison to this study. This model estimates required contingency fuel from historical trip-fuel deviations, calculated over two years of operational data for each aircraft type and origin–destination pair, under the assumption that deviations follow a normal distribution. Different coverage levels (e.g., 90% or 99%) are assigned based on destination runway configuration, visibility/ceiling forecasts, and an internal congestion classification.

While operationally well integrated, the system has several limitations. It relies on a narrow set of inputs and incorporates weather and traffic effects only indirectly. First, the assumed distribution does not hold across all aircraft–route pairs, and the method cannot be easily adapted to new routes or aircraft types with limited data. Second, the low outlier threshold risks discarding valuable cases where unusually, but realistic, high trip-fuel deviations occurred, thereby losing information that could improve predictions. Furthermore, the baseline approach draws on only a limited feature set and gives little explicit weight to factors such as weather conditions and traffic congestion—features that pilots and dispatchers routinely consider when making fuel-loading decisions.

VI. Results

The following section presents the results of our analysis. We first report the performance of the proposed model in Subsection VI-A, evaluating its predictive accuracy against established baselines. We then examine the explainability of the model outputs in Subsection VI-B, with a focus on how the results can be interpreted in an operational context by pilots and dispatchers.

A. Model Evaluation

Separate models are trained for different quantiles for $\tau \in \{0.950, 0.952, \ldots, 0.998\}$ (95.0–99.8% in 0.02 percentage-point increments), training on the training set and tuning hyperparameters on the validation set (final values in Section IV). Higher quantiles penalize under-prediction more heavily and thus yield more conservative fuel predictions. We evaluate models using both machine-learning statistics and business KPIs, benchmarking against the current baseline implementation. As a statistical measure, we report the quantile goodness-of-fit developed by Koenker and Machado [26]:

$$R^1_\tau = 1 - \frac{\hat{V}(\tau)}{\tilde{V}(\tau)}, \tag{5}$$

where $\hat{V}(\tau)$ denotes the model's pinball loss and $\tilde{V}(\tau)$ the pinball loss of an constant τ-quantile predictor, always predicting the sample τ-quantile of the training target. An R^1_τ value of 0 indicates no improvement over this baseline predictor, values closer to 1 indicate better fit, and $R^1_\tau < 0$ reflects worse performance.

We compare the quantile regression model with the baseline SCF model. As the latter reports performance for different coverage levels of 90%, 95%, and 99%, we select the quantiles from our model whose *empirical test-set coverage*—the amount of flights where the predicted fuel uplift exceeded the amount required to meet safety reserves—most closely matches SCF levels. Results are shown in Table III.

TABLE III. Quantile regression goodness-of-fit and empirical coverage (model τ matched to SCF targets by coverage).

Quantile / Model	R^1_τ	Coverage
97.8	0.2494	0.9672
SCF90	-0.5329	0.9672
98.8	0.2924	0.9804
SCF95	-0.2307	0.9782
99.8	0.2975	0.9946
SCF99	0.1962	0.9939

The different model quantiles achieve a higher goodness-of-fit at matched coverage, implying lower pinball losses. The baseline's lower goodness-of-fit reflects conservative over-prediction, consistent with its objective of modelling trip–fuel deviation rather than required contingency usage. Segmentation of the results by region indicates that model performance is strongest on shorter, European operations and decreases for the longer intercontinental flights. Overall, these results suggest that quantile-based models can deliver the same safety coverage as SCF with less contingency uplift.

Figure 1 relates the cost-to-carry (CTC) of excess contingency fuel to the percentage of abnormal phases (ABNPH). The y-axis reports CTC, which is calculated from the *excess contingency fuel* (uplifted contingency minus the model's required contingency). Since unused contingency fuel can be carried forward and used in the next leg's block, not all of the excess fuel is considered wasted.

The red dot represents the solution of the baseline model. In Figure 1, for example, the τ=98.6, 98.8, and 99.0 quantile models achieve both lower CTC and fewer abnormal phases than the baseline SCF model. The proposed model can optimise the current operations in two possible ways:

1) **Economical policy:** maintains ABNPH at current levels while reducing CTC through lower contingency uplift. CTC falls from 2,136,211 kg to 1,900,846 kg (235,364 kg, ≈11%), with ABNPH declining by 0.03 percentage points. At an average jet-fuel price of $689.30 per tonne,[1] this corresponds to a saving of $162,236. Using an emissions factor of 3.16 kg CO_2/kg fuel [27], the reduction equates to 743t CO_2, avoiding an additional $69,526 in internal carbon costs [2], for a total estimated saving of **$231,762** over the four-month period.

[1]https://www.iata.org/en/publications/economics/fuel-monitor/ (accessed 31 August 2025).

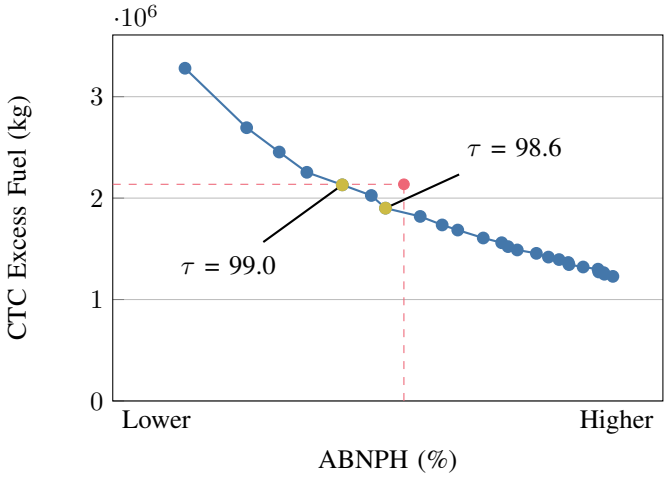

Figure 1. Cost-to-carry of excess fuel as a function of the percentage of flights experiencing abnormal phase events. Exact values on the horizontal axis are omitted due to data sensitivity.

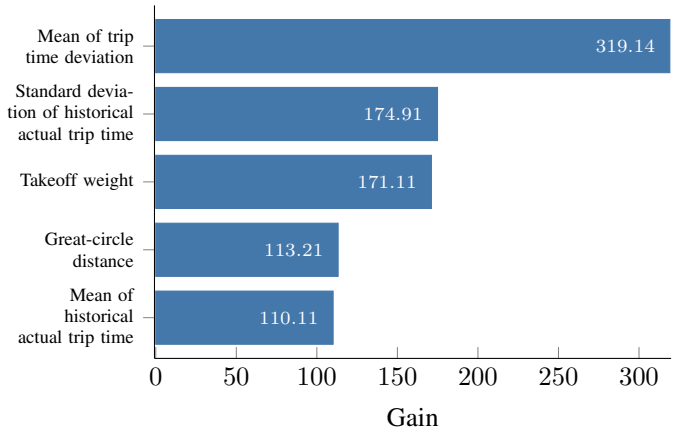

Figure 2. Top features by gain importance.

2) **Safety policy:** maintains CTC at current levels while reducing the number of abnormal phases entered by ≈22.2%.

B. Explainability

Decision trees are often described as interpretable models because their structure can, in principle, support both global and local explanations—global summaries highlight dataset-level patterns, while local analyses trace the decision path behind individual predictions [21, 22]. However, this inherent interpretability only holds when the tree remains small. Ribeiro et al. [28] argues, decision-rule explanations quickly become less interpretable as model complexity increases, for example, when the number of trees or their depth grows. The trained models are substantially more complex than what is typically considered human-interpretable.

We illustrate potential explainability metrics using model agnostic explanations, LightGBM's built-in feature importance and SHapley Additive exPlanations (SHAP) [29]. These provide global and local explanations that can be more easily understood by end users.

1) Global Explainability: Figure 2 shows gain-based feature importance from LightGBM. Three of the top predictors are historical airborne-time metrics, capturing route congestion and operational conditions; the most influential is the mean trip-time deviation. Great-circle distance and take-off weight also rank highly, consistent with longer and heavier flights being more likely to require contingency fuel.

2) Local Explainability: Local explanations are illustrated in the SHAP waterfall plot (Figure 3). This visualization plots how each feature shifts the prediction from the model baseline to the final output: blue bars raise the estimate, red bars lower it. Such plots highlight the specific drivers behind an individual flight's contingency fuel recommendation, providing pilots and dispatchers with transparent, case-by-case insights.

VII. Discussion and Recommendations

While this study demonstrates the potential of machine learning for contingency fuel estimation, several areas merit further work. The models show promise in reducing both excess contingency carriage and abnormal phases by improving prediction accuracy. Building on these findings, this section reflects on the results and their implications for both practice and future research. We begin in Subsection VII-A by comparing the proposed approach against the current Statistical Contingency Fuel baseline. We then consider operational aspects of deploying the model in Subsection VII-B. The discussion continues in Subsection VII-C, where we highlight methodological and practical constraints of this study, before turning to potential avenues for improvement and extension in Subsection VII-D.

A. Comparison with Baseline

The current baseline SCF method already delivers strong performance, and our approach does not bring large contingency fuel savings. The immediate benefit lies less in direct fuel savings but in (1) reducing the need to compensate contingency deficiency with discretionary fuel, and (2) offering pilots more transparent, data-driven reasoning. A well-calibrated model may even increase contingency in specific high-risk cases where needed. With richer predictors and clearer feedback loops, pilot and dispatcher confidence may grow, supporting reductions in discretionary fuel.

B. Operational Usage

The existing SCF model is able to provide predictions at the outset of flight planning. The updated machine learning approach developed here, however, relies on features extracted from the flight plan, which introduces a feedback loop between the flight planning and the model's outputs. Addressing this dependency will be important for successful operational integration.

C. Current Limitations

One current limitation is the use of a single quantile across all flights. The SCF model already applies different coverages

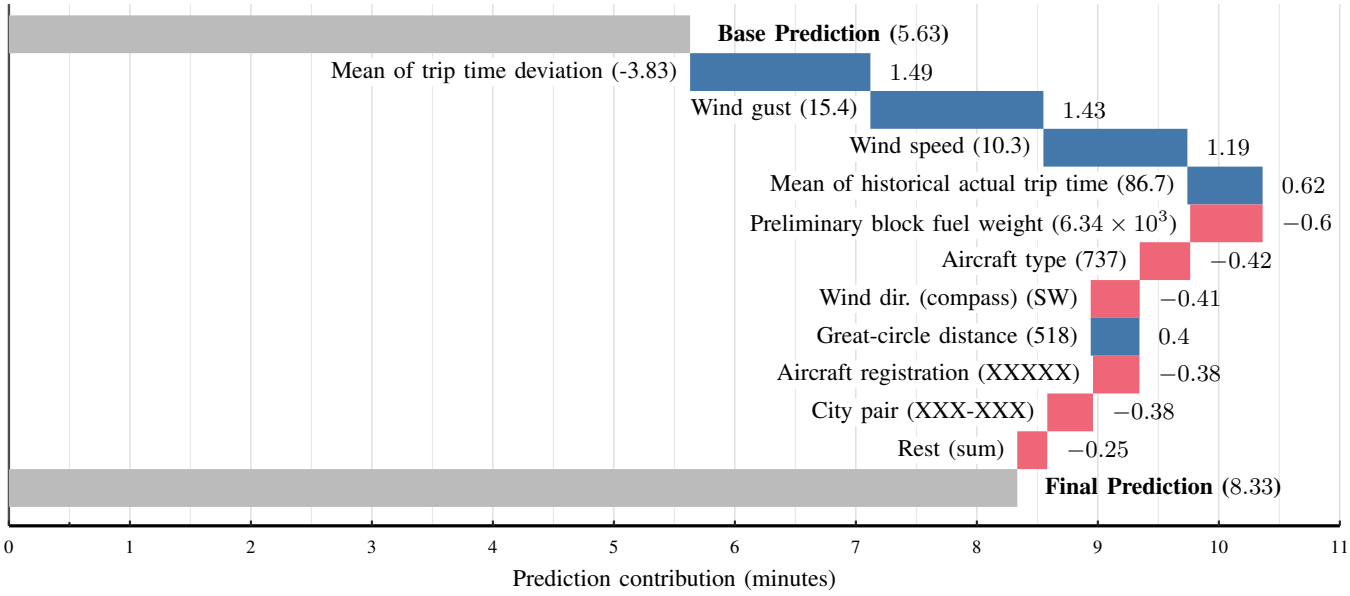

Figure 3. SHAP waterfall. Bars show each feature's contribution from the base value to the final prediction.

by risk profile, and further gains could come from adapting model quantiles to flight risk levels. Lower quantiles should be applied to flights assessed as low risk (good forecast weather and low traffic at the destination), whereas higher quantiles are reserved for high-risk flights (deteriorating weather, heavy arrival flows, or runway constraints). Current risk levels are defined by a simple decision tree using destination weather, congestion, and runway availability; more advanced classifiers, as in Zhang and Mahadevan [30], could assign categories more effectively.

Furthermore, the choice of target variable warrants further investigation. While the revised definition better aligns with operational needs, analyses using alternative abnormal-phase definitions indicate that performance is definition-dependent. The current target performs poorly on abnormal phase definitions based solely on the trip fuel deviation.

D. Future Improvements

A goal for future research is to expand beyond quantile regression to neural networks and other advanced learning architectures. The larger volume of available data could support such models, but challenges remain. First, the data are highly imbalanced—only a small fraction of flights actually require contingency fuel—which complicates training. Second, complex models such as neural networks and stacked ensembles may reduce explainability.

While the operational dataset enabled a model that generalizes across both regional European and intercontinental flights, performance is consistently weaker on long-haul routes than on short-haul operations. This likely reflects greater variability on long-haul sectors (e.g., weather forecast uncertainty) that is not fully captured by the current predictors. Future work should assess whether region–specific models (Europe vs. Intercontinental) or region-aware post-hoc calibration improve performance.

Finally, the model predictors relied only on TAF forecasts, which carry more uncertainty for long-haul flights. Incorporating METAR observations, as suggested by Dalmau and Attia [31], could better capture actual weather–fuel requirement relationships.

VIII. Conclusion

This study has demonstrated the potential of machine learning–based approaches to improve contingency fuel prediction, addressing a notable gap in operational fuel management research. Using a quantile regression LightGBM model, we showed that the cost-to-carry excess fuel can be reduced by approximately 10% while maintaining comparable safety margins. A more conservative configuration reduces fuel savings but decreases abnormal flight phases by around 22.2%, highlighting the trade-offs between efficiency and operational robustness. Compared to current methods used in real-world operations, the quantile regression approach achieved stronger goodness-of-fit at equivalent coverage levels, offering a more accurate and flexible representation of fuel needs without compromising safety.

We also introduce interpretable methods that can be used to communicate model outputs to pilots and dispatchers, with the aim of building trust and reducing discretionary fuel loading. This integration of explainability into operational decision-making represents an important step toward practical adoption.

Future research should address limitations of the current approach, particularly the assumption of a uniform quantile across all flights. Developing adaptive strategies that allow coverage levels to vary with route characteristics, weather risk, or operational constraints may further enhance both safety and efficiency.

ACKNOWLEDGMENT

The authors would like to thank Frans Huisman for his guidance and support during the initial stages of the project. His perspective as a pilot provided valuable insights into how contingency and discretionary fuel management are approached operationally. The authors also acknowledge Frans Vossen for his crucial role in assisting in the preparation and explanation of the operational data.

REFERENCES

[1] A4E, ACI-Europe, ASD, CANSO, and ERA, "Destination 2050 - A Route to Net Zero European Aviaition," tech. rep., Royal Netherlands Aerospace Centre, SEO Amsterdam Economics, 2021.

[2] KLM, "Climate Action Plan," tech. rep., Amsterdam, The Netherlands, 2023.

[3] M. S. Ryerson, M. Hansen, L. Hao, and M. Seelhorst, "Landing on empty: estimating the benefits from reducing fuel uplift in US Civil Aviation," *Environmental Research Letters*, vol. 10, p. 094002, Sept. 2015.

[4] L. Kang, M. Hansen, and M. S. Ryerson, "Evaluating predictability based on gate-in fuel prediction and cost-to-carry estimation," *Journal of Air Transport Management*, vol. 67, pp. 146–152, Mar. 2018.

[5] L. Kang and M. Hansen, "Improving airline fuel efficiency via fuel burn prediction and uncertainty estimation," *Transportation Research Part C: Emerging Technologies*, vol. 97, pp. 128–146, Dec. 2018.

[6] L. Kang and M. Hansen, "Quantile Regression–Based Estimation of Dynamic Statistical Contingency Fuel," *Transportation Science*, vol. 55, pp. 257–273, Jan. 2021.

[7] L. Hao, M. Hansen, and M. S. Ryerson, "Fueling for contingencies: The hidden cost of unpredictability in the air transportation system," *Transportation Research Part D: Transport and Environment*, vol. 44, pp. 199–210, May 2016.

[8] EASA, "Fuel Implementation Manual V0.1," tech. rep., European Union Aviation Safety Agency, 2022.

[9] ICAO, "Annex 6 - Operation Of Aircraft - Part I - International Commercial Air Transport - Aeroplanes," tech. rep., Nov. 2018.

[10] E. M. Tramper, "Improving Fuel Decisions by Airline Pilots in Flight Preparation and Execution," Master's thesis, Delft University of Technology, Delft, The Netherlands, Jan. 2015.

[11] A. C. Trujillo, "Uncertainties that Flight Crews and Dispatchers Must Consider When Calculating the Fuel Needed for a Flight," NASA Technical Memorandum, NASA, Langley Research Center, Hampton, Virginia, May 1996.

[12] Y. Qian and T. Long, "Contingency Fuel Consumption Analysis and Optimization Based on QAR Data," in *Proceedings of 2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCASIT 2019): October 17-19, 2019, Kunming, China*, (Piscataway, NJ), IEEE Press, 2019.

[13] K. A. Kuts and G. V. Kovalenko, "Boeing 777 Fleet Statistical Contingency Fuel Determination on Fixed Routes," *Russian Aeronautics*, vol. 64, pp. 583–590, Oct. 2021.

[14] K. Atchariyachanvanich, W. Kruaklai, N. Chaipatchareekorn, N. Sukteab, and S. Yooyen, "A statistical model for estimating statistical contingency fuel," in *2022 20th International Conference on ICT and Knowledge Engineering (ICT&KE)*, (Bangkok, Thailand), pp. 1–5, IEEE, Nov. 2022.

[15] D. C. Schneider, *An exploratory analysis of commercial airline contingency fuel calculations: with forecasting and optimization*. PhD, The George Washington University, May 2009.

[16] X. Zhu and L. Li, "Flight time prediction for fuel loading decisions with a deep learning approach," *Transportation Research Part C: Emerging Technologies*, vol. 128, p. 103179, July 2021.

[17] R. Koenker and G. Bassett, "Regression Quantiles," *Econometrica*, vol. 46, p. 33, Jan. 1978.

[18] R. Koenker, *Quantile regression*. No. no. 38 in Econometric Society monographs, Cambridge New York: Cambridge University Press, 2005.

[19] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, vol. 29, Oct. 2001.

[20] A. A. Freitas and A. A. Freitas, "Comprehensible Classification Models – a position paper,"

[21] N. Burkart and M. F. Huber, "A Survey on the Explainability of Supervised Machine Learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, Jan. 2021.

[22] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Mar. 2017. arXiv:1702.08608 [stat].

[23] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," 2017.

[24] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (Anchorage AK USA), pp. 2623–2631, ACM, July 2019.

[25] R. Dalmau-Codina and L. Gabagnou, "METAFORA," May 2024.

[26] R. Koenker and J. A. F. Machado, "Goodness of Fit and Related Inference Processes for Quantile Regression," *Journal of the American Statistical Association*, vol. 94, pp. 1296–1310, Dec. 1999.

[27] International Air Transport Association, "IATA Carbon Offset Program - Frequently Asked Questions," Tech. Rep. 10.2, International Air Transport Association (IATA), Apr. 2022.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," Aug. 2016. arXiv:1602.04938 [cs].

[29] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 2017. arXiv:1705.07874 [cs].

[30] X. Zhang and S. Mahadevan, "Ensemble machine learning models for aviation incident risk prediction," *Decision Support Systems*, vol. 116, pp. 48–63, Jan. 2019.

[31] R. Dalmau and J. Attia, "A Collection of Machine Learning Models for Improved Airport Operations Amidst Adverse Weather Conditions," *European Journal of Transport and Infrastructure Research*, vol. 25, pp. 133–159, Feb. 2025.