Second US-Europe Air Transportation Research and Development Symposium (ATRDS2026)

Probabilistic Forecasting of Aircraft Transit Time within a Flight Information Region: A Case Study of Schiphol Airport

Egemen Süülker, 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*—During the transition from the en-route phase to landing, an aircraft's flight time is subject to significant uncertainty. This uncertainty arises primarily from unpredictable weather, varying aircraft performance characteristics, and the human element in executing ATC instructions. Improved estimation of the approach phase duration could yield significant benefits for airline fuel planning and flight scheduling, yet current practice still largely relies on fixed deterministic buffers. Existing work on arrival delay prediction focuses on deterministic models at smaller forecast horizons during the airborne phase. This paper develops and validates an explainable probabilistic forecasting model for flight duration within the Amsterdam Schiphol (AMS) Flight Information Region (FIR), with a forecast moment in the pre-departure phase. The primary objective is to forecast the duration within the AMS FIR using information available at planning, while providing interpretable contributors of delay that can be clearly communicated to flight dispatchers and pilots. Results show that our model achieves a reduced MAE of 33% and a reduced RMSE of 26% relative to the current operational baseline, while capturing about one third of the variance in FIR duration ($R^2 = 0.33$). Additionally, quantile-based transit time forecasts can provide airlines with a more risk-aware basis for fuel and schedule planning than fixed deterministic buffers. However, the relatively low R^2 shows that a substantial share of the variation in FIR duration remains unexplained, largely associated with tactical ATC interventions that occur under otherwise acceptable weather and capacity conditions.**

***Keywords*—Flight Information Region; Airline Planning; Explainable Machine Learning; Quantile Regression; Tree Based Learning Algorithms**

I. Introduction

Global airspace is divided into Flight Information Regions (FIRs), three-dimensional areas of defined dimensions within which flight information and alerting services are provided by the designated authority [1]. Within a given FIR, airspace is further organized into sectors and classifications, including the terminal areas surrounding major airports. The Terminal Maneuvering Area (TMA) is the controlled airspace where air traffic transitions between the en-route phase and the final approach, or conversely, from departure to en-route. TMAs are crucial for managing high-density traffic around major hubs, ensuring both safety and operational efficiency [2]. In the approach phase of flight, aircraft frequently encounter capacity constraints that require sequencing, vectoring, or holdings. These measures often result in airborne delays or, more specifically, additional Arrival Sequencing and Metering Area (ASMA)[1] times, as defined by EUROCONTROL [3].

Predicting delays within the FIRs, during the planning phase of a flight, offers significant advantages for fuel optimization, punctuality, and passenger satisfaction. Accurate forecasts allow for more informed contingency fuel loading, thereby reducing the 'fuel-to-carry-fuel' penalty [4]. In terms of punctuality, these predictions can help to mitigate the effects of reactionary delays. By identifying arrival delays hours in advance, airlines can proactively adjust ground handling or aircraft rotations to prevent the propagation of delay into subsequent departures. There is also a clear interest in shifting this predicted additional time to pre-departure ATFM (Air Traffic Flow Management) ground delay which reduces excess fuel burn [5]. Consequently, delay predictions within the arrival airspace involve diverse stakeholders from Air Traffic Control (ATC) to airlines and passengers.

The current state of the art often overlooks delays in the approach phase. Most pre-tactical delay prediction studies aggregate arrival delay rather than focusing on the time within the arrival airspace [6]. Some works achieve impressive error metrics, typically in the low single-digit minute range [7], [8]. However, their forecast moments lie within the airborne phase (close to FIR entry), and do not consider data constraints during earlier planning. Additionally, studies forecasting holding times use Gradient-Boosted Decision Trees (GBDTs) and Long Short-Term Memory (LSTM) networks [7], [9], which provide a point estimate. From an airline perspective, a indication of prediction confidence, i.e. probabilistic forecasting, is also necessary for trustworthiness. While a common strategy for this issue is to combine a regression model with a classifier [10], [11], in this study we make use of a more recent strategy used in aviation: quantile regression to obtain probability

[1]ASMA is defined as a cylindrical airspace with a 40-nautical-mile radius surrounding an airport. The "additional ASMA time" metric estimates the average delay aircraft experience within this zone.

distributions [6], [12].

This study utilizes real operational dataset from a major European airline, integrating historical operational data with specific flight plan information. The primary objective is to extend upon the current planning of flight durations within the Amsterdam Schiphol (AMS) Flight Information Region (FIR), referred as AMS FIR in the remainder of this paper. Beyond achieving a sufficient forecast accuracy, the model should also be explainable so that the underlying logic can be clearly communicated to promote trust among flight dispatchers and pilots.

This paper is divided as follows: Section II introduces the problem providing more information on the FIR area, and respective delays, of Schiphol Airport. Section III introduces the state-of-the-art of TMA and FIR delay estimation, highlighting the contribution of this paper. The forecast method employed is then defined in Section IV. A detailed breakdown of the data sources and the comparative baseline models is described in Section V. Section VI presents the forecasting results with the previous data. The explainability of the model is explored in Section VII. Section VIII discusses the contributions, limitation of this work as well as future work. Finally, Section IX concludes this paper.

II. Problem Definition

Aircraft flight time inside the arrival airspace is dynamically uncertain: strongly driven by air traffic density, runway configuration, weather conditions, and arrival sequencing. To manage these flows, the controller utilizes Arrival Management (AMAN) systems, which generate an Estimated Approach Time (EAT) based on the aircraft's assigned Standard Instrument Arrival (STAR) and holding fixes [13]. These EATs serve as target timestamps for crossing specific points, such as approach fixes or the runway threshold. Nevertheless, they remain advisory and are frequently adjusted due to evolving traffic and operational constraints. As these processes are managed dynamically, deterministic predictions are insufficient. Instead, probabilistic models are more suitable for estimating not only the expected duration but also the confidence of it. This approach provides airlines with a better view of uncertainty, allowing more risk-aware decision-making in their planning.

The target dataset for this research consists of arrival flights operated by a major international airline. The primary goal is to provide insight into arrival transit times during the pre-departure phase, preferably 90 minutes before the aircraft takes off (when the flight plans are typically filed). A major challenge in the approach phase is the unpredictability of ATC interventions, such as vectoring or speed adjustments, which are significant sources of airborne delay.

As illustrated in Figure 1, traffic density heat maps reveal general insights into arrival patterns within the AMS FIR and the TMA. A key observation is that classical teardrop holding patterns are rare. Instead, flights typically follow Standard Instrument Arrivals (STARs) from FIR entry points (such as *TOPPA*, *MOLIX*, or *EEL*) toward Initial Approach Fixes (IAFs) like *SUGOL*, *ARTIP*, or *RIVER*. However, once past the IAF, aircraft are frequently subject to ATC vectoring, resulting in irregular trajectories and sequencing behaviors.

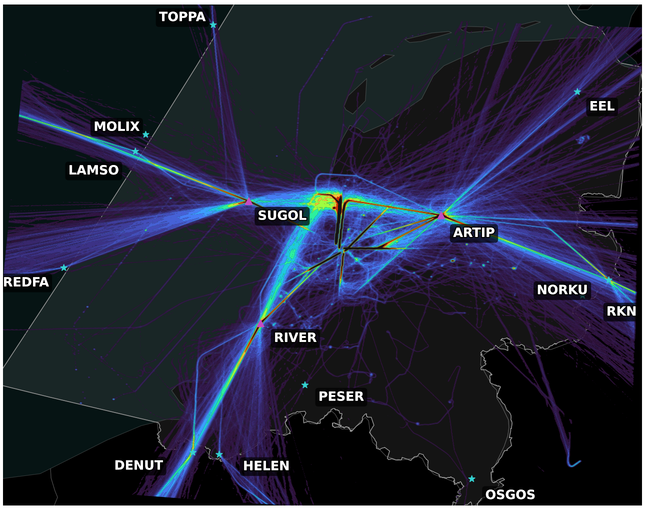


Figure 1. Heat map showing arrival traffic density over the AMS FIR. Blue indicates low-density traffic; Red indicates high-density traffic.

While the majority of previous research has focused on predicting flight times within the TMA, this study instead targets the transit time from FIR entry to touchdown. This choice is motivated by several practical considerations. First, FIR entry points are operationally well-defined and consistently recorded in flight planning data, providing a reliable and standardized reference for model development. In contrast, TMA boundaries vary across airports and regions, making it difficult to define a consistent spatial reference. By using the FIR, the resulting model becomes more readily transferable to other airports and airspaces, enhancing its generalizability. Furthermore, since operational flight planning already references FIR entry points, aligning the prediction horizon with these waypoints ensures closer integration with existing airline workflows.

III. Related Work

In this review, we first focus on studies related to FIRs and TMA areas, which are discussed in Section III-A. Next, Section III-B reviews research on the key factors that significantly influence delays. Finally, Section III-C identifies the research gaps and outlines the objective of this paper.

A. FIR/TMA Delay Forecast

Zhang et al. [7] derived two novel features to forecast the flight time within the TMA. They used multiple features, such as (1) the total number of arrival aircraft already present within the TMA when a flight enters the TMA, (2) the distribution of arrival traffic. They were able to achieve an MAE (Mean Absolute Error) between 55 to 100 seconds for different regression models.

Lui et al. [11] proposed a two-stage gradient boosting framework. In the first stage, a binary classifier estimates the probability that a flight will enter a holding pattern. In the second stage, this probability is combined with the trajectory and weather features to predict the remaining time to landing from the forecast moment of TMA entry. A significant contribution is the introduction of comprehensive feature engineering, i.e. route-specific rainfall intensity metrics and Bayesian weather-induced traffic features, which were demonstrated to outperform the traditional weather features. The best performing configuration, two-stage XGBoost (eXtreme Gradient Boosting), achieved a 6% reduction in Mean Absolute Percentage Error (MAPE) under extreme weather scenarios compared to a single XGBoost regressor.

Dalmau et al. [6] aimed to predict the arrival delays pre-tactically using only flight schedules and historical data. The model makes use of input features available during the pre-tactical phase, such as the airline, aircraft type, or expected number of passengers, to provide predictions of the delay distribution several days before operations. The quantile regression model using CatBoost achieved better error metrics across all quantiles when compared to the current operational and statistical baselines from an airline. Specifically, it achieved a MMQPE (Mean Multi-Quantile Pinball Error) of 30.3 minutes.

Vella et al. [9] developed supervised learning models to predict holding durations 15 to 60 minutes prior to TMA entry. Utilizing ADS-B surveillance data from the OpenSky Network, the study employed a LightGBM regression model and a Long Short-Term Memory (LSTM) network to forecast airborne holding times. The results demonstrated that LSTM achieved better performance with a minimum Root Mean Squared Error (RMSE) of 2.25 minutes and a MAE of 1.5 minutes. The most important features were found to be the current queue length (traffic pressure) and wind speed.

B. Trajectory Delay Prediction

Dalmau et al. [14] developed a multi-stage decision support system combining four LightGBM-based models to improve airport operations under adverse weather across the 50 busiest European airports. The first two models predict arrival and departure peak service rates, defined as the 99th percentile of observed hourly throughput, grouped on weather features, runway configuration, and an airport identifier. Compared against Airport Corner tables[2], the machine-learning models consistently achieve lower mean pinball error on the test set, with particularly large gains at capacity-constrained hubs such as Amsterdam Schiphol Airport. A third model estimates the probability of ATFM regulation due to weather using weather and calendar (hour, weekday, month). For airports with at least 50 regulated hours in the test set, this model obtains precisions between 0.33 and 0.88, and recalls between 0.19 and 0.63. The fourth model predicts the regulation entry rate (arrivals per hour) when a weather-related regulation is in force. On these positive cases, the MAE typically ranges between about 1.5 and 8.6 movements per hour, depending on the airport.

Wang et al. [8] proposed a hybrid machine learning model for short-term 4D trajectory prediction in the TMA. First, ADS-B arrival trajectories are preprocessed and then clustered into a small number of typical arrival patterns. DBSCAN is used to find clusters and to remove noisy trajectories such as holdings and strong vectoring. Second, a Multi-Cells Neural Network (MCNN) is trained, where each cluster has its own "cell" that predicts the runway ETA. A decision tree assigns each new flight to one of the clusters based on its entry position, and the corresponding network produces the ETA prediction. Multiple Linear Regression, with and without clustering, is used as a baseline. The proposed methodology achieved an average MAE of 135.62 seconds and RMSE of 210.21 seconds over five folds with an R^2 (Coefficient of Determination) of 0.8.

[2]An 'Airport Corner table' refers to a EUROCONTROL reference table that lists predefined capacity values for an airport under different runway configurations and operating conditions [15].

C. Research Gaps and Contribution of this Paper

From the related work on forecasting arrival delays, we can identify several research gaps:

- Existing work on delay prediction either focuses on pre-tactical models that forecast aggregate airborne delay [6] (e.g the arrival delay) or on tactical models that predict time in arrival airspaces (FIR/TMA) during the airborne phase [7]–[9], [11]. As a result, there is currently a limited number of studies that forecast time within arrival airspaces (FIR/TMA) before departure while considering practical airline planning constraints regarding data availability across forecast horizons.
- Most studies rely on publicly available data, such as ADS-B and published weather reports, even though airlines have access to richer internal day-of-operations information. Although airline-specific features have been used for departure and en route delay forecasting [16] [17], they have not been applied to arrival airspace duration. Thus, probabilistic FIR duration forecasting using airline-internal planning data remains unexplored.

The objective of this paper is thus formulated: to produce a probabilistic forecast of the transit time in the FIR, generated 90 minutes prior to aircraft departure.

IV. Methodology

The following section outlines the methodology proposed to estimate flight duration within the FIR. Section IV-A defines the target variable to be forecasted, namely the transit duration within the FIR. Section IV-B describes the development of the predictive model used to estimate this quantity. Finally, Section IV-C presents the feature engineering process and details the input variables employed by the model.

A. Target Variable

The target variable, T_{FIR}, is defined as the duration between the FIR entry timestamp and the touchdown time. This scope ensures operational relevance, as FIR entry points are planned during the pre-departure phase, whereas Initial Approach Fixes (IAFs) are often assigned tactically by ATC. Furthermore, choosing the entire FIR-to-touchdown window allows the model to capture the majority of airborne delays, as most sequencing manoeuvres and tactical deviations occur after the aircraft has passed the IAF.

The empirical distribution of T_{FIR} and its deviation from the planned duration, P_{FIR}, is illustrated in Figure 2. The distribution reveals that over half of the flights arrive in less time than the planned durations, highlighting a systemic tendency toward over-provisioning fuel and buffer time for the arrival phase. Conversely, the 95th percentile delay ($\approx$ 7 min) underscores the persistent risk of *unexpected* disruptions. By generating quantile predictions, the proposed model incorporates real-time weather and congestion features to explicitly quantify this tail risk, rather than relying on static buffers used during flight planning.

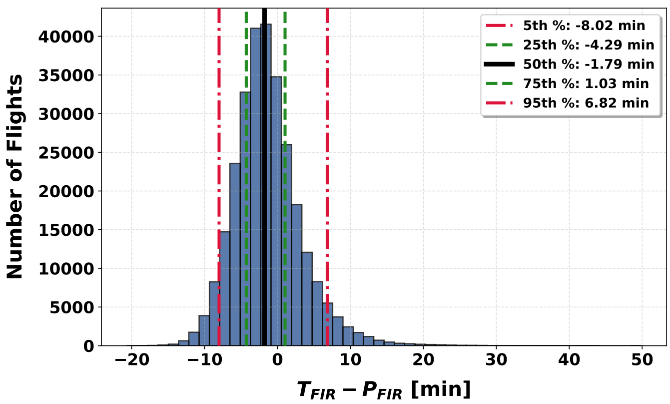


Figure 2. Distribution of the difference between planned (P_{FIR}) and actual (T_{FIR}) FIR durations.

B. Model Development and Selection

Due to the complex and stochastic nature of operations within the arrival airspace, quantile regression models [18], [19] are employed to estimate the conditional quantiles, $Q_\tau(Y \mid X)$, of the flight duration in the FIR. Quantile regression offers a robust alternative to traditional mean-based models, as it is inherently more resilient to heteroscedasticity and the outliers frequently encountered in historical flight datasets. Rather than providing a single point estimate, this approach captures the broader distributional characteristics of T_{FIR}. In particular, it provides critical insights into the upper tails of the distribution, representing potential delays, significantly enhancing strategic fuel planning and decision-making during the flight planning phase.

Quantile regression is implemented with Gradient Boosting Machines (GBMs) [20], which combine decision trees into an ensemble by iteratively fitting to residuals, aiming to reduce the quantile loss. Tree-based learners are able to capture complex feature interactions while preserving interpretability through rule-based representations [21]–[23], an important property for safety-critical tasks such as fuel prediction.

Hyperparameter tuning was done using the Optuna framework [24], utilizing its default Tree-structured Parzen Estimator (TPE) sampler to find the optimal hyperparameters. This tuning process was done using the validation dataset. The optimal configuration, detailed in Table I, is specific to the 95th quantile ($\tau = 0.95$) model.

C. Feature Engineering

Recursive Feature Elimination (RFE) was used to identify the most influential predictors. The final set of selected features is discussed in detail below and can be broadly categorized into three primary groups:

*a) **Flight Plan Features**:* The *Planned Cost Index*, *Planned Runway*, and *Aircraft Type* are extracted from the flight plan. To reduce dimensionality, the *FIR Entry Group* feature aggregates individual entry points into clusters sharing common Standard Arrival Routes (STARs). As previously illustrated in Figure 1, entry points are mapped to their respective IAFs: *RKN*, *NORKU*, and *EEL* are grouped into *ARTIP*; *MOLIX*, *LAMSO*, *REDFA*, and *TOPPA* into *SUGOL*; and *DENUT* and *HELEN* into *RIVER*. The *Great-Circle Distance* is calculated using the Haversine formula [25] relative to the origin airport. Additionally, the *Arrival Hour* and *Arrival Month* are included as categorical variables derived from the planned arrival time. While sinusoidal transformations were evaluated to account for the cyclical nature of these features, they yielded no significant performance gain over a simpler categorical representation. Consequently, the latter was selected to maintain lower model complexity.

*b) **Weather Features**:* Observed weather statistics—including *Visibility, Ceiling, Wind Gusts*, and calculated *Tail/Headwind* and *Crosswind* components—are derived from METAR reports. During training, observed weather is used to ensure that the model captures the relationship between environmental conditions and duration; at inference, these are replaced by the most recent TAF available at dispatch. METAR and TAF messages are parsed using the METAFORA library [26], selecting the validity period corresponding to the planned arrival. Wind components are decomposed relative to the runway heading following standard trigonometric conventions [27].

*c) **Traffic and Congestion Features**:* Airfield congestion is captured via *Arrival Count* (N_{arr}), *Departure Count* (N_{dep}), and *Mean Arrival Delay* ($\bar{D}_{\text{arr}}$). The counting windows for N_{arr} and N_{dep} were empirically optimized using Spearman correlation analysis on the training dataset. The correlation analysis yielded forward looking window of size of $W_{\text{arr}} =$ 100 minutes and backward-looking window $W_{\text{dep}} = 20$ minutes. Specifically, a forward-looking window was selected for arrivals to represent future demand, while a backward-looking window was chosen for departures to represent recent throughput. The planned arrival and departure is used to

TABLE I. LightGBM Hyperparameters ($\tau = 0.95$)

Hyperparameter	Value
n_estimators	273
learning_rate	0.03
num_leaves	413
max_depth	4
min_child_samples	18
subsample	0.47
colsample_bytree	0.98
reg_alpha	1.76
reg_lambda	0.0
objective	Quantile

determine the arrival and departure counts.

To capture latent delay trends, the Mean Arrival Delay ($\bar{D}_{\text{arr}}$) is calculated as the average delay of the N_{past} preceding landings relative to the forecast moment (t_{entry} or t_{forecast} at inference):

$$\bar{D}_{\text{arr}}(t_{\text{entry}}) = \frac{1}{N_{\text{past}}} \sum_{i=1}^{N_{\text{past}}} \delta_i, \tag{1}$$

where δ_i represents the arrival delay of the i-th preceding flight. To calibrate the mean delay feature, an empirical search was conducted on the training dataset, identifying $N_{\text{past}} = 55$ flights as the most effective look-back depth. It should be noted that the practical value of this feature may be restricted, particularly in the context of long-haul operations. For such flights, the airport conditions at the time of dispatch may not accurately reflect the operational environment upon arrival several hours later. Conversely, for short-haul or regional flights, the conditions at departure often serve as a reliable proxy for the arrival state, capturing tactical trends that are highly relevant to crew and dispatch decision-making. An overview of the final features used in the model is presented in Table Table II.

Figure 3 illustrates the correlations between the numerical features selected for model training. Although the *Cost Index* shows a moderate correlation with *Great-Circle Distance*, it is retained because it reflects the strategic intent of the airline, whereas distance remains a strictly physical variable. This approach aligns with findings by Achenbach et al. [28], who demonstrated that while cost index is often a function of distance, it captures unique tactical information regarding fuel costs and delay management. Similarly, the moderate correlation between *Visibility* and *Ceiling* is expected due to their shared meteorological nature, just as the high correlation of *Wind Gust* with other wind metrics.

TABLE II. Model Features

Feature	**Unit**	**Type**	**Example**
Flight Plan Features			
Aircraft type	[-]	Cat.	777
Cost index	[-]	Num.	200
Great-circle distance	[NM]	Num.	697
Arrival hour	[-]	Cat.	X
Arrival month	[-]	Cat.	X
Planned runway	[-]	Cat.	18R
FIR entry group	[-]	Cat.	ARTIP
Traffic & Congestion Features			
Arrival count	[-]	Num.	75
Departure count	[-]	Num.	14
Mean arrival delay	[min]	Num.	-5
Weather Features			
Visibility	[m]	Num.	9000
Ceiling	[m]	Num.	3,048
Tail/Headwind	[kt]	Num.	5.8
Crosswind	[kt]	Num.	6.5
Wind gust	[kt]	Num.	25

V. Case Study

The performance of the proposed machine learning framework is evaluated using a comprehensive, large-scale operational dataset. Subsection V-A outlines the data sources and collection process, including the specific preprocessing steps taken to ensure data quality and consistency. To establish a comparative framework, the model is benchmarked against both the current operational baseline and a more rigorous statistical method; these are detailed in Subsection V-B and Subsection V-C, respectively.

A. Data Collection and Processing

The model was developed using data from internal operational flight records, airport schedule data, and meteorological reports. Additionally, open source ADS-B data from the OpenSky Network [29], is used for filling missed actual touchdown timestamps. The retrieved data is temporarily split as follows: the training set covers flights from January 2023 through September 2024 (232,517 flights). The validation set consists of flights from October 2024 to the end of that year (41,354 flights). The test set includes flights operating between January and October 2025, totalling 45,232 flights. The data processing methods applied are outlined below:

1) Missing Data Handling: For flights with missing FIR entry or touchdown timestamps, values are derived from the ADS-B database. These records are merged using the ATC callsign and the nearest timestamp within a predefined tolerance of the planned arrival time. To ensure data quality, merging is performed selectively to prevent the inclusion of erroneous durations caused by the reuse of callsigns for subsequent flights. For flights with missing planned FIR entry point, values are imputed using the most frequent entry point associated with the specific departure airport.

2) Outlier Removal: Initially, the z-score of the target variable was explored to identify and remove outliers. This explored removing flights that had unusually high T_{FIR} when the weather was within VFR (Visual Flight Rules) limits [30] and the traffic counts was below the season averages. However, this approach presented a significant risk of discarding valid rare events that are crucial for learning the tail behaviour of the delay distribution. Since quantile regression is inherently robust to outliers (unlike standard regression which is sensitive to extreme errors) [18], this strategy was not implemented at the end. Instead, data cleaning was limited to removing only physically implausible flights.

B. Operational Baseline

In the target variable introduction, Figure 2 presented the distribution of actual durations relative to the planned durations. It was observed that, for most instances, the planning was conservative by approximately 2 minutes. The planned duration was derived from the scheduled time between the destination and the last waypoint that includes one of the FIR entry points. The planned duration is sometimes manually adjusted rather than taken directly from the internal planning system. For each FIR Entry–Runway pair, an upper quantile of the observed durations is selected as a realistic, slightly conservative reference. When this quantile-based value is

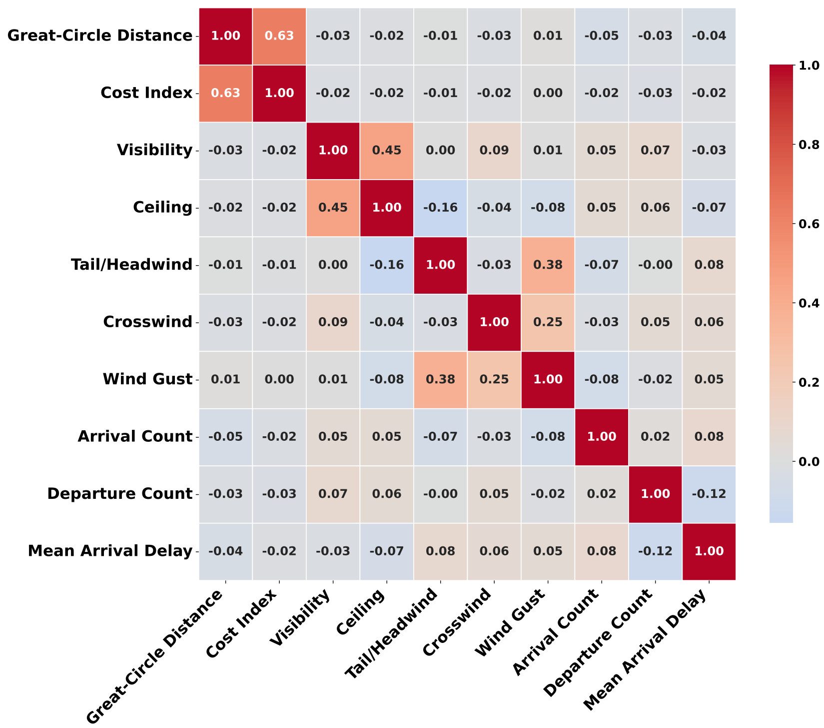


Figure 3. Feature correlation for the numerical features used in the models

substantially lower than the duration provided by the planning system, the planned duration in the schedule is overridden accordingly.

C. Statistical Baseline

To establish a more robust benchmark, a stratified historical baseline was developed. This approach estimates the conditional quantiles ($\tau \in \{0.05, 0.25, 0.50, 0.75, 0.95\}$) of T_{FIR} by grouping historical observations into distinct operational contexts. The stratification is based on three primary dimensions:

- **Temporal:** The day is split into three periods: *Morning* (03:00–12:00), *Afternoon* (12:00–17:00), and *Evening* (17:00–22:00). Seasonal variations are captured by splitting the year into *Summer* and *Winter* based on the daylight saving time schedule.
- **Operational:** The *FIR Entry Group* and *Planned Runway* align with the LightGBM features. *Flight Product* is included to account for the flight's origin distance (Short, Medium, or Long Haul).
- **Meteorological:** Weather conditions are discretized into *Low Visibility* ($< 1500\,\text{m}$), *High Wind* ($> 25\,\text{kts}$), or *Normal* conditions.

To maintain statistical significance, a minimum threshold of 300 observations per cohort was required. If a specific combination of features resulted in an under-represented group, a hierarchical fallback mechanism was triggered to ensure a valid prediction. The model first attempts a *Specific* level match using all available features (FIR Entry, Season, Hour, Runway, Product, and Weather). If insufficient data exists, it filters on an *Operational* level (excluding Weather and Product), otherwise falls-back to a *General* level (FIR Entry and Hour only).

VI. Results

In the subsequent section the resulting model performance is evaluated. This chapter first presents the methods used for assessment of the models Subsection VI-A. Subsequently, Subsection VI-B benchmarks the model against the established baselines.

A. Model Evaluation

Five separate LightGBM models are created for $\tau \in \{0.05, 0.25, 0.50, 0.75, 0.95\}$, with the model performance being evaluated on the test set. The model's performance is evaluated using both point and probabilistic prediction assessments. For point predictions ($\tau = 0.5$), the standard performance metrics for regression models are used, namely the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). For probabilistic predictions, we measure the Mean Pinball Error (MPE). To further assess the performance of the prediction intervals, two additional metrics from the literature [31] are employed:

- Prediction Interval Coverage Probability (PICP): The percentage of actual observations that fall within the interval defined by two quantile models, indicating how well the nominal coverage of the interval is achieved.
- Mean Prediction Interval Width (MPIW): The average absolute difference between the predictions of two quantile models, capturing the width of the prediction intervals.

A high-performing probabilistic model should achieve high coverage (high PICP) while maintaining narrow intervals (low MPIW).

B. Model Comparison

The final models are compared with both the operational baseline and the statistical model. As the Operational Baseline outputs a single deterministic prediction, it is assessed using only the point metrics, whereas the Statistical and LightGBM models are evaluated on both point and probabilistic performance. The results from the evaluation are presented in Table III.

The performance increase when moving from the Operational Baseline to the Statistical Model is much higher than the increase from the Statistical Model to the Proposed Model. Even though the gain in point performance is relatively marginal, LightGBM's superiority lies in its quantile performance. Since the primary objective was probabilistic forecasting, the MPE serves as the main assessment criterion, where the Proposed Model consistently achieved the lowest values across all quantiles. This improvement is most notable at the tails ($\tau = 0.05$, $\tau = 0.95$). The negative R^2 of Operational Baseline reflects the inherent difficulty of using static planning to capture dynamic flight variability. The MPE distribution for the Operational Baseline confirms that static planning is heavily skewed toward the upper tail of the distribution. Effectively, the planned duration serves as a high-percentile buffer rather than a central estimate, resulting in systematic over-provisioning.

C. Dynamic Model Performance

The dynamic performance will be assessed through different forecast horizons. The previous performance results were given for the forecast moment 90 minutes before departure. After this, the performance of the model will be assessed at different times of the day.

TABLE III. Model Performance Comparison

Metric	**Operat. Baseline**	**Statist. Baseline**	**Proposed Model**
Point Metrics ($\tau = 0.50$)			
MAE [min]	3.38	2.84	2.25
RMSE [min]	4.20	3.75	3.10
R^2 [-]	-0.35	0.011	0.33
Interval Metrics ($\tau = 0.05$–0.95)			
PICP [%]	-	86.0	91.0
MPIW [min]	-	12.26	9.51
Interval Metrics ($\tau = 0.25$–0.75)			
PICP [%]	-	47.0	50.53
MPIW [min]	-	4.29	3.68
Mean Pinball Loss (MPE)			
MPE ($\tau = 0.05$)	2.71	0.357	0.267
MPE ($\tau = 0.25$)	2.26	1.110	0.864
MPE ($\tau = 0.75$)	1.13	1.207	0.986
MPE ($\tau = 0.95$)	0.67	0.487	0.384

1) Forecast Horizon: The model uses the latest TAF available at dispatch for weather features. All other flight plan features are fixed at 90 minutes before departure, leaving *Mean Arrival Delay* as the only feature that updates dynamically at the actual departure time. Feature uncertainty analysis showed that this 90-minute gap did not significantly degrade the quality of the *Mean Arrival Delay* feature compared to the forecast horizon for different haul types. Therefore, only the effect of the forecast horizon is assessed based on the flight distance. Figure 4 illustrates the evolution of point-prediction performance across varying origin distances to AMS. A general upward trend in both MAE and RMSE is observed as the great-circle distance increases. This pattern reflects the increased uncertainty inherent in long-haul operations. Furthermore, the localized spikes in error at the extreme ranges (e.g., $\geq 6000\,\text{NM}$) are impacted by data sparsity.

2) Hourly Performance: The performance of the model is likely to vary throughout the operational day due to fluctuating traffic levels and the transition between night and day regimes. To assess the robustness of the model under these varying conditions, the prediction errors are analysed across different hours in a day. As shown in Figure 5, the error is lowest during the deep night hours (03:00–04:00 UTC), where traffic complexity is minimal. A sharp increase in RMSE is observed at 05:00 UTC, coinciding with the lifting of the night curfew and the onset of the morning arrival peak at AMS. The second peak occurs during the evening where the traffic uncertainty is highest.

Crucially, Figure 5 demonstrates the model's ability to adapt to this dynamic uncertainty. The MPIW closely mirrors the error trends, expanding from approximately 7.5 minutes at night to nearly 10 minutes during peak hours. This adaptive widening allows the model to maintain a stable PICP hovering near the theoretical target of 90% (dashed green line) throughout the entire day. This behaviour confirms that the LightGBM model successfully learns to quantify the increased risk associated with high-density daytime operations

VII. Model Explainability

The explainability of the multi-quantile LightGBM model is addressed using the SHapley Additive exPlanations (SHAP) framework. The Shapley value for a feature represents its marginal contribution to a prediction. Since computing exact Shapley values is computationally expensive, this study uses TreeExplainer [32], a specialized method for tree-based models such as GBDTs that enables efficient estimation of Shapley values.

In Figure 6, the vertical axis lists the feature names, ordered from top to bottom by their mean absolute Shapley value (global importance). Each dot along the horizontal axis represents the Shapley value of a given feature for a single observation. The color of the dot indicates the magnitude of that feature for that observation: red for high values and blue for low values. For categorical features, such as runway or aircraft type, the color encoding has no direct interpretation. For clarity, the plots are shown only for the lower-tail model ($\tau = 0.05$), the median model ($\tau = 0.50$), and the upper-tail model ($\tau = 0.95$). Some features, such as *Cost Index* or

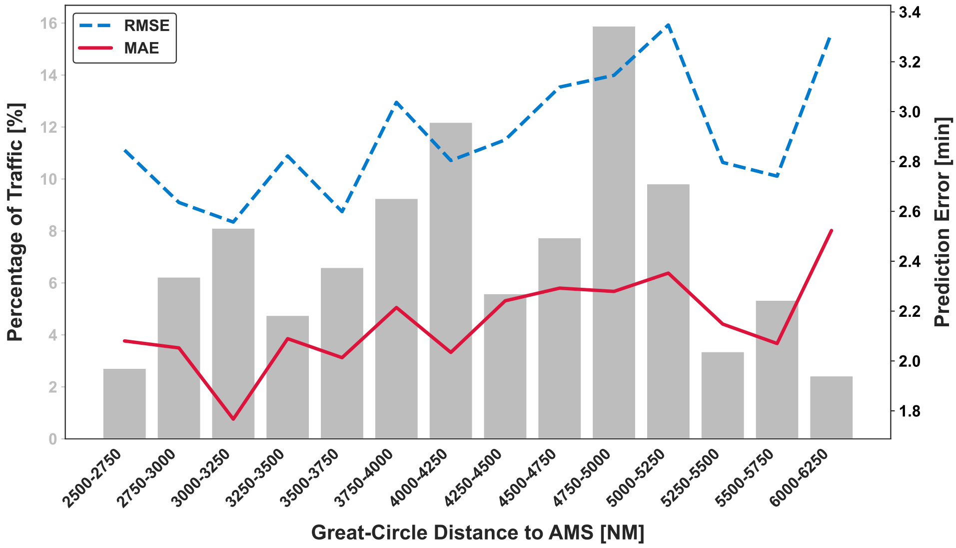
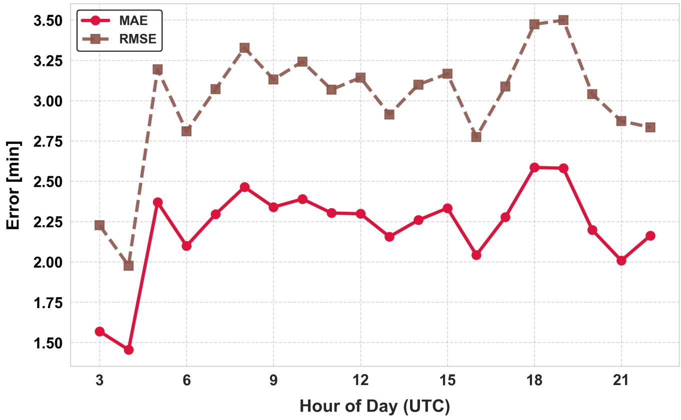
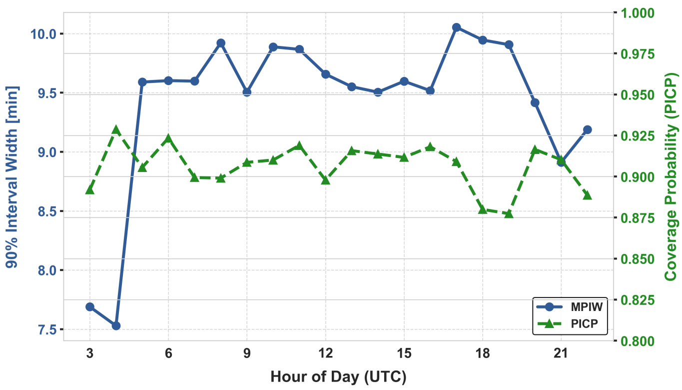


Figure 4. Point performance for different flight distances



Point Prediction Performance



Probabilistic Performance

Figure 5. Hourly evolution of model performance metrics on the test dataset

Ceiling, are treated as categorical by TreeExplainer despite being numerical; this occurs because these variables have low cardinality (a small number of unique values).

Comparing the 5th, 50th, and 95th quantile models shows that, while core features such as *FIR Entry Group* and *Planned Runway* are important across all quantiles, the influence of weather and traffic features increases toward the upper tail. For example, *Tail/Headwind* is influential in every model, but its Shapley values grow from approximately $+1$ minute in the 5th quantile model to about $+10$ minutes in the 95th quantile model. Similarly, weather features such as *Visibility* and *Wind Gust* move up in the importance ranking for higher quantiles, reflecting their stronger role in driving extreme delays.

Overall, the SHAP summary plots indicate that there is no single "golden feature": predictions arise from a combination of factors whose relative importance shifts by quantile. Although this may seem trivial, it is reassuring to see that the model predicts longer durations for conditions such as low visibility and high crosswind or tailwind.

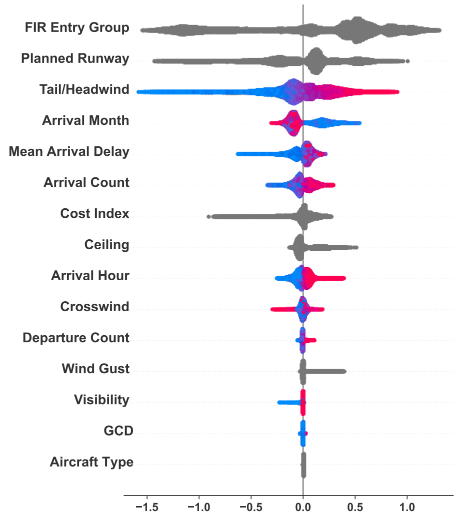
VIII. Discussion and Recommendations

In the following Subsection VIII-A, we reflect on the research outcomes. After, Subsection VIII-B provides an overview of suggested recommendations to be considered in future research.

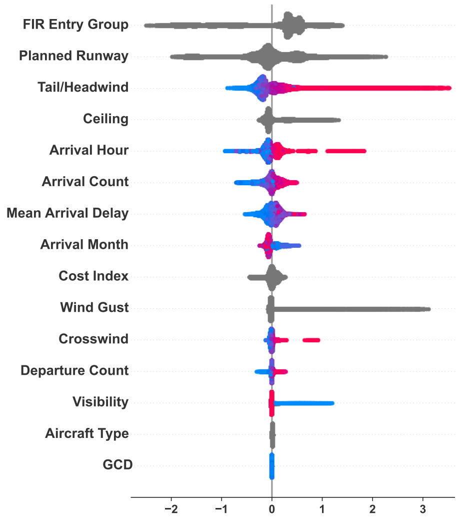
A. Discussion of Research Outcomes and Implications

The results demonstrate that a pre-departure, probabilistic FIR duration forecasting model based on quantile regression is both feasible and operationally meaningful. From a numerical perspective, the LightGBM quantile model yields a substantial improvement over the current operational baseline: point errors are reduced by more than 33% in MAE and 26 % in RMSE. However, the model explains around only one third of the variance in T_{FIR} ($R^2 \approx 0.33$) mainly due to the long forecast horizon and the dynamic nature of arrival operations.

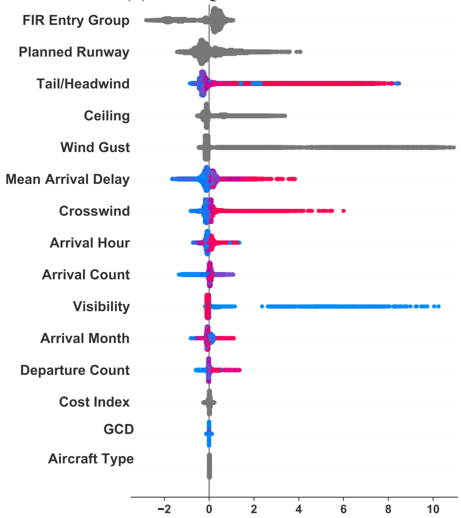
The proposed model is able to provide calibrated prediction intervals. The 90% interval between $\tau = 0.05$ and $\tau = 0.95$ achieves a Prediction Interval Coverage Probability of approximately 91 %, while maintaining a mean prediction interval of about 9.5 minutes. Additionally, there is a clear division of



(a) 5th Quantile Model



(b) 50th Quantile Model



(c) 95th Quantile Model

Figure 6. SHAP summary plots for different quantile models. Features are ordered by overall importance (top to bottom). Red indicates high feature values and blue indicates low feature values. GCD denotes the *Great-Circle Distance*.

roles: flight plan and traffic features, such as *FIR Entry Group*, *Planned Runway*, *Arrival Count*, and *Mean Arrival Delay*, dominate the central tendency of T_{FIR}, whereas weather features, including *Visibility*, *Ceiling*, *Tail/Headwind*, and *Wind Gust*, play a vital role for capturing tail risk and maintaining well-calibrated upper quantiles.

The results show that performance is strongest in the high-density central regime, roughly 17–23 minutes of FIR time, which covers the majority of flights. Errors increase for very short durations associated with shortcut cases and for long durations associated with elongations. For elongated flights, more than half of the largest misses cannot be attributed to weather or congestion features, pointing to unobserved drivers such as tactical ATC decisions and non-standard operations. Among the observed drivers, mis-forecasted low visibility is the most consistent correlate of under-prediction, with wind-gust surprises playing a secondary role. Shortcut flights, by contrast, predominantly arise when conditions are better than forecast, with higher visibility, lighter traffic, and more favourable winds than anticipated. In these cases, the model tends to over-predict because it cannot foresee opportunistic shortcuts that controllers only offer under nominal realised conditions.

B. Recommendations for Model Improvement

The presented work highlights several promising directions for future research. One natural extension is multi-airport generalization. The framework developed for AMS is conceptually agnostic, but applying it to other hubs would help distinguish which features are airport-specific.

Additionally, the relatively low R^2 further indicates that more informative or refined features are needed. A particular important element to add in the future is the likelihood of holding manoeuvrers. Moreover, ATC intent may be modelled more effectively with richer operational inputs, such as declared airport capacity, runway-configuration plans, or Arrival Manager (AMAN)-derived demand indicators. Inbound-peak capacities per runway configuration and visibility could allow the model to distinguish more clearly between nominal and near-capacity states rather than purely historical patterns.

Finally, the TAF data merging process could be enhanced by incorporating dispatcher input. TEMPO and PROB periods, in particular, may benefit from additional expert judgement. To adopt a more conservative approach with respect to forecast errors, one could select the period with the worst ceiling and visibility, even if this period has a broader validity range. Additionally, specific weather phenomena, such as fog or precipitation, could be explored as additional categorical weather features.

IX. Conclusion

Delays within the arrival airspaces, such as Flight Information Regions (FIRs) or Terminal Maneuvering Areas (TMAs), are highly dynamic and costly for airlines. However, current planning still relies largely on fixed deterministic estimates. This study addressed that gap by developing an explainable probabilistic forecasting model for flight time within the Amsterdam Schiphol (AMS) FIR, with a forecast moment in the pre-departure phase, approximately 90 minutes before take-off. Five LightGBM quantile models $\tau \in \{0.05, 0.25, 0.50, 0.75, 0.95\}$ were trained by minimising the Mean Pinball Error, and evaluated on a holdout set from 2025.

The resulting multi-quantile LightGBM model outperformed the current operational planning baseline. In point terms, it achieved reduced MAE and RMSE by more than 33% and 26% compared to current planning and explained about one third of the variance in T_{FIR} despite the long look-ahead horizon. In probabilistic terms, the central 90% interval between $\tau = 0.05$ and $\tau = 0.95$ achieved a Prediction Interval Coverage Probability of approximately 91% with a Mean Prediction Interval Width of 9.51 minutes. The prediction intervals adapt meaningfully to operating conditions, widening during high-traffic and adverse-weather periods and tightening under normal operating conditions.

Nevertheless, the analysis highlighted important limitations. The model remains constrained by the quality of its inputs and by the absence of explicit information on tactical ATC behaviour. More than half of the most extreme elongations could not be attributed to weather or congestion features and instead appear to be driven by unobserved operational factors and forecast failures, particularly in visibility and gusts. These findings underline that probabilistic models cannot eliminate uncertainty. However, they can structure and quantify it in a way beneficial to operations.

Overall, this study shows that moving from deterministic to quantile-based TMA time forecasts offers airlines a practical lever for more efficient and resilient operations. The proposed approach provides a clear, explainable baseline for risk-aware pre-departure planning at major hubs such as AMS. By planning against appropriate upper quantiles, dispatchers can explicitly account for tail-risk delays that simple averages miss. In benign situations, the same framework supports safe reductions in excess fuel. Nevertheless, further work is needed to incorporate richer operational context and forecast-uncertainty information, the low R^2 indicating that there is still room for improvement. Adding features such as runway configuration and capacity, and the likelihood of holding manoeuvres likely to improve efficiency on the model.

References

[1] SKYbrary, "Flight information region (fir)." https://skybrary.aero/articles/flight-information-region-fir, 2024. Accessed: 2026-02-09.

[2] H. G. Visser, "Terminal area traffic management," *Progress in Aerospace Sciences*, vol. 28, no. 4, pp. 323–368, 1991.

[3] EUROCONTROL, "Additional ASMA time: Performance indicator methodology," 2024. Accessed: April 2, 2026.

[4] M. S. Ryerson, M. Hansen, J. W. Bonn, and S. Karasaki, "Landing on empty: estimating the benefits from reducing fuel uplift in US civil aviation," *Environmental Research Letters*, vol. 10, p. 094002, sep 2015.

[5] Performance Review Commission, "Performance Review Report (PRR) 2024: An Assessment of Air Traffic Management in Europe," technical report, EUROCONTROL, Brussels, Belgium, March 2024.

[6] R. Dalmau, P. De Falco, M. Spak, and J. D. R. Varela, "Probabilistic pre-tactical arrival and departure flight delay prediction with quantile regression: A case study for geneva international airport using operational data," in *Fifteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2023)*, (Brétigny-Sur-Orge, France), June 2023.

[7] J. Zhang, Z. Peng, C. Yang, and B. Wang, "Data-driven flight time prediction for arrival aircraft within the terminal area," *IET Intelligent Transport Systems*, vol. 16, no. 2, pp. 263–275, 2022.

[8] Z. Wang, M. Liang, and D. Delahaye, "A hybrid machine learning model for short-term estimated time of arrival prediction in terminal manoeuvring area," *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 280–294, 2018.

[9] M. Vella, J. Gauci, and A. Dingli, "Data-driven prediction of aircraft holding times using opensky data," *Journal of Open Aviation Science*, vol. 2, 2025.

[10] R. Dalmau, P. Very, and G. Jarry, "On the causes and environmental impact of airborne holdings at major european airports," *Journal of Open Aviation Science*, vol. 1, pp. 1–13, 2024.

[11] G. N. Lui, C. H. Nguyen, K. Y. Hui, K. K. Hon, and R. P. Liem, "Enhancing aircraft arrival transit time prediction: A two-stage gradient boosting approach with weather and trajectory features," *Journal of the Air Transport Research Society*, vol. 4, p. 100062, 2025.

[12] P. Lothaller, M. Ribeiro, J. Sun, J. de Wilde, and A. Piva, "Contingency fuel prediction with explainable machine learning for airline operations," in *Proceedings of the SESAR Innovation Days 2025*, (Delft, The Netherlands), 2025. Operations & Environment, Aerospace Faculty, Delft University of Technology; ATM Strategy and Tactical Planning, KLM Royal Dutch Airlines.

[13] EUROCONTROL, "Arrival manager (aman) implementation guidelines and lessons learned," guideline, EUROCONTROL, December 2010. Cooperative Network Design.

[14] 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, no. 1, pp. 133–159, 2025.

[15] EUROCONTROL, "Airport corner." https://www.eurocontrol.int/tool/airport-corner, 2025. Accessed: 2025-02-09.

[16] M. Beltman, "Dynamically forecasting airline departure delay probability distributions for individual flights using supervised learning," MSc thesis, Delft University of Technology, Delft, Netherlands, December 2023.

[17] C. Dolman, "Forecasting of airline en route delay for individual flights using supervised machine learning," MSc thesis, Delft University of Technology, Delft, Netherlands, December 2024.

[18] R. Koenker and G. Bassett Jr, "Regression quantiles," *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

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

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

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

[22] 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.

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

[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 and Data Mining*, pp. 2623–2631, 2019.

[25] Scikit-learn Developers, "sklearn.metrics.pairwise.haversine_distances." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.haversine_distances.html, 2025. Accessed: 2025-12-18.

[26] R. Dalmau, "Metafora: A python library for parsing meteorological aerodrome reports (metars) and terminal area forecasts (tafs)," 2023.

[27] International Virtual Aviation Organisation, "Crosswind and headwind calculation." https://wiki.ivao.aero/en/home/training/documentation/Crosswind_and_Headwind_calculation, n.d. IVAO Documentation Library. Accessed: 2025-12-18.

[28] A. Achenbach and S. Spinler, "Prescriptive analytics in airline operations: Arrival time prediction and cost index optimization for short-haul flights," *Operations Research Perspectives*, vol. 5, pp. 265–279, 2018.

[29] M. Schäfer, M. Strohmeier, V. Lenders, I. Martinovic, and M. Wilhelm, "Bringing up opensky: A large-scale ads-b sensor network for research," in *IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks*, pp. 83–94, IEEE, 2014.

[30] Luchtverkeersleiding Nederland, "ENR 1.2 Visual Flight Rules." Electronic Aeronautical Information Publication (eAIP) Netherlands,

AIRAC AMDT 01-2026, 2026. Effective 22 January 2026. Accessed: 2026-01-09.

[31] E. V. Aretos and D. G. Sotiropoulos, “SEF: A method for computing prediction intervals by shifting the error function in neural networks,” *arXiv preprint arXiv:2409.05206*, 2024.

[32] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable ai for trees,” *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.