Issue 25(1), 2025, pp.45-66
ISSN: 1567-7141
TU Delft OPEN Publishing

EJTIR

SI: Sesar Innovation Days 2023
https://ejtir.tudelft.nl
DOI: 10.59490/ejtir.2025.25.1.7488

Combining Machine Learning Models to Improve Estimated Time of Arrival Predictions

Ramon Dalmau[1], Aymeric Trzmiel[2], Stephen Kirby[2]

[1] corresponding author ramon.dalmau-codina@eurocontrol.int, EGSD/INO/ENG, EUROCONTROL, France; https://orcid.org/0000-0003-3587-7331
[2] EGSD/INO/ENG, EUROCONTROL, France

Keywords
Machine learning
Flight predictability
Estimated time of arrival

Publishing history
Submitted: 12 April 2024
Revised date(s): 2 May 2024, 15 September 2024
Accepted: 17 September 2024
Published: 9 January 2025

Cite as
Dalmau, R., Trzmiel, A., & Kirby, S. (2025). Combining Machine Learning Models to Improve Estimated Time of Arrival Predictions. *European Journal of Transport and Infrastructure Research, 25*(1), 45-66.

Abstract

All aviation stakeholders require accurate estimated times of arrival in order to run flight operations as efficiently as possible. The time of arrival, however, is difficult to predict because it is affected by the uncertainties of the previous flight phases, with take-off time variability being the most significant contributor. At present, estimated time of arrival predictions are computed by the Enhanced Traffic Flow Management System, which collects data from a variety of sources to provide the best estimate throughout the entire duration of the flight. This paper introduces a novel approach that leverages existing machine learning models to enhance the accuracy of estimated time of arrival predictions, also during the pre-departure phase. More specifically, the first model (*Knock-on*) anticipates rotational reactionary delays arising from unrealistic available turn-around times; the second model (*FADE*) forecasts the evolution of air traffic flow management delays for regulated flights; and the third model, *AirborneTime*, was trained to identify systematic discrepancies between reported and actual airborne times. Using a dataset comprised of historical traffic and meteorological data collected during one year, this paper presents a comprehensive evaluation of this ensemble of models, referred to as PETA, against the current predictions across various time horizons, ranging from 6 hours before departure to the moment of take-off. The results indicate that the proposed solution surpasses the existing system in approximately two-thirds of the predictions. When the proposed solution performs better, the average and median improvements are 14 minutes and 7 minutes, respectively. However, when it underperforms, the average and median deteriorations are 7 minutes and 4 minutes, respectively. The optimal time frame appears to be between 2 and 6 hours before the departure time. This quantitative data is supported by feedback from European airlines, air navigation service providers and airports who used PETA in a live trial.

List of Abbreviations and Acronyms

A-CDM: Airport Collaborative Decision Making
ANSP: Air Navigation Service Provider
AOBT: Actual Off-Block Time
API: Application Programming Interface
ATA: Actual Time of Arrival
ATC: Air Traffic Control
ATFM: Air Traffic Flow Management
ATM: Air Traffic Management
ATT: Available Turn-Around Time
CASA: Computer-Assisted Slot Allocation System
DLY: ATFM Delay
DPI: Departure Planning Information
EATIN: EUROCONTROL Air Transport Innovation Network
ECAC: European Civil Aviation Conference
EFD: ETFMS Flight Data Message
EOBT: Estimated Off-Block Time
ETA: Estimated Time of Arrival
ETFMS: Enhanced Tactical Flow Management System
E/TMA: Extended/Terminal Manoeuvring Area
FADE: Forecast of ATFM Delay
FS: Filed Slot Allocated
FSA: First System Activation
IFP: Initial Flight Plan
MAE: Mean Absolute Error
METAR: Meteorological Aerodrome Report
NM: Network Manager
PDLY: Predicted Departure Delay
PETA: Predicted ETA
POBT: Predicted Off-Block Time
PTOT: Predicted Take-Off Time
SI: Slot Issued
TAF: Terminal Area Forecast
TOBT: Target Off-Block Time
TSAT: Target Start-Up Approval Time
TXOT: Taxi-Out Time

1 Introduction

In air traffic management (ATM) terminology, the time of arrival refers to the landing (or touchdown) time, whereas the in-block time refers to the event when the aircraft arrives at the parking position and the parking brakes are activated.

Accurate estimated time of arrival (ETA) is crucial for all aviation stakeholders, serving as a key input for numerous processes throughout a flight. From the airline's perspective, both the flight operations centre and the airline operations centre rely on accurate ETAs for efficient ground operations. These operations include gate and stand utilisation, ground handling, and staff planning. Accurate ETA predictions also enhance passenger connections and improve customer satisfaction. From an airport's perspective, inbound ETAs trigger the airport collaborative decision-making (A-CDM) process, optimising the flow of passengers and luggage, and the use of airport resources such as runways, taxiways, and gates, as well as ground services like transportation.

Lastly, for air navigation service providers (ANSPs), accurate ETA predictions significantly enhance arrival traffic management. By preventing congestion in the extended/terminal manoeuvring area (ETMA/TMA), these precise predictions allow for a more streamlined flow of incoming flights. This reduces the need for holding patterns, which in turn minimises delays and lowers fuel consumption.

At present, ETA predictions are offered to all stakeholders through the Enhanced Tactical Flow Management System (ETFMS). These predictions, however, remain subject to various uncertainties throughout different flight states due to factors such as air traffic flow management (ATFM) measures, weather conditions, air traffic control (ATC) practices, ATFM delay changes stemming from the Computer-Assisted Slot Allocation System (CASA) algorithm, as well as runway usage, for instance.

Figure 1 illustrates the main factors affecting ETA predictions. It should be noted that this figure is not exhaustive, but it nevertheless demonstrates the wide range of uncertainties that ETA predictions face. Furthermore, it is presupposed that the prediction of ETA becomes increasingly uncertain the farther the flight is from its actual time of arrival (ATA). Therefore, the development and availability of enhanced ETA predictions, in comparison to ETFMS estimations, across various look ahead times, would assist stakeholders in operating more efficiently, improving planning, enhancing predictability, and increasing punctuality.

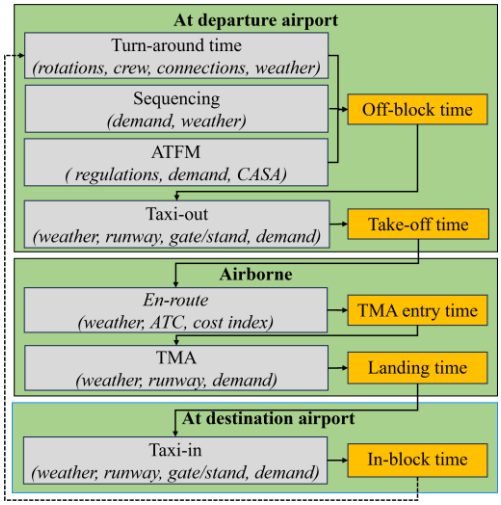


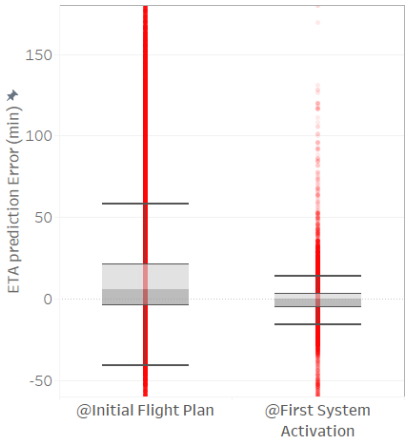
Figure 1. Flight states (green), events (orange), processes (grey) and sources of uncertainty within a process (italic).

To gain insight into the current situation, a comprehensive analysis of the ETFMS ETA predictive accuracy was conducted. This analysis encompassed 6 months of flight data from and to the 50 busiest airports in the European Civil Aviation Conference (ECAC) area, spanning from January to March and June to August 2022. The dataset comprised approximately 2M intra-ECAC flights monitored by the ETFMS. The ETA prediction error of each flight was computed as the difference between the ATA and the ETA as reported by the ETFMS. Therefore, positive values indicate that the ETFMS prediction was overly optimistic, i.e., the flight arrived later than the predicted ETA. The computation was made at two specific and representative events of the flight:

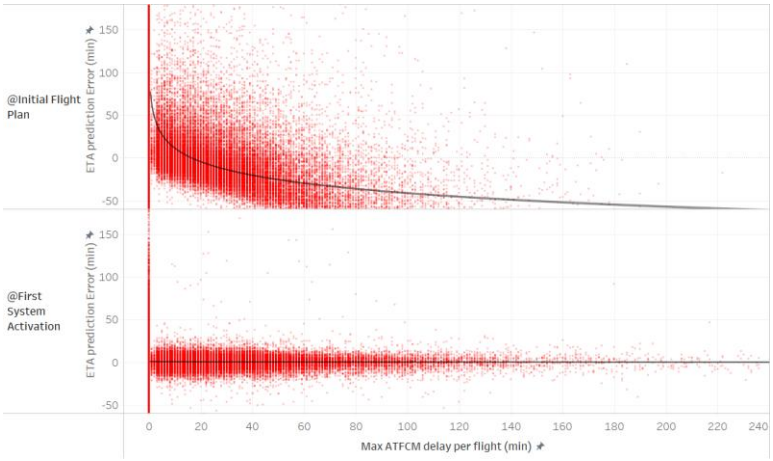
- at the submission of the initial flight plan (IFP), typically between 3 and 9 h before take-off, and
- at first system activation (FSA), i.e., when the first ATC message is received, typically right after take-off.

Figure 2(a) shows that the ETA prediction is more accurate and with less dispersion at FSA than at IFP. These results highlight that the prediction of the airborne time, which is equivalent to the ETA

prediction at FSA, is relatively accurate and that most of the current ETA prediction error is attributable to the take-off time uncertainty.



(a) Current ETA prediction error at IFP and FSA events.



(b) Correlation between current ETA prediction error and maximum ATFM delay.

Figure 2. Current difference between ATA and ETA (i.e., ETA prediction error) at IFP and FSA events, considering the ETA predicted by ETFMS. Positive values indicate flights arrived later than predicted, while negative values indicate they arrived earlier.

To further assess the possible cause of ETA uncertainties before take-off, Fig. 2(b) shows the ETA prediction error values as a function of the maximum ATFM delay assigned to the flight (if applicable, otherwise set to 0) from IFP to FSA. It should be noted that the ATFM delay for a flight can vary over time as CASA optimises slots to reduce ATFM delays. Therefore, the "maximum ATFM delay" here refers to the highest ATFM delay recorded for that flight from IFP to FSA. The results shown in Fig. 2(b) indicate that, as expected, the ETA prediction error at FSA (i.e., the airborne time prediction error) is independent of the maximum ATFM delay. Figure 2(b) also shows that the maximum ATFM delay impacts the ETA prediction error at IFP: As the maximum ATFM delay increases, the ETA error becomes negative (i.e., the error decreases, but not in absolute terms), indicating that flights arrive earlier than expected by the ETFMS. This behaviour occurs because the ETFMS initially predicts a significant delay and, consequently, a later ETA. However, as CASA improves the ATFM slots, the actual departure time is earlier than initially anticipated, resulting in an earlier arrival. These findings align with expectations, as the ATFM delay typically decreases due to the true revision process of the CASA algorithm.

Significant efforts have been made in recent years to improve ETA predictions, particularly within the academic and research communities. A multitude of models have been developed, with the majority relying on machine learning techniques and encompassing diverse types of data (e.g., flight information, weather, surveillance data for trajectory clustering), as well as various designs (e.g., artificial neural networks and gradient-boosted decision trees) (Strottmann, 2015; Ayhan, 2018; Wang, 2020a; Wang, 2020b; Christien, 2021). In this regard, the performance of several machine learning models has also been assessed and compared (Silvestre, 2021; Zhang, 2022). Recently, researchers proposed novel methods for accurately predicting ETAs in Beijing TMA (Ma, 2023) and for a multi-airport system (Wang, 2023). Specifically, the authors exploited spatio-temporal features based on clustering analysis of trajectory patterns, drawing on methodologies proposed in the previous research.

While demonstrating outstanding predictive capabilities, these studies primarily focused on addressing the ETA prediction challenge in the context of airborne time, that is, without considering uncertainties related to take-off times. Filling this gap, two machine learning models to improve ground delay predictions were developed under the EUROCONTROL Air Transport Innovation Network (EATIN) framework: (1) the so-called *Knock-on* model (Dalmau, 2024), which predicts the reactionary rotational delay for non-regulated flights, and (2) *FADE* (forecast of ATFM delay), which predicts the evolution of the ATFM delay for regulated flights (Dalmau, 2021).These models have made significant advances in predicting off-block and, as a result, take-off times. When combined with the *AirborneTime* model, which corrects residual errors in airborne time predictions, they could form the foundation of an advanced, data-driven ETA prediction system, effective from flight plan submission several hours before take-off.

This research paper offers two primary contributions. The first is the introduction of the prediction of ETA (PETA) algorithm, a novel approach that enhances ETA predictions, even before take-off, by utilising established machine learning models. Summarising the three machine learning models used by the PETA algorithm (*Knock-on*, *FADE*, and *AirborneTime*) is an integral part of this initial contribution. The benefits of employing existing and independent models are manifold. Firstly, it allows users greater control in determining which ETA contributors should be predicted using machine learning and which should rely on real-time information from ETFMS. Secondly, it circumvents the need for maintaining a large, complex model. Lastly, the performance of PETA inherently improves whenever any of its constituent models are retrained, eliminating the need for additional adjustments.

Regarding the second contribution, this paper provides a detailed performance evaluation of PETA's predictions by comparing them to the predictions of the current system (i.e., the ETFMS) over the course of a year, encompassing all intra-ECAC flights. This year includes three months of testing and eight months of live trial, during which several stakeholders requested PETA predictions via a dedicated application programming interface (API). The evaluation focuses primarily on ETA predictions made prior to take-off, as early as 6 hours in advance, when ETA uncertainty is at its highest level.

2 Methodology

The PETA system comprises the integration of several machine learning models, each specialising in predicting the duration of a specific process, as illustrated in Figure 3. At the time of writing this document, however, the model specialised in predicting the taxi-out time was not yet available, and ETFMS predictions were used instead.

Model	Knock-on	FADE	Current	Airborne time
Flight state	**At departure airport**			**Airborne**
Process	Turn-around Sequencing	ATFM	Taxi-out	En-route TMA
Output of the model	Reactionary delay	ATFM delay	Taxi-out time	Airborne time
Predicted event	Off-block time without ATFM delay →	Off-block time with ATFM delay →	Take-off time →	**Time of arrival (landing)**

Figure 3. PETA: combination of models to predict the ETA.

The decision to utilise machine learning over traditional methods, such as conventional linear regression, in this study is motivated by the complexity of the problem and the extensive amount of data collected by the Network Manager (NM). This data was instrumental in training effective models.

2.1 Individual models

The three models that constitute PETA are based on gradient-boosted decision trees, specifically the *LightGBM* implementation by Microsoft. Several factors influenced the choice of this type of model: (1) they are simple to train, (2) they can handle high-cardinality variables like airports, airlines or aircraft types, (3) they are robust to missing values, and (4) they consistently perform well with tabular datasets.

First, the *Knock-on* model predicts the rotational reactionary delay by taking various factors into account. These include the available turn-around time (ATT), specific flight attributes such as departure and destination airports, and the aircraft operator, as well as essential calendar features such as hour of the day, day of the week, and month of the year. Among all the features, the ATT is particularly important as it captures the inbound delay. Additionally, the model incorporates weather variables at the departure airport around the estimated off-block time (EOBT). It utilises variables such as wind speed, wind direction, visibility, cloud ceiling, and the presence of thunderstorms or snow, sourced from terminal area forecasts (TAFs) during inference and meteorological aerodrome reports (METARs) during training. The reasoning for using observations during training will be explained in more detail later. However, weather features are generally not the most significant factors on average, as weather conditions are often favourable. The main goal of this model is to improve off-block time predictions for non-regulated flights (i.e., flights not subject to ATFM regulations).

Second, the primary goal of *FADE* is to predict the final ATFM delay (DLY), right before departure, of a regulated flight. It should be noted that *FADE* does not predict which flights are going to be regulated, but just the expected delay of already regulated flights. In other words, a flight needs to be regulated to benefit from *FADE* predictions. Furthermore, *FADE* does not capture the scenario where a change in ATFM delay leads to missing a slot due to reactionary delay. Similar to the *Knock-on* model, its predictions are conditioned on several flight attributes, including the departure and destination airports. *FADE* also considers the current ATFM delay and the parameters of the ATFM regulation that determines the delay (i.e., the most penalising regulation), including the reference location, its reason (e.g., ATC capacity, weather, or industrial action) and the duration.

Third, the *AirborneTime* model was developed from scratch with the goal of improving airborne time predictions. This entails estimating the time it will take from take-off to landing. The *AirborneTime* model considers a variety of flight attributes, such as origin and destination airports, aircraft operator, and tactical flight data, as well as factors such as departure delay and ATT for the subsequent rotation operated by the same aircraft registration. Analogously to the *Knock-on* model, it includes calendar-related features and considers the expected weather conditions at the

destination airport around the ETA predicted by the ETFMS. The underlying hypothesis here is that these various features can collectively contribute to identifying systematic shortcuts along the route, airline-specific time buffers and speed adjustments to compensate for delays or save fuel, and additional airborne time required in the destination airport's TMA due to traffic congestion and/or adverse weather. It is important to note that the most crucial feature of this model is the airborne time as predicted by the ETFMS at that time; in other words, the ETFMS prediction itself is a key feature of the model. Consequently, this model heavily relies on ETFMS predictions and aims primarily to correct them.

Table 1 provides an overview of the top 10 most important features identified by Shapley analysis for the three models that compose the PETA ensemble. This table details each feature's type (numerical or categorical) and its variability — whether it changes over time for a flight from IFP to FSA or remains static. For instance, the ATT feature may change if the inbound flight updates the ETA and/or if the outbound flight changes the EOBT. A brief discussion of the Shapley analysis results will follow in Section 3.1.1. Note that none of the weather variables are not among the top 10 most important features in the *Knock-on* model, but wind speed is among the top features in the *AirborneTime* model.

Table 1. Overview of the top 10 most important features identified by Shapley analysis for the different models within PETA.

Model	Feature Description	Type	Variability
Knock-on	Available turn around time	Numerical	Dynamic
	Aerodrome of departure	Categorical	Static
	Aircraft operator callsign	Categorical	Static
	Aircraft type	Categorical	Static
	Current taxi time	Numerical	Dynamic
	Month	Categorical	Static
	Hour	Categorical	Static
	Aerodrome of destination	Categorical	Static
	Day of week	Categorical	Static
FADE	(Current ATFM) Delay	Numerical	Dynamic
	Number of regulations	Numerical	Dynamic
	Time to estimated off-block time	Numerical	Dynamic
	Late update	Categorical	Dynamic
	Protected location type	Categorical	Dynamic
	Protected location ID	Categorical	Dynamic
	Aerodrome of departure	Categorical	Static
	Aircraft operator callsign	Categorical	Static
	Flight state	Categorical	Dynamic
AirborneTime	ETFMS predicted fly time	Numerical	Dynamic
	Route distance	Numerical	Static
	Aerodrome of destination	Categorical	Static
	Aerodrome of departure	Categorical	Static
	Aircraft type	Categorical	Static
	Aircraft operator callsign	Categorical	Static
	Wind speed	Numerical	Dynamic
	Hour	Categorical	Static
	Available turn around time	Numerical	Dynamic
	Month	Categorical	Static

2.2 *PETA: combined models*

The idea behind the PETA system is illustrated in Algorithm 1. In this algorithm representation, each model is treated as a non-linear function of several parameters, $y = f(x_1, x_2, \dots)$, where the ellipsis (…) indicates "and other features", like those listed in Table 1. For instance, *Knock-on* predicts the rotational reactionary delay as a function of the ATT and other (…) features, thus *Knock-on*(ATT, ...) represents the reactionary rotational delay predicted by this model for a given set of input feature values. Lines 13-15 show how *Knock-on* and *FADE* are combined to predict the departure delay (PDLY). This doublet of models is expected to provide more accurate off-block time predictions (POBT) for both regulated and non-regulated flights.

Subsequently, in the absence of a model to predict the taxi-out time, the predicted take-off time (PTOT) is obtained by adding the taxi-out time (TXOT) as reported by the ETFMS. Finally, the airborne time as predicted by the *AirborneTime* model is added to the PTOT, resulting in the predicted time of arrival (PETA). The PETA is used to estimate the ATT of the next flight in the sequence, and the process is repeated until all flights have been processed. It is worth noting that, in contrast to standard ATM notation, and due to the absence of readily available taxi-in information, the ATT used by the *Knock-on* was defined as the difference between EOBT and the time of arrival of the previous flight, not the in-block time. This implies that *Knock-on* implicitly predicts the taxi-in time from the information provided in the inputs.

Algorithm 1. **PETA: Propagates predictions along flights operated by the same aircraft registration r to improve ETAs. The models are highlighted in blue.**

$\mathcal{M} \leftarrow$ Latest message of all flights operated by r
$\mathcal{M} \leftarrow$ Remove cancelled flights from $\mathcal{M}$
$\boldsymbol{f} \leftarrow$ Sequence of flights operated by r, sorted by EOBT
For $i = 1, \dots, |\boldsymbol{f}|$ **do:**
    If $\boldsymbol{f}(i)$ is a terminated flight:
        PETA(i) ← ATA(i)
    Else:
        If $\boldsymbol{f}(i)$ has departed:
            POBT(i) ← AOBT(i)
            PDLY ← AOBT(i) − EOBT(i)
        Else:
            ATT ← EOBT(i) − PETA($i-1$)

            PDLY ← *Knock−on*(ATT, …)
            If $\boldsymbol{f}(i)$ is regulated:
                PDLY ← max(PDLY, *FADE*(DLY(i),...))
            POBT(i) ← EOBT(i) + PDLY
        PTOT(i) ← POBT(i) + TXOT(i)
        PETA(i) ← PTOT + *AirborneTime*(PDLY, …)
End For

2.3 *Data and training*

The three datasets used to train the three machine learning models that compose PETA, respectively, were constructed using ETFMS flight data messages (EFDs) and weather observations from METARs. Each training dataset covers the period from January 1st, 2022, to February 28th, 2023, spanning slightly more than one year.

The EFDs are triggered, for example, when the CASA-assigned ATFM delay changes, when the airspace user updates the flight's route, and when a departure planning information (DPI) message

is sent when departing from a CDM airport. Each EFD includes the most up-to-date information for the emitting flight, such as the EOBT, ETA, aircraft type and registration, current taxi time, as well as the current ATFM delay and the ATFM regulations affecting the flight (if any).

Concerning the weather data, when deciding whether observations (i.e., METARs) or forecasts (i.e., TAFs) are more suitable for training a model, it is crucial to understand the context and goals of the model. Both approaches have their merits, and the choice depends on the specific question the model aims to answer.

If the goal is to learn the true cause-and-effect relationship between ETA variability and weather, training the model with weather variables extracted from METARs is preferable. For instance, if a flight was delayed due to a thunderstorm reported in the METAR, the model would learn to associate such conditions with delays. Ultimately, it was the thunderstorm itself, not the prediction of it (from a TAF), that caused the delay. One could argue that thunderstorm forecasts might influence the behaviour of agents within the system, potentially causing delays even if the thunderstorm does not occur. However, this philosophical discussion is beyond the scope of this paper.

On the other hand, if the objective is to learn the relationship between ETA variability and weather as predicted at a given look-ahead time, the model should be trained using TAFs. Although this method aligns with the common data science practice of "train the model with the data available at inference time", it has some drawbacks. First, it introduces a dependency on the accuracy of the forecasts. For example, if a TAF predicts clear skies but a storm causing delays occurs instead, the model trained on TAFs would be misled, attempting to identify a correlation between (predicted) clear skies and delays, which does not make sense. Second, the model would implicitly learn the errors inherent in the forecasting system. If the forecast accuracy improves over time, the model might still compensate for errors that no longer apply, reducing its performance unless it is retrained with updated forecasts. Third, the model will be valid for only one look-ahead time, as it will essentially learn the ETA variability given a predicted weather at a specific look-ahead time, since the predicted weather changes with the look-ahead time.

Training with METARs has the advantage of allowing the model to make predictions at any time, as long as a reliable weather forecast of the expected observations is available. This flexibility is particularly valuable because it ensures that as weather forecasting systems improve, the model's performance can also improve without needing retraining. For example, if new forecasting technology reduces errors in TAFs so that TAFs and METARs become more closely aligned, a model trained on METARs will automatically benefit from the more accurate forecasts during inference.

This consideration has been carefully addressed in the proposed models, which utilise METARs for training to capture the actual impact of weather on flight delays, while using TAFs during inference to make forward-looking predictions. This approach ensures that the errors within the models that we can control are kept separate from the weather forecast errors that are beyond our control. The raw METARs (for training) and TAFs (for inference) were processed to extract the weather variables with the open-source library *metafora*[1].

[1] https://github.com/ramondalmau/metafora

3 Results

This section presents the outcomes of an evaluation conducted using historical flight and meteorological data. Two distinct datasets were utilised to assess performance. The first dataset encompasses the four-month period designated as the test set, while the second dataset spans the eight-month period when the model was provided to some airlines, airports and ANSPs during a live trial. Both datasets include all intra-ECAC flights operated by aircraft listed in the Base of Aircraft Data, which accounts for 95% of all aircraft types.

Each observation within these datasets corresponds to an EFD transmitted by a flight. Consequently, the reader should keep in mind that the dataset contains more observations than flights, and a flight may contribute to the performance metrics multiple times. Furthermore, each observation (i.e., EFD) contains the most up-to-date flight information available at the moment of its emission. This information serves as input for the three sequential machine learning models. Additionally, it is important to emphasise that the *Knock-on* model used in the experiment incorporates weather information at the departure airport around the EOBT. Similarly, the *AirborneTime* is dependent on weather data at the destination, which is based on the current ETA as predicted by the ETFMS. As discussed in the previous section, METARs at EOBT/ETA were used for training, while the latest TAFs before the submission of the EFD, covering the EOBT/ETA, were used for inference.

The performance of the individual models that compose PETA, as well as the entire ensemble, on the test and live trial sets is consistently compared against the predictions of the ETFMS at the exact same time. The ETFMS predictions are extracted from the EFD used to populate the input features for the models. For clarity, the ETFMS predictions will be referred to as "Current" throughout the remainder of this paper.

3.1 Test set (from March 1st to June 30th, 2023)

Section 3.1.1 provides an overview of the performance metrics for the individual models in the test set, each predicting its respective target independently. Specifically, for flights that were not regulated from submission of the IFP to termination, it entails comparing the predicted departure delay according to the *Knock-on* model with the actual departure delay value. Similarly, this section compares the predicted ATFM delay by *FADE* in each observation of a regulated flight with the actual ATFM delay just prior to departure. It also includes a comparison of the predicted airborne time, as generated by the *AirborneTime* model, with the actual airborne time for all flights, regardless of whether they were regulated or not.

Section 3.1.2, on the other hand, delves into the collective performance of the PETA system in the test set. In this evaluation, predictions are still generated on an observation-by-observation basis, but the ensemble's inputs are based on predictions made for the previous flight operated by the same aircraft recursively. As a result, the distinction between regulated and non-regulated flights may be obscured, as predictions are made on an aircraft registration basis rather than for each individual flight.

To clarify, when a flight is subject to one or more ATFM regulations, the *FADE* model predicts the ATFM delay specifically for that flight. This predicted ATFM delay, along with the predicted reactionary rotational delay, contributes to a ground delay that can propagate to subsequent flights operated by the same aircraft, as outlined in Algorithm 1. Consequently, even flights not directly subject to any ATFM regulation may experience delays due to the cascading effect of ATFM regulations affecting previous flights of the same aircraft. The propagation of predictions is illustrated in Fig. 4.

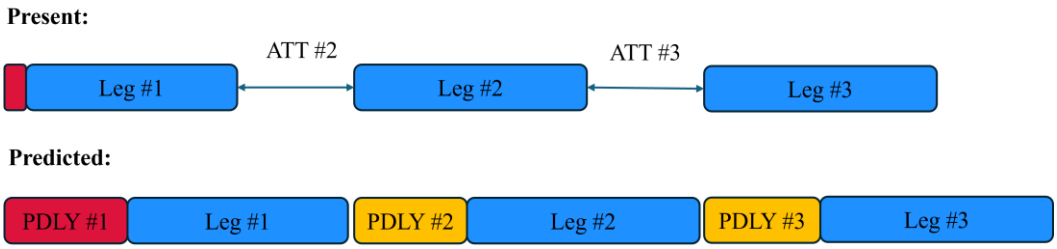


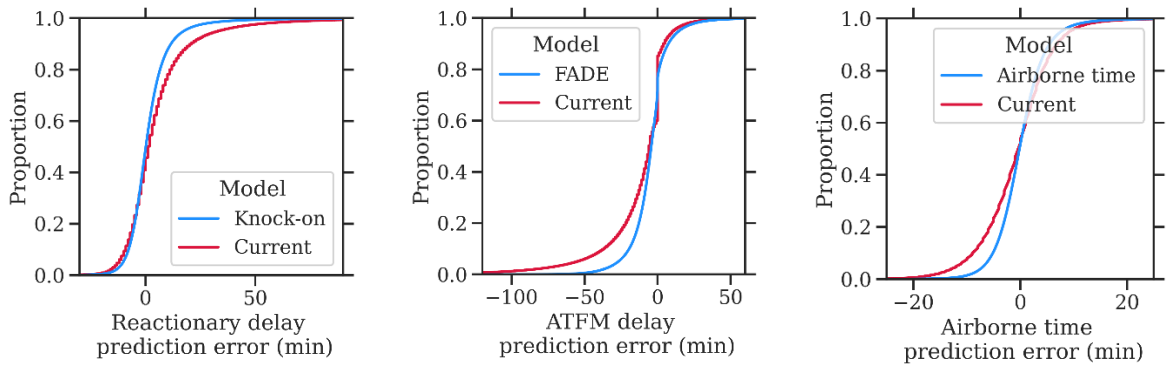
Figure 4. Illustrative example showing how flight legs not directly affected by ATFM delays could still be predicted to experience ground delays due to the anticipated ATFM delay of preceding legs. The blue boxes represent the block time from off-block to in-block; the red box indicates the ATFM delay; and the orange box denotes the reactionary rotational delay. Flight legs are shown in chronological order, from the left to the right. The first flight is regulated and FADE predicts an increase of ATFM delay, which propagates to the subsequent legs as rotational reactionary delay.

Thus, to avoid the introduction of arbitrary classification rules for predictions, no distinction was made between regulated and non-regulated flights. As a result, the ensemble's performance metrics shown in Section 3.1.2 apply to all flights, regardless of their ATFM status, when predicting the ETA.

3.1.1 Individual models

This section thoroughly examines the performance of each of the three models in predicting their respective targets. To streamline the presentation and maintain conciseness, performance metrics are aggregated for the entire test set without differentiating between various look-ahead times. Nevertheless, it is important to note that the performance of the models may vary with the look-ahead horizon due to the dynamic nature of some features. For instance, one of the most important features of *FADE* is the time to EOBT, which means its performance is likely to change with different prediction horizons. Similarly, the ATT of the *Knock-on* model depends on the ETA and EOBT of the inbound and outbound flights, which may change with time. An analysis of how performance degrades with look-ahead time for the entire PETA ensemble, due to the quality of the input data, will be presented in subsequent sections.

Figure 5 shows the (signed) cumulative prediction error distribution of both current and machine learning models. Complementing this figure, Table 2 presents the key metrics of the absolute prediction error distribution for the machine learning models in comparison to the current predictions.



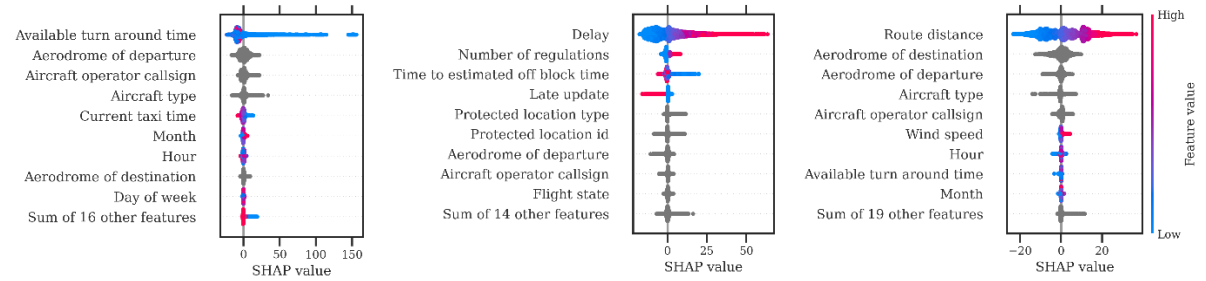
(a) Reactionary delay *(b) ATFM delay* *(c) Airborne time*

Figure 5. Empirical cumulative distribution function of the various prediction errors in the test set.

Table 2. Absolute prediction error distribution metrics (min) in the test set.

Output	Reactionary delay		ATFM delay		Airborne time	
Model	Current	*Knock-on*	Current	*FADE*	Current	*AirborneTime*
Mean	10.3	6.6	15.2	9.7	5.0	3.4
Std.	15.0	8.8	23.2	10.7	4.5	3.4
5th Perc.	6.0	4.4	0.0	0.0	4.0	2.6
25th Perc.	0.9	0.4	0.0	2.5	0.3	0.2
Median	3.0	2.0	8.0	6.7	1.9	1.2
75th Perc.	12.0	8.1	19.0	13.2	7.0	4.6
95th Perc.	34.0	18.6	56.0	30.6	13.6	9.3

Figure 6 displays the Shapley value distribution for the different models. In short, the Shapley values represent the average marginal contribution of each feature in the model across all possible combinations of features. This means that the Shapley value for a given feature is the average difference in the model's prediction when that feature is included versus when it is not, considering all possible subsets of the other features. This provides a measure of the importance of each feature in the model's predictions. For more information on Shapley values, please see (Lundberg, 2020) and the references therein. In this kind of figure, the y-axis indicates the name of the features, in order of mean absolute Shapley value from the top to the bottom. Each dot in the x-axis shows the Shapley value of the associated feature on the prediction for one observation, and the colour indicates the magnitude of that feature: red indicates high, while blue indicates low. By definition, positive (resp. negative) Shapley values increase (resp. decrease) the prediction with respect to the expected value of the target in the train set. The specific results will be discussed in their respective sections.



(a) Reactionary delay *(b) ATFM delay* *(c) Airborne time*

Figure 6. Distribution of Shapley values for PETA in the test set. For the (c) Airborne time, the ETFMS predicted airborne time is the most important feature. This key feature has been omitted in this figure because its Shapley values are an order of magnitude higher than those of the other features, which would overshadow their importance.

Knock-on

The prediction error of this model is computed as the difference between the actual off-block time (AOBT) and the POBT. Positive values indicate that the model is overly optimistic, predicting less reactionary delay than what actually occurred, whereas negative values indicate that the flight departed earlier than expected.

The *Knock-on* predictions are compared against the off-block time as reported in the EFD. It is important to note that the off-block time may undergo updates during the flight's course, often prompted by delay messages from the aircraft operator. Similarly, for CDM airports, more precise

off-block time estimations can be provided in the form of target off-block time (TOBT) or target start-up approval time (TSAT). The current model (i.e., the ETFMS) effectively takes in to account these updates.

In terms of absolute off-block time prediction error, Table 2 shows that the *Knock-on* model reduces the mean value by roughly 30% (from 10.3 to 6.6 min). This reduction is also visible in the remaining distribution metrics, albeit with slightly different absolute and relative figures. While a 30% improvement is certainly significant, it raises an important question: What is the operational impact of a 30% enhancement in off-block time prediction? This crucial question, which also applies to the other models and thus PETA, will be further explored in the conclusion of this paper.

Figure 5(a) shows that, when compared to the current model, the *Knock-on* model consistently improves off-block time predictions for non-regulated flights. This improvement is mostly visible on the positive side of the distribution, indicating the *Knock-on* model's ability to anticipate reactionary delays well before the aircraft operator updates the off-block time information in the system with more precise values.

The significant improvement observed can be attributed to a critical distinction: the current model does not include the minimum turn-around time when identifying overlapping consecutive flight plans operated by the same aircraft registration. In practical terms, this could result in scenarios where the arrival time of a flight aligns unrealistically closely with the off-block time of the subsequent flight, operated by the same aircraft registration, until the aircraft operator provides more accurate timing information. The *Knock-on* model, on the other hand, excels at identifying these scenarios by leveraging historical observations to learn about the minimum turn-around time, which effectively becomes a latent variable of the model. This figure also shows that, albeit to a lesser extent, the *Knock-on* model demonstrates the ability to identify flights that systematically depart earlier than expected. This, in turn, helps to mitigate the negative tail of the cumulative prediction error distribution, diminishing overly pessimistic predictions.

Figure 6(a) reveals, unsurprisingly, that the most significant feature of the *Knock-on* model, in terms of mean absolute Shapley values, is the ATT. Because the ATT is computed based on the ETA of the inbound flight, this consequently implies that the performance of the *Knock-on* is highly sensitive to the quality of the previous ETA prediction. Lower values of this feature are associated with a high rotational reactionary delay. Other notable features include the aerodrome of departure, the aircraft operator, and the aircraft type. Weather-related features also play a role, although their significance is considerably less.

FADE

The prediction error of this model is computed as the difference between the actual ATFM delay right before departure and the predicted one. Like the *Knock-on* model, positive values indicate that the model was overly optimistic, predicting too much ATFM delay improvement, whereas negative values indicate that the flight departed with less ATFM delay than that assigned by CASA at the prediction time. In this case, the current model, in absence of a more elaborated baseline, consists of using the current ATFM delay assigned by CASA (reported in the EFD) as the best prediction.

The performance of *FADE* depends significantly on the look-ahead time, with the time to EOBT being a crucial feature. *FADE* was trained using various ATFM delay updates for each flight, meaning each training observation corresponds to a specific message about the flight rather than the flight itself. These messages can vary greatly, with some sent 5 hours before EOBT and others just 30 minutes prior. The same methodology applies to the evaluation process. During *FADE*

evaluation, aggregated metrics are computed based on these individual messages rather than entire flights, disregarding the look-ahead time to ensure consistency with the evaluation of the other two models. For a detailed assessment of performance relative to the prediction time horizon, please refer to the evaluation of the PETA ensemble.

In terms of the absolute ATFM delay prediction error distribution, as shown in Table 2, *FADE* manages to reduce the mean error by approximately 5.5 min (36%). It is worth noting that a large portion of this reduction is due to observations on the negative side of the signed ATFM delay prediction error distribution, as discussed in the previous paragraph. Furthermore, other key distribution metrics show significant improvements, with a particular emphasis on the 95th percentile, which is reduced by 25.4 min (45%). These findings suggest that the evolution of ATFM delay can be predicted to a certain extent using historical observations.

Figure 5(b) shows that *FADE* outperforms the current model in the negative tail of the cumulative prediction error distribution. This fact is consistent with expectations, given that the ATFM delay assigned to flights is frequently reduced due to the CASA algorithm's optimisation efforts. CASA attempts to improve the ATFM slots of regulated flights, ensuring that they depart as close to their EOBT as possible, through the so-called true revision process. As a result, current model's predictions of ATFM delay are generally pessimistic, particularly when made well ahead of the EOBT.

Figure 5(b) also highlights an important point: *FADE* faces difficulties in determining whether the ATFM delay will remain stable or increase. In these scenarios, the current model outperforms *FADE*. This gap can be attributed to *FADE*'s lack of network awareness, as it generates predictions based solely on flight-specific information, without taking into account other ATFM regulations present in the network even if not directly affecting the flight. These unaccounted-for regulations could potentially have a greater impact on the flight, causing drastic changes in its delay. To effectively address this issue, future work should focus on developing a network-aware model for *FADE*. Such a model should be capable of identifying situations in which the ATFM delay remains unchanged or increases due to the complex interaction between regulations, allowing for more accurate predictions. The authors believe that expanding the training dataset laterally (i.e., to include more features) would bring more performance benefits than expanding it vertically (i.e., to add more observations).

Figure 6(b) demonstrates that the most significant feature of the *FADE* model, in terms of mean absolute Shapley values, is the current ATFM delay. This is followed by the number of regulations affecting the flight and the look-ahead time of the prediction relative to EOBT. The protected location (be it airspace, airport, group of airports, or point) where the most penalising regulation affecting the flight is implemented also plays a non-negligible role.

AirborneTime

The prediction error of this model is computed as the difference between the actual airborne time and the predicted value. Notably, unlike the previous models, positive values in this context indicate that the model was overly optimistic, predicting a shorter duration than the actual flight time, whereas negative values indicate that the flight completed its journey in less time than anticipated. In the case of the current model, the prediction is based on the difference between the EFD's ETA and estimated take-off time (ETOT) at prediction time.

The metrics presented in Table 2, particularly the absolute airborne time prediction error distribution, clearly show that the improvement with respect to current values remains somewhat modest in absolute terms (measured in min). However, it is important to note that the relative

improvement is not insignificant, amounting to approximately 30% when the MAE is considered. Unlike *FADE* and the *Knock-on* models, which achieve significant reductions in MAE by several min, the gains achieved by the airborne model are expected to be more limited. Notably, ETA predictions made by the current system when the flight is already in flight or very close to take-off are very accurate. As a result, there is little room for improvement in such scenarios, and the majority of research efforts aimed at improving ETA predictions should be directed towards improving take-off time predictions.

Figure 5(c) shows that the *AirborneTime* model is effective at improving current predictions at both ends of the distribution, with the most notable improvements occurring on the negative side. This finding indicates that the *AirborneTime* model succeeds at identifying flights that consistently complete the journey in less time than current estimates. Such deviations can occur as a result of a variety of factors such as time buffers, speed adjustments, or ATC shortcuts, among others. Furthermore, the minor improvement observed on the positive side suggests that the airborne model has a greater ability to identify flights that will spend more time in the air than the current model originally predicted. This could be attributed to factors such as bad weather or recurrent traffic congestion at the destination airport.

Figure 6(c) indicates that, based on mean absolute Shapley values, the distance of the planned route appears as the most significant feature of this model. However, it is crucial to highlight that the model also considers the predicted airborne time according to ETFMS, which is actually the most important feature. This key feature has been omitted from Fig. 6(c) because its Shapley values are an order of magnitude higher than those of the other features, which would overshadow their importance. Following this, the city-pair (origin and destination airports), aircraft operator, and aircraft type are also important features. Similar to the *Knock-on* model, features related to the weather conditions at the destination are utilised by the model, but they do not significantly contribute to the prediction. This is again in terms of mean absolute values.

3.1.2 PETA: combined models

This section presents the quantitative results of PETA predictions (i.e., the amalgamation of the three models) on the test set. The ETA predictions generated by PETA will be compared to those of the current system under identical conditions. It is important to note that, as in the previous section, the term "Current" refers to the ETFMS.

To begin, Fig. 7 shows the distribution of (signed) ETA prediction errors in the test set, allowing for a comparison with the results presented for the individual models (see Fig. 5). These errors are computed as the actual time of arrival (ATA) minus the predicted ETA, consistent with previous evaluations. Consequently, positive values denote unexpected delays, whereas negative values indicate that the flight arrived at the destination airport earlier than predicted.

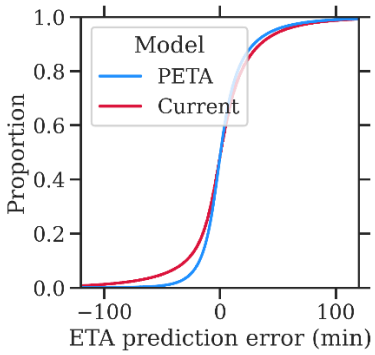


Figure 7. Empirical cumulative distribution function of the ETA prediction error in the test set. Remember that "Current" refers to ETFMS predictions.

Figure 7 closely aligns with the cumulative distributions previously shown for the individual models. The reader should keep in mind that accurate predictions have a cascading effect, positively influencing predictions for subsequent flights in the sequence, thereby amplifying the overall improvement in ETA predictions. Additionally, the shortcomings observed in *FADE*'s performance on the positive side of the ATFM delay prediction error distribution are partially offset by the advantages offered by the *Knock-on* model in that region. Specifically, the *Knock-on* model excels in predicting flights with delayed departures resulting from rotational reactionary delays, thereby contributing to a more balanced performance.

Complementing these results, Figure 8 presents a histogram illustrating the differences in absolute ETA prediction errors between the current system and PETA within the test set. Each bar in this histogram represents the frequency of a particular difference in error values. Specifically, each observation corresponds to a single prediction, and the value for each observation was determined by first calculating the absolute difference between the actual time of arrival (ATA) and the current system's estimated time of arrival (ETA), and then subtracting the absolute difference between the ATA and PETA's ETA. As a result, the positive side of the distribution shows the number of instances where PETA's predictions were more accurate than those of the current system, in terms of absolute error. Conversely, the negative side indicates the number of instances where the current system outperformed PETA in predicting the ETA.

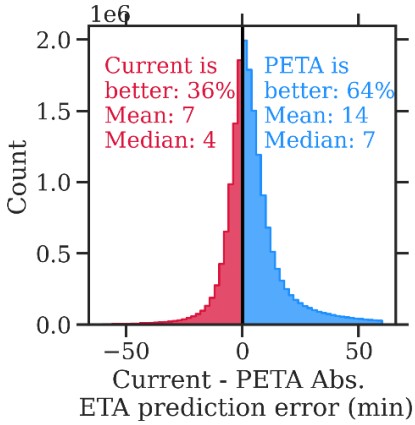


Figure 8. Histogram of the differences between the current and PETA absolute ETA prediction errors in the test set. For completeness, the mean and median absolute values of both the blue and red distributions are also included.

Figure 9 presents the same values (absolute ETA prediction error difference between the current system and PETA) but shows the average error for the top 50 airports with most intra-ECAC arrivals.

Figure 9 shows that, on average, PETA provided more accurate ETA predictions than the current system for all considered airports, ranging from 2 min better for Catania airport (LICC) to 13 min better for Alicante airport (LEAL). More detailed analysis is required to understand the large differences between airports. As an example, a relatively high percentage of regulated flights for a given destination airport might positively impact the PETA predictions, allowing *FADE* to improve upon the current system's predictions.

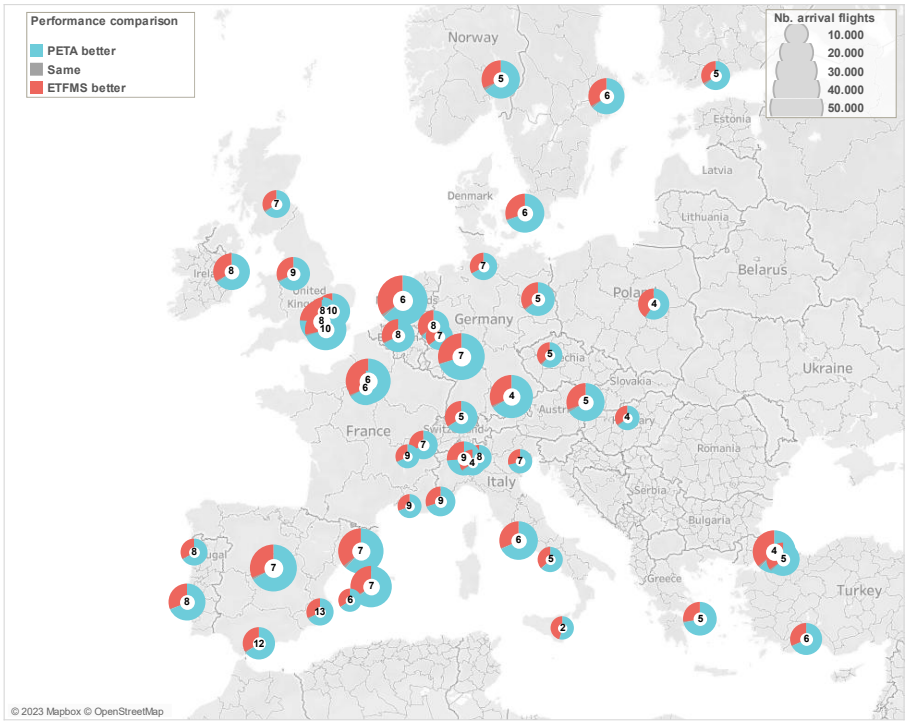


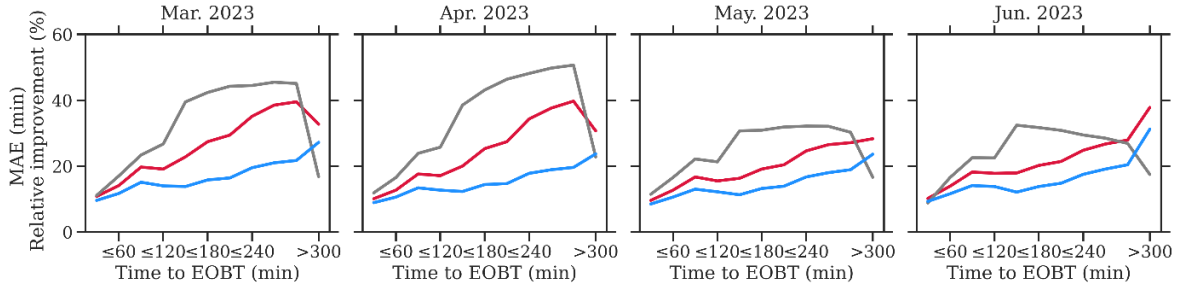
Figure 9. Differences between the current and PETA absolute ETA prediction errors per airport in the test set. (Each circle is an airport with its size proportional to the number of arrivals considered.)

3.2 *Live trial set (from July 1st, 2023, to February 29th, 2024)*

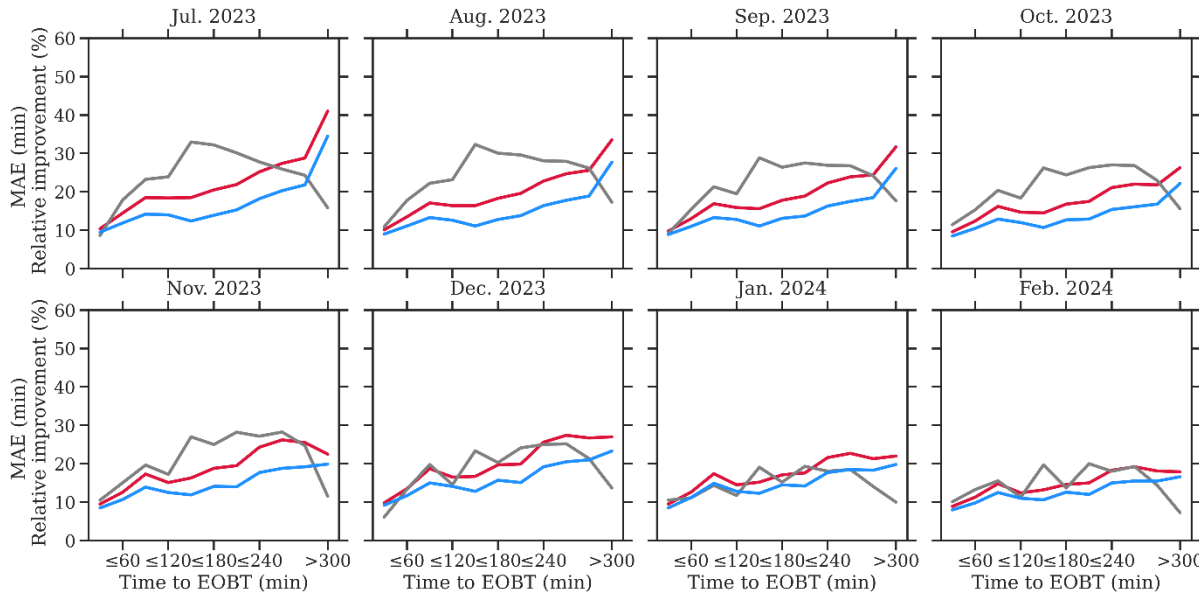
This section goes beyond hypothetical performance on the test set and discloses the results of comparing PETA predictions to those of the current system over an eight-month period of real-world operation. Throughout this period, PETA was frequently utilised by a variety of stakeholders via a dedicated API. The subsequent qualitative feedback from users reinforces the quantitative results.

3.2.1 Quantitative results

Figure 10 shows the mean absolute error (MAE) of the ETA predicted by PETA and the current system, grouped by month, as a function of the time to EOBT. This figure also shows the relative improvement (expressed as a percentage) of PETA with respect to the current system. It should be noted that, despite the primary goal of this section is to present the results from the live trial, the authors believed it was important to include equivalent figures from the test set for comparative analysis. The reader will quickly notice that this comparison aids comprehension of the observed evidence.



(a) Test set



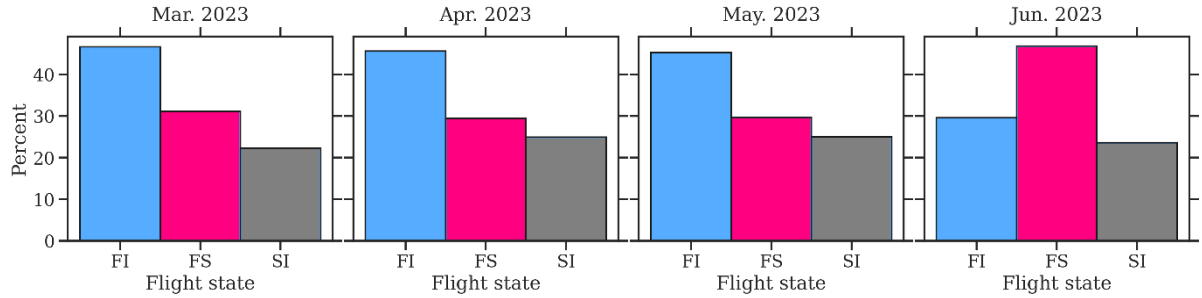
(b) Live trial set

Figure 10. Mean absolute error and relative improvement as a function of the time to EOBT. Being consistent with the notation, blue represents PETA, red means current, and grey indicates the relative improvement.

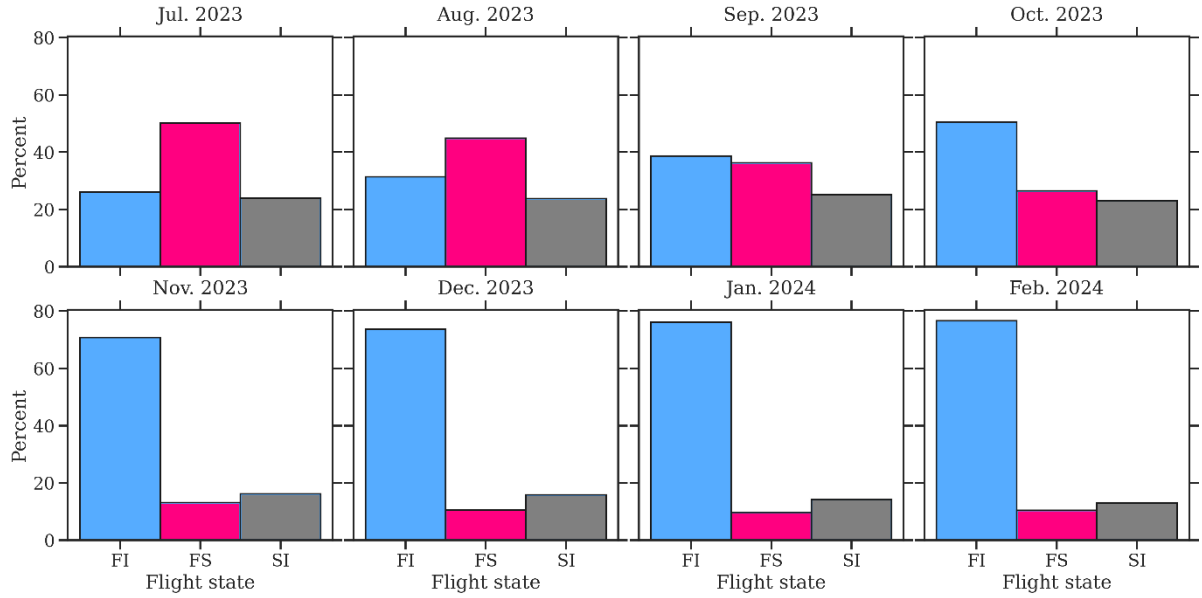
Figure 10 demonstrates that an ensemble of machine learning models working collaboratively to improve ETA predictions consistently outperforms current predictions across different look-ahead times. This observation holds particular significance within the look-ahead times ranging from 2 to 6 h before the EOBT. As one approaches EOBT, existing predictions tend to be already quite accurate, leaving limited room for improvement. Conversely, when further away from EOBT, the information feeding into the machine learning models becomes more uncertain, consequently affecting the predictions made by the ensemble. It is crucial to bear in mind that, just as accurate predictions have a cascading positive effect on performance, any inaccuracies (e.g., stemming from unreliable input data far from EOBT) can have a detrimental impact on overall performance. These findings suggest that the proposed ensemble could provide the most significant operational benefits between 2 and 6 h before EOBT, and that its usage outside of this time frame may not be as beneficial.

An intriguing observation stemming from Fig. 10 is that the relative improvement during the test set and the initial months of the live trial was notably higher than in the later months of the trial. One might initially attribute this trend to a data drift issue necessitating re-training of the constituent models within PETA. However, upon closer examination, it became evident that the relative improvement of PETA is intricately linked to the severity and volatility of ATFM delays within the network: in the absence of severe and dynamic ATFM delays, *FADE* offers no significant benefits. Furthermore, the reduction in rotational reactionary delays, triggered by primary delays such as ATFM delays, makes the *Knock-on* model less essential in the ensemble. In such optimistic scenarios for flight operations, the current system performs admirably, and any potential relative improvement looks small in comparison.

The assertion made earlier gains support from Fig. 11, which depicts the distribution of flight states precisely at the moment of prediction, independently of the look-ahead time. In simpler terms, for every individual prediction (i.e., EFD), the flight's specific state was captured to create this visual representation.



(a) Test set



(b) Live trial set

Figure 11. Distribution of flight states. FI: Filed, FS: Filed slot allocated, SI: Slot issued.

To clarify, a flight is in the state "Filed" (FI) after ETFMS received its flight plan. A flight is in the state "Filed, Slot Allocated" (FS) after slot allocation was applied to it due to an ATFM regulation. A flight is in the state "Filed, Slot Issued" (SI) after slot allocation was applied to it and the corresponding slot allocation message was sent. The Slot Allocation Message is sent two hours before the EOBT of each pre-allocated flight, known as Slot Issue Time 1. An allocated slot cannot be taken by another flight, unless the regulation is deep rectified, and the calculated take-off time has not been forced. Moreover, the slot allocated to a flight may be improved by the true revision process of CASA. Thus, flights in FS and SI state are subject to ATFM regulations, while flights in FI state are not.

Figure 11 provides valuable insights. Notably, during months marked by a greater relative improvement (as observed in Fig. 10), like March and April 2023, a substantial proportion of predictions were made for flights in the FS and SI states. Conversely, in months with lower reported relative improvement, like February 2024, most predictions corresponded to flights in FI state (i.e., not regulated). This intriguing pattern suggests that PETA delivers its most significant benefits during challenging network conditions.

It should be noted that the performance of the PETA ensemble, presented in this section, is a cumulative result of the contributions from three distinct models. An initial analysis, which involved selectively deactivating individual models within the ensemble to assess their marginal

contribution on the overall performance, revealed that *Knock-on* and *FADE* are the primary contributors to PETA's performance. Interestingly, their relative contributions are situation-dependent: on days with a high volume of ATFM regulations, *FADE* takes precedence, while on regular days, *Knock-on* proves to be more important. In contrast, the *AirborneTime* model contributes little to overall performance.

Finally, it is important to note that the three PETA models operate in a cascading fashion and along a flight sequence. This means that any incorrect prediction of one model may have negative consequences for subsequent predictions (for the same or next flights). For example, as illustrated in Fig. 4, if the ATFM delay of the first flight leg had remained unchanged at a relatively low value and assuming no other operational disruptions, not only would the predicted ETA for the first flight leg be incorrect, but the predicted ETAs for subsequent flight legs would also be wrong due to the predicted propagation of ground delays -- an outcome that did not actually occur. The sensitivity of each model to errors in their inputs, which should not be confounded with the marginal contribution discussed in the previous paragraph, remains unquantified. This will be the focus of future research.

3.2.2 Qualitative results

As mentioned in Section 3.2, the PETA stakeholders have been provided with PETA predictions all along the live trial period (from July 2023 to February 2024) either through offline data (monthly predictions) or via dedicated API internally developed. The motivation was to assess how the PETA predictions could be used in current and live situations and what benefit it could bring to stakeholders in their daily duties. This section describes stakeholders' subjective feedback collected via an online survey and based on six answers: three from airlines, two from ANSPs and one from airport.

Over the four of participants who used PETA, the level of usage was reported from low (3 ratings) to medium (1 rating) with half of the users who reported to send more than 500 API request per week. The main usage was to get PETA predictions for operational use or for post operation analysis. However, no one made operational decisions based on PETA predictions.

Subjective feedback on PETA performance is consistent with quantitative results previously reported: The absolute performance of PETA was found overall high (4 ratings out of 6 in and all the participants found that PETA predictions are better than the ones provided by NM or by their internal tool). Finally, considering the valuable PETA predictions, some of the stakeholders express the need to be incorporated into the NM systems to feed their internal tools (e.g. airport Demand and Capacity Balancing tool).

4 Conclusions

Results have shown that PETA's ETA predictions are better (have a smaller absolute error) than the current system's ETA predictions for about two thirds of the flights in the test set. The current system is better for the remaining third of flights. In the test set, when PETA performs better, the average and median improvements are 14 minutes and 7 minutes respectively. However, when it under-performs, the average and median deterioration is 7 minutes and 4 minutes respectively. The optimal time frame in terms of relative improvement, with respect to the current system, appears to be between 2 and 6 hours before the departure time.

As of now, we are conducting an investigation to understand why the current system occasionally produces superior predictions. The insights gained from this study may contribute to future enhancements in PETA. Our approach involves a detailed examination of extreme discrepancies, both positive and negative, to deepen our understanding and refine the model. Still, PETA gives more accurate predictions on average at different lookahead times in the six time-bins we

considered prior to departure. Additionally, PETA's improvements over the current system are generally more substantial (as evidenced by the longer tail for PETA's improvements in Fig. 8).

The current version of PETA takes predicted taxi times from ETFMS. If these predictions could be improved with a dedicated model, PETA's ETA predictions could be improved further. The TITOP project within the EUROCONTROL EATIN framework has, in fact, started to develop models for taxi times for a selection of the busiest ECAC airports. A future development could be to incorporate TITOP into PETA. The potential performance improvement, however, is still unknown.

Looking at individual model results, the absolute prediction error of the current system (according to its target) is largest for ATFM delay, then for reactionary delay then for airborne time (see Table 2). Given that the *Knock-on*, *FADE* and *AirborneTime* models each show an approximate improvement over the current system of 30%, this suggests that the most beneficial component of PETA could be *FADE*, then *Knock-on*, then finally the *AirborneTime* model. In principle, we would expect PETA's performance to be best when it uses all three models. However, because the three models are not independent, a comprehensive analysis of the marginal contributions of each model to overall PETA predictions is required to confirm this.

Although our paper does not provide a detailed analysis of how prediction errors vary with the number of flight legs in the rotation, we hypothesise based on preliminary observations that prediction errors may increase as the number of flight legs increases. This potential degradation in performance could be attributed to the accumulation of error at each flight leg, compounding over time. A comprehensive evaluation of these dynamics is warranted, and we recommend that future work focus on conducting such analyses to better understand and mitigate these effects.

Regarding the qualitative results of the live trial, the authors acknowledge that the sample size is relatively small. Conducting live trials and securing participant involvement present numerous logistical challenges. However, despite the limited sample size, the insights gained are valuable and provide a foundation for future studies. We plan to expand these trials in subsequent research to gather more extensive data and further validate our findings.

An issue that has not yet been addressed is how to assess the operational benefit of PETA in the live trial and beyond. This paper shows significant average performance improvement over the current system, yet how does this translate into operational benefit? Given there will be a financial cost to users to implement PETA in their operational systems, will the implementation costs for users be sufficiently outweighed by the cost-savings delivered by PETA? One possible approach would be to monetise the error (accuracy) of ETA predictions, but this is a large project and falls outside of the scope of the current work.

Data Access Statement

The data that support the findings of this study are confidential and not publicly available due to privacy restrictions. Therefore, they cannot be shared for reproducibility purposes.

Contributor Statement

Ramon Dalmau: Conceptualization, Methodology, Formal Analysis, Data Curation, Software, Writing – Original Draft, and Visualization. Aymeric Trzmiel: Conceptualization, Data Curation, Software, Validation, and Writing – Review & Editing. Stephen Kirby: Conceptualization, Supervision, Resources, Project Administration, and Writing – Review & Editing.

Conflict of Interest (COI)

There are no conflict of interest.

References

Ayhan, S., Costas, P., & Samet, H. (2018). Predicting estimated time of arrival for commercial flights. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 33–42. https://doi.org/10.1145/3219819.3219874

Christien, R., Favennec, B., Pasutto, P., Trzmiel, A., Weiss, J., & Zeghal, K. (2021). Predicting arrival delays in the terminal area five hours in advance with machine learning. *14th USA/Europe Air Traffic Management Research and Development Seminar (ATM2021)*.

Dalmau, R. (2024). Probabilistic and explainable tree-based models for rotational reactionary flight delay prediction. *CEAS Aeronautical Journal*. https://doi.org/10.1007/s13272-024-00750-w

Dalmau, R., Genestier, B., Anoraud, C., Choroba, P., & Smith, D. (2021). A machine learning approach to predict the evolution of air traffic flow management delay. *14th USA/Europe Air Traffic Management Research and Development Seminar (ATM2021)*.

EUROCONTROL. (2024). EATIN - EUROCONTROL Air Transport Innovation Network. https://www.eurocontrol.int/project/eatin

Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Ma, Y., Du, W., Chen, J., Zhang, Y., Lv, Y., & Cao, X. (2023). A spatiotemporal neural network model for estimated-time-of-arrival prediction of flights in a terminal maneuvering area. *IEEE Intelligent Transportation Systems Magazine, 15*(1), 285–299. https://doi.org/10.1109/MITS.2021.3132766

Silvestre, J., Santiago, M., Bregón, A., Martínez-Prieto, M. A., & Álvarez-Esteban, P. C. (2021). On the use of deep neural networks to improve flights estimated time of arrival predictions. *Engineering Proceedings, 13*(1), Article 3. https://doi.org/10.3390/engproc2021013003

Strottmann Kern, C., Medeiros, I. P., & Yoneyama, T. (2015). Data-driven aircraft estimated time of arrival prediction. *2015 Annual IEEE Systems Conference (SysCon)*, 727–733. https://doi.org/10.1109/SYSCON.2015.7116837

Wang, G., Liu, K., Chen, H., Wang, Y., & Zhao, Q. (2020). A high-precision method of flight arrival time estimation based on XGBoost. *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, 883–888. https://doi.org/10.1109/ICCASIT50869.2020.9368723

Wang, L., Mao, J., Li, L., Li, X., & Tu, Y. (2023). Prediction of estimated time of arrival for multi-airport systems via "bubble" mechanism. *Transportation Research Part C: Emerging Technologies, 149*, 104065. https://doi.org/10.1016/j.trc.2023.104065

Wang, Z., Liang, M., & Delahaye, D. (2020). Automated data-driven prediction on aircraft estimated time of arrival. *Journal of Air Transport Management, 88*, 101840. https://doi.org/10.1016/j.jairtraman.2020.101840

Zhang, J., Peng, Z., Yang, C., & Wang, B. (2022). Data-driven flight time prediction for arrival aircraft within the terminal area. *IET Intelligent Transport Systems, 16*(2), 263–275. https://doi.org/10.1049/itr2.12142