PETA: Combining Machine Learning Models to Improve Estimated Time of Arrival Predictions

Ramon Dalmau, Aymeric Trzmiel & Stephen Kirby
EGSD/INO/ENG
EUROCONTROL Innovation Hub (EIH)
Brétigny-Sur-Orge, France

***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 (FADE) forecasts the evolution of air traffic flow management delays for regulated flights; the second model (KNOCK-ON) anticipates rotational reactionary delays arising from unrealistic available turn-around times; and the third model was trained to identify systematic discrepancies between reported and actual airborne times. Using a dataset comprised of historical traffic and meteorological data collected from March to June 2023, 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.**

***Index Terms*—machine learning; flight predictability; estimated time of arrival**

I. INTRODUCTION

Estimated time of arrival (ETA) prediction is an important yet challenging issue for the aviation sector. The ETA is important for all aviation stakeholders because it serves as a trigger/input for several air traffic management (ATM) processes throughout the flight. Airlines, airports, air navigation service providers (ANSPs), and the Network Manager (NM) all require accurate ETAs to run flight operations as efficiently as possible. The earlier that data or forecasts are available, the better one can plan ahead. At present, ETA predictions are offered to all stakeholders through the Enhanced Traffic 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, computer assisted slot allocation (CASA) system and runway usage, for instance.

Fig. 1 illustrates the main factors affecting ETA predictions.

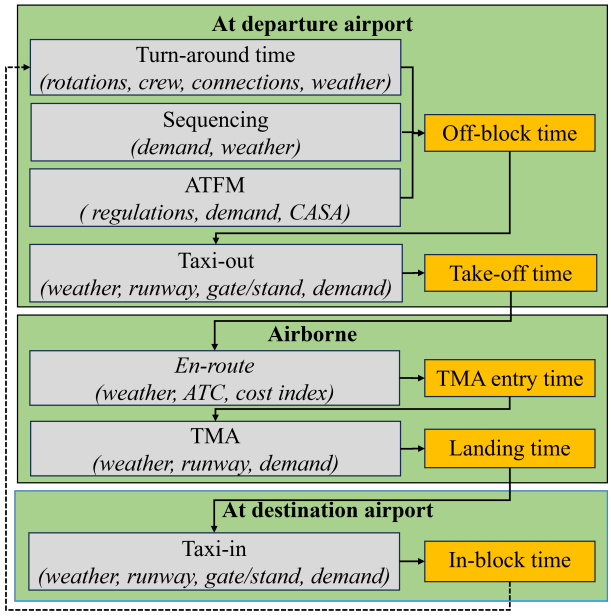


Figure 1: Flight states (green), events (orange), processes (grey) and sources of uncertainty within a process (italic).

In addition, 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.

The results of a data exploratory analysis using current ETFMS predictions revealed that most of the ETA prediction error is due to uncertainty in take-off time. In a point of fact, right after take-off, the ETA prediction (which at that moment corresponds to the airborne time prediction) error follows a Gaussian distribution with a median of approximately 0 min and a relatively low (though not insignificant) dispersion.

Significant efforts have been made in recent years to improve ETA predictions, particularly within the academic and research communities. These efforts have resulted in a plethora of models, most of which are based on machine learning.

The primary focus of these models, however, has been on predicting airborne time, frequently overlooking an important aspect that heavily contributes to the negative operational impact of ETA uncertainty: the variability in take-off times.

Several projects under the EUROCONTROL Air Transport Innovation Network (EATIN) framework[1], including curfew collaborative management [1] and the forecast of ATFM delay evolution (FADE) [2], have made advancements in the prediction of off-block times, and consequently, take-off times.

These models, when combined with a machine learning model aimed at mitigating the aforementioned residual errors in airborne time predictions, could serve as fundamental building blocks for the development of an advanced and data-driven ETA prediction system applicable from the time the flight plan is submitted, several hours before take-off.

The contribution of this paper is twofold. Firstly, it presents the so-called prediction of ETA (PETA) algorithm, which leverages existing machine learning models to improve the ETA predictions. Secondly, it provides a detailed performance evaluation of PETA's predictions by comparing them to the predictions of the current system (i.e., the ETFMS), conducted over a recent three-month period encompassing all intra-European Civil Aviation Conference (ECAC) flights. The evaluation focuses primarily on ETA predictions prior to take-off, as early as 6 h in advance, when uncertainty is at its peak.

Following a state-of-the-art assessment detailing motivations and current situation, the paper describes the methodology and then presents the results and discussion/perspectives.

II. State-of-the-art

A. Definitions and assumptions

In 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[2]. Thus, for the remainder of this document, ETA will refer to the estimated landing time.

Various sources of data are available for stakeholders to calculate ETAs according to their specific needs (e.g., advanced-surface movement guidance and control system for tower/airport, correlated position report). These ETAs serve a variety of functions depending on who owns them. It should be noted, however, that the most accurate (among many) ETA at a specific flight phase is not always shared with all stakeholders, which can prevent the ATM community from fully capitalising on its accuracy. As a result, the focus of this study and the developed models is primarily on improving the ETFMS ETA predictions to make them universally accessible.

B. Motivation

Overall, the ETAs are used by aviation stakeholders to plan and optimise their operations according to their needs: from the airline's perspective, both flight operating centre and airline operating centre need ETAs for overall ground operations, which include aspects such as stand or gate utilisation, ground handling, and staff planning. Furthermore, improved ETA predictions would not only allow for more efficient passenger connections, but would also increase customer satisfaction. From the airports' perspective, inbound ETA is used as a trigger to the airport collaborative decision making (A-CDM) process to optimise flow of passengers and luggage, and for the use of airport resources (e.g., runways, taxiways, gates) and ground services (e.g., ground transportation).

Last but not least, from the ANSPs' perspective, accurate ETA predictions would allow smoother and more efficient arrival delay/traffic management. For instance, it would allow the better planning of arrival traffic, thus reducing or even avoiding arrival bunching in the extended / terminal manoeuvring area (ETMA/TMA) that cause extra delays and fuel consumption (e.g., from holding patterns, level-offs, vectoring).

C. Current situation

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 the 50 busiest airports in ECAC, 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 actual time of arrival (ATA) and the ETA as reported by the ETFMS. Therefore, positive values indicate that the ETFMS prediction was overly optimistic, meaning the flight arrived later than the predicted ETA (i.e., positive $\rightarrow$ delay). The computation was made at two specific and representative events of the flight:

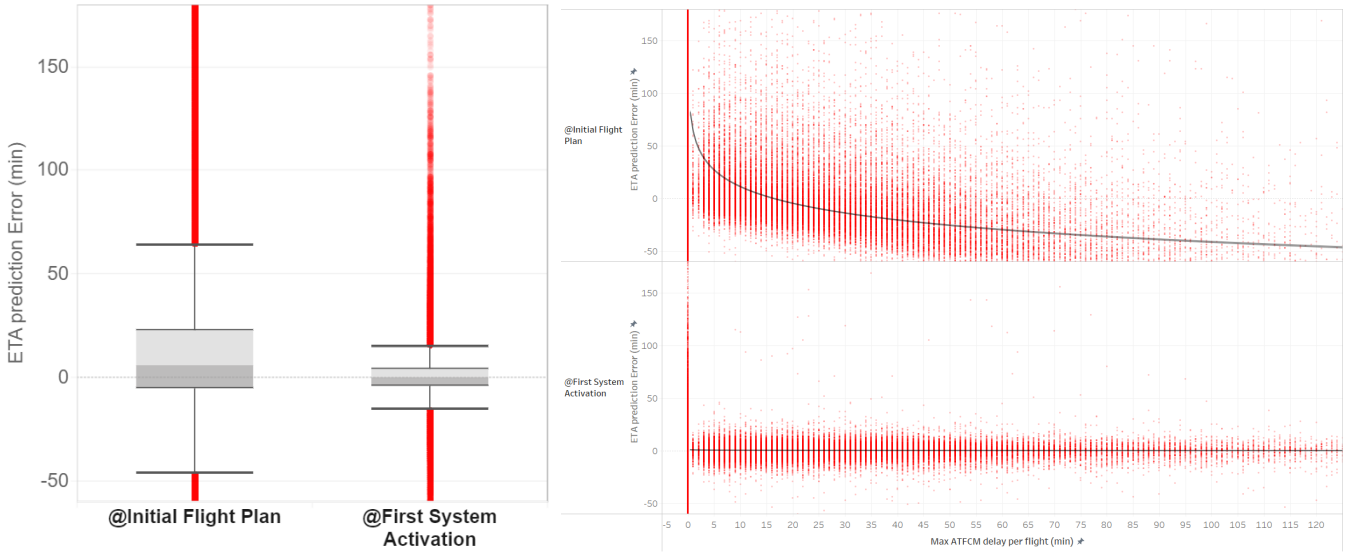
1) at the submission of the initial flight plan (IFP), typically between 3 and 9 h before take-off, and
2) at first system activation (FSA), i.e., when the first ATC message is received, typically right after take-off.

Figure 2a shows that the ETA prediction is more accurate and with less dispersion at FSA (median is 0, with 50% of values within the [-4, 4] min range), when compared to the same values at IFP (median is 6 min, with 90% of values within the [-5, 23] min range). These results highlight that the prediction of the airborne time (used as ETA prediction at FSA) is relatively accurate and that most of the current ETA prediction error is attributable to the take-off time uncertainty.

To further assess the possible cause of ETA uncertainties, Fig. 2b 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. Results indicate that, as expected, the ETA prediction error at FSA is independent from the maximum ATFM delay. Figure 2b also shows, however, that the maximum ATFM delay value has an impact on the ETA prediction error at IFP. The trend exhibits a decrease below 0 as the maximum delay increases, indicating that the prediction of the current system was overly pessimistic. These findings align with expectations, as the ATFM delay typically decreases due to the true revision process of CASA.

[1] https://www.eurocontrol.int/project/eatin

[2] It is worth noting that the in-block time can be calculated by adding the taxi time to the landing time.



(a) ETA prediction error distribution

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

Figure 2: Results of the analysis with current ETA predictions at IFP and FSA events.

In summary, the most significant part of the current ETA prediction errors originates from take-off uncertainties. Airborne uncertainties, though not insignificant, tend to be low.

D. Literature review

The ETA prediction problem has drawn the attention of many researchers in the recent years. 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) [3]–[7]. In this regard, the performance of several machine learning models has also been assessed and compared [8], [9]. Recently, researchers proposed novel methods for accurately predicting ETAs in Beijing TMA [10] and for a multi-airport system [11]. 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, EUROCONTROL developed two machine learning models to improve ground delay predictions: (1) the so-called KNOCK-ON model, which predicts the reactionary delay for non-regulated flights, and (2) FADE, which predicts the evolution of the ATFM delay for regulated flights [2]. KNOCK-ON and FADE, combined with a third machine learning model that tackles the airborne time prediction problem, are the three building blocks of the PETA system that follows.

III. METHODOLOGY

The general 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.

The decision to utilise machine learning over traditional methods in this study is motivated by the complexity of the problem and the extensive amount of data collected by NM. This data was instrumental in training effective models.

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.

A. 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 or airlines, (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. Furthermore, the model takes into account the weather forecast

associated with the estimated off-block time (EOBT) at the departure airport. The primary goal of this model is to improve off-block time predictions for non-regulated flights.

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.

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 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. 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.

B. PETA: combined models

The idea behind the PETA system is illustrated in Algorithm 1. 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. Thereafter, 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 PETA. The PETA is used to estimate the ATT of the next flight in the sequence, and the process is repeated. 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.

C. Data and training

The three machine learning models were trained using ETFMS flight data messages (EFDs) and meteorological aerodrome reports (METARs) to form the dataset, encompassing data from January 1st, 2022, to February 28th, 2023. The raw

Algorithm 1 PETA: propagates predictions along flights operated by the same aircraft registration r to improve ETAs.

1: $\mathcal{M} \leftarrow$ Latest message of all flights operated by r
2: $\mathcal{M} \leftarrow$ Remove cancelled flights from $\mathcal{M}$
3: $\boldsymbol{f} \leftarrow$ Sequence of flights operated by r, sorted by EOBT
4: **for** $i = 1, \ldots, |\boldsymbol{f}|$ **do**
5: **if** $\boldsymbol{f}(i)$ is a terminated flight **then**
6: $\text{PETA}(i) \leftarrow \text{ATA}(i)$
7: **else**
8: **if** $\boldsymbol{f}(i)$ has departed **then**
9: $\text{POBT}(i) \leftarrow \text{AOBT}(i)$
10: $\text{PDLY} \leftarrow \text{AOBT}(i) - \text{EOBT}(i)$
11: **else**
12: $\text{ATT} \leftarrow \text{EOBT}(i) - \text{PETA}(i-1)$
13: $\text{PDLY} \leftarrow \text{KNOCK-ON}(\text{ATT}, ...)$
14: **if** $\boldsymbol{f}(i)$ is regulated **then**
15: $\text{PDLY} \leftarrow \max(\text{PDLY}, \text{FADE}(\text{DLY}(i), ...))$
16: **end if**
17: $\text{POBT}(i) \leftarrow \text{EOBT}(i) + \text{PDLY}$
18: **end if**
19: $\text{PTOT}(i) \leftarrow \text{POBT}(i) + \text{TXOT}(i)$
20: $\text{PETA}(i) \leftarrow \text{PTOT}(i) + \text{AIRBORNETIME}(\text{PDLY}, ...)$
21: **end if**
22: **end for**

(textual) METARs were processed with the open-source library `metafora` (https://github.com/ramondalmau/metafora).

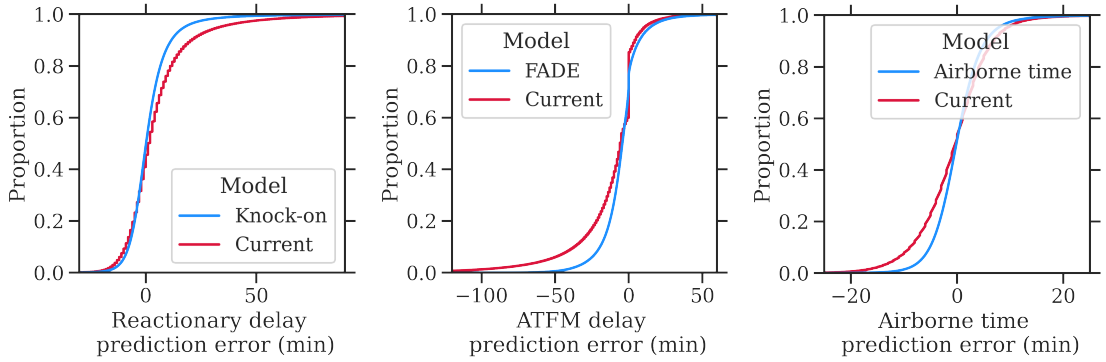
IV. RESULTS

This section presents the outcomes of an evaluation undertaken using historical flight and meteorological data. The dataset used to assess performance spans the period from March 1st to June 30th, 2023, and it includes all intra-ECAC flights operated by aircraft listed in the base of aircraft data, which accounts for 95% of all aircraft types. Each observation in this dataset corresponds to an EFD message sent by a flight.

EFD messages 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. Consequently, the reader should keep in mind that the dataset contains more observations than flights.

Section IV-A provides an overview of the performance metrics for the individual models, each predicting its respective target independently. Section IV-B, on the other hand, delves into the collective performance of the PETA system.

The performance assessment presented in Section IV-A assumes perfect knowledge of weather conditions at the departure and destination airports, i.e., the closest METAR to EOBT and ETA. In contrast, in order to ensure a realistic assessment of the PETA performance in future operations, the performance assessment presented in Section IV-B only considers information available at the prediction time while applying the following rule: when the time difference between



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

Figure 4: Empirical cumulative distribution function of the various prediction errors.

the predicted milestone and the prediction time is <3 h, the latest METAR is used; otherwise, the latest TAF is considered.

A. Individual models

This section thoroughly examines the performance of each of the three models in predicting their respective targets.

Figure 4 shows the (signed) cumulative prediction error distribution of both current and machine learning models. Complementing this figure, Table I presents the key metrics of the absolute prediction error distribution for the machine learning models in comparison to the current predictions. The specific results will be discussed in their respective sections.

TABLE I: ABSOLUTE PREDICTION ERROR DISTRIBUTION METRICS (MIN).

Output	Reactionary delay		ATFM delay		Airborne time	
Model	Current	KNOCK-ON	Current	FADE	Current	Airborne time
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

1) 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 takes in to account these updates.

In terms of absolute off-block time prediction error, Table I 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.

Figure 4a 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 realistic 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.

2) 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 consists of using the current ATFM delay assigned by CASA as the best prediction.

In terms of the absolute ATFM delay prediction error distribution, as shown in Table I, FADE manages to reduce the mean error by approximately 5.5 min (36%). It's 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%).

Figure 4b 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 works diligently 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 4b 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.

3) AIRBORNETIME*:* To assess the predictive power of the model, the prediction error can be calculated 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 I, 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. These findings are further supported by the results of the analysis presented in Section II-C, implying that a significant portion of ETA uncertainty is caused by factors prior to take-off. 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 4c 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.

B. PETA: combined models

The ETA predictions generated by the PETA system will be juxtaposed with those of the current system under identical conditions. It is important to highlight that, in the case of regulated flights, the ETA provided by the current system effectively takes into account the (current) ATFM delay.

The distribution of (signed) ETA prediction errors is depicted in Fig. 5. These errors are calculated as the actual time of arrival (ATA) minus the predicted ETA, as in the previous evaluations. As a result, positive values indicate unexpected delays, while negative values indicate that the flight arrived at the destination airport earlier than predicted.

The results are further enriched by Table II, which presents various metrics related to the distribution of absolute ETA prediction errors. Notably, in contrast to previous evaluations, the results are grouped based on the time to EOBT.

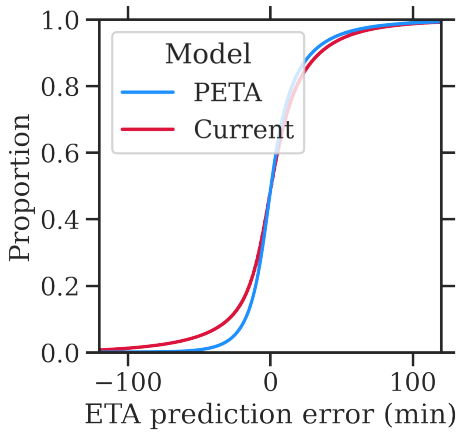


Figure 5: Empirical cumulative distribution function of the ETA prediction error.

Figure 5 closely aligns with the cumulative distributions previously showcased for the individual models. Interestingly, a substantial portion of the overall improvement can be attributed to FADE's capability to forecast regulated flights for which the ATFM delay is expected to decrease in the near future. This particularly benefits the negative side of the

TABLE II: ABSOLUTE ETA PREDICTION ERROR DISTRIBUTION METRICS (MIN) GROUPED BY TIME TO EOBT.

Metric	Mean		Std.		25th Perc.		Median		75th Perc.	
Model	Current	PETA	Current	PETA	Current	PETA	Current	PETA	Current	PETA
≤30 min	14.5	12.1	21.5	20.0	3.9	3.0	8.5	6.7	16.9	13.7
(30 min, 1 h]	15.9	12.7	21.7	19.7	4.3	3.4	9.8	7.6	19.2	14.8
(1 h, 2 h]	19.2	14.3	24.7	20.1	5.0	4.0	11.7	8.8	24.0	17.2
(2 h, 4 h]	26.2	17.6	32.9	22.8	7.0	5.1	16.0	11.2	33.0	21.7
(4 h, 6 h]	30.4	19.9	37.6	25.1	7.6	5.7	18.0	12.7	39.0	24.7
>6 h	23.1	19.6	29.3	26.9	5.7	5.1	13.0	11.5	28.8	23.5

signed ETA prediction error distribution. 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.

Table II demonstrates that an ensemble of machine learning models working collaboratively to improve ETA predictions consistently outperforms existing 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 may not be beneficial outside of this time frame.

Finally, Fig. 6 shows the histogram of differences in absolute ETA prediction error between the current system and PETA. Each observation in this histogram corresponds to one prediction, and the associated value was computed as ABS(ATA – Current system's ETA) - ABS(ATA – PETA). Accordingly, the positive side of the distribution includes the observations in which PETA was better than the current system, in absolute terms, whereas the negative side contains cases in which PETA was worse than the current system.

According to this figure, PETA outperformed the current system in roughly two thirds or the predictions.

Figure 7 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. This figure shows that, on average, PETA provided more accurate ETA predictions than the current

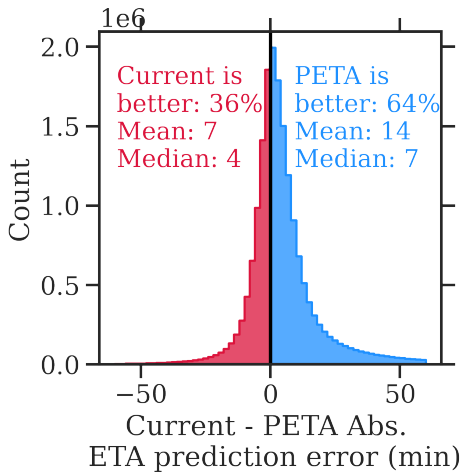


Figure 6: Histogram of the differences between the current and PETA absolute ETA prediction errors. For completeness, the mean and median absolute values of both the blue and red distributions are also included.

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.

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.

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). 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.

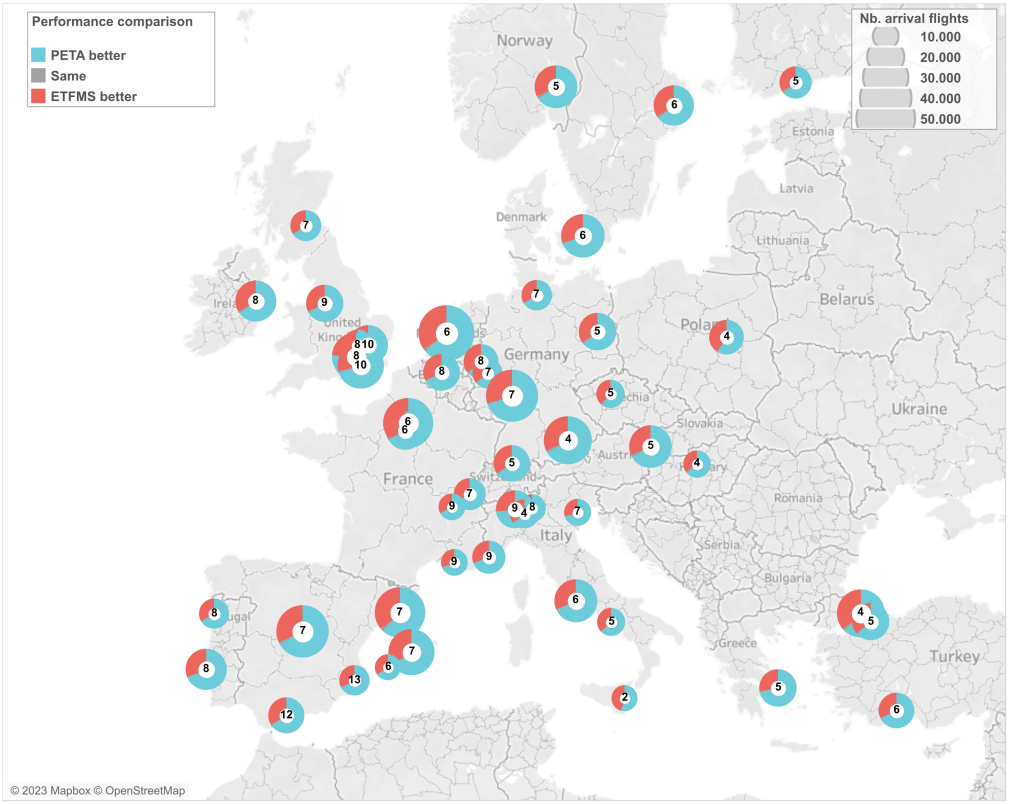


Figure 7: Differences between the current and PETA absolute ETA prediction errors per airport. (Each circle is an airport with its size proportional to the number of arrivals considered.)

V. 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. When PETA 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.

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 than the current system's improvements over PETA (as evidenced by the longer tail for PETA's improvements in Fig. 6).

The latest 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 I). 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 AIRBORNETIMEairborne time model. In principle, we would expect PETA's performance to be best when it uses all three models. However, this has to be confirmed by a comprehensive analysing marginal contributions of each model on overall PETA predictions, the three models not being independent.

The EATIN programme is focused on delivering relatively quick-return operational benefits to users. The next significant step is to make PETA available to a small number of users through an informal live trial, which will be achieved by providing an API for authorised users to access PETA on EUROCONTROL's Cloudera development platform. The TAF is one of the inputs to PETA for creating weather features. PETA can make an ETA prediction for a flight without a TAF, but the API will allow users to provide their own raw TAFs if they wish. Analysis has shown that the accuracy of PETA's

ETA predictions is reduced if no TAF is provided, but the reduction in accuracy (in terms of mean absolute error) is in the tenths of min, so is of minor significance.

An issue that has not yet been addressed is how to assess the operational benefit of PETA in the live atrial 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 monetize the error (accuracy) of ETA predictions, but this is a large project and falls outside of the scope of the current work!

Data Availability 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.

Acknowledgements

This study was proposed by Austrian & ANS CR for the 5th EUROCONTROL Air Transport Innovation Network (EATIN) cycle in December 2022. The project caught the attention of partners that expressed their interest and joined the project. It includes the following partners:

- Airports: Brussels, Prague, Heathrow;
- Airspace Users: Swiss, KLM, Austrian, TAP;
- ANSPs: NATS, ANS CR.

References

[1] R. Dalmau, G. Murgese, Y. De Wandeler, R. Correia, and A. Marsden, "Early Detection of Night Curfew Infringements by Delay Propagation with Neural Networks," in *14th USA Europe Air Traffic Management Research and Development Seminar*, (Virtual Event), 2021.

[2] R. Dalmau, B. Genestier, C. Anoraud, P. Choroba, and D. Smith, "A Machine Learning Approach to Predict the Evolution of Air Traffic Flow Management Delay," in *14th USA Europe Air Traffic Management Research and Development Seminar*, (Virtual Event), 2021.

[3] C. Strottmann Kern, I. P. de Medeiros, and T. Yoneyama, "Data-driven aircraft estimated time of arrival prediction," in *2015 Annual IEEE Systems Conference (SysCon) Proceedings*, (Vancouver, BC), pp. 727–733, 2015.

[4] S. Ayhan, P. Costas, and H. Samet, "Predicting Estimated Time of Arrival for Commercial Flights," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (New York, NY), pp. 33–42, 2018.

[5] Z. Wang, M. Liang, and D. Delahaye, "Automated data-driven prediction on aircraft Estimated Time of Arrival," *Journal of Air Transport Management*, vol. 88, p. 101840, 2020.

[6] G. Wang, K. Liu, H. Chen, Y. Wang, and Q. Zhao, "A High-precision Method of Flight Arrival Time Estimation based on XGBoost," in *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT*, (Weihai, China), pp. 883–888, 2020.

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

[8] J. Silvestre, M. de Santiago, A. Bregon, M. A. Martínez-Prieto, and P. C. Álvarez Esteban, "On the Use of Deep Neural Networks to Improve Flights Estimated Time of Arrival Predictions," *Engineering Proceedings*, vol. 13, no. 1, 2021.

[9] 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.

[10] Y. Ma, W. Du, J. Chen, Y. Zhang, Y. Lv, and X. Cao, "A Spatiotemporal Neural Network Model for Estimated-Time-of-Arrival Prediction of Flights in a Terminal Maneuvering Area," *IEEE Intelligent Transportation Systems Magazine*, vol. 15, no. 1, pp. 285–299, 2023.

[11] L. Wang, J. Mao, L. Li, X. Li, and Y. Tu, "Prediction of estimated time of arrival for multi-airport systems via "Bubble" mechanism," *Transportation Research Part C: Emerging Technologies*, vol. 149, p. 104065, 2023.