Dynamically forecasting airline departure delay probability distributions for individual flights using supervised learning

Beltman, Maarten; Ribeiro, Marta; de Wilde, Jasper; Sun, Junzi

DOI
10.1016/j.jairtraman.2025.102788

Publication date
2025

Document Version
Final published version

Published in
Journal of Air Transport Management

Citation (APA)
Beltman, M., Ribeiro, M., de Wilde, J., & Sun, J. (2025). Dynamically forecasting airline departure delay probability distributions for individual flights using supervised learning. *Journal of Air Transport Management*, *126*, Article 102788. https://doi.org/10.1016/j.jairtraman.2025.102788

Journal of Air Transport Management 126 (2025) 102788

Contents lists available at ScienceDirect

Journal of Air Transport Management

journal homepage: www.elsevier.com/locate/jairtraman

Dynamically forecasting airline departure delay probability distributions for individual flights using supervised learning

Maarten Beltman [a], Marta Ribeiro [a,*], Jasper de Wilde [b], Junzi Sun [a]

[a] Faculty of Aerospace Engineering, Delft University of Technology, Delft, The Netherlands
[b] Koninklijke Luchtvaart Maatschappij N.V., The Netherlands

ARTICLE INFO

Keywords:
Departure Delay
Supervised learning
Random Forest
CatBoost
Deep Neural Networks

ABSTRACT

Punctuality is a key performance indicator for any airline, especially hub-and-spoke airlines, given their focus on short passenger connections. Flights that are delayed at departure need to compensate for lost time whilst airborne. Because fuelling takes place well before scheduled departure, predicted departure delays determine the planned fuel amounts for en-route speed optimization. To prevent unnecessary fuel burn, airlines benefit from highly accurate departure delay predictions. This study aims to extend previous work on airline departure delay forecasting to a dynamic and probabilistic domain, whilst incorporating novel day-of-operations airline information to further minimize prediction errors. Random Forest, CatBoost, and Deep Neural Network models are proposed for a case study on departure flights of a major hub-and-spoke airline from its hub airport between 1 January 2020 and 1 August 2023. The Random Forest model is selected for its probabilistic performance and high accuracy in predicting delays between 5 and 25 min, for which en-route speed optimization has the largest effect. At the 90 min prediction horizon, the model reaches a Mean Absolute Error of 8.46 min and a Root Mean Square Error of 11.91 min. For 76% of flights, the actual delay is within the predicted probability distribution range. Finally, this study puts a strong emphasis on explainability. Flight dispatchers are therefore provided with the main factors impacting the prediction, explaining the context of the flight. The versatility of the model is demonstrated in two shadow runs within the procedures of an international airline, where delays caused by familiar and unfamiliar factors were successfully predicted.

1. Introduction

Punctuality is a key performance indicator for airlines, especially hub-and-spoke airlines, given their focus on short passenger connections. A significant factor that affects punctuality is the departure delay of the flight. The latter refers to the difference between the actual off-block time and scheduled off-block time, where off-block marks the moment when an aircraft begins to push back from the gate or parking position to initiate taxiing for departure. Costly passenger compensation and experienced discomfort are not the only incentives for airlines to minimize such delays. To ensure on-time arrivals, flights that were delayed at departure beyond pre-planned slacks in the scheduled travel time, have to compensate for the lost time whilst airborne, thereby increasing fuel consumption which results in higher costs and emissions.

Required fuel amounts are indicated in a flight's final flight plan, which is usually issued by the airline's flight dispatcher around 90 min before scheduled departure time. Accurate departure delay predictions greatly benefit fuel amount calculations. In case of underpredicted departure delays, insufficient fuel is carried for compensating time, possibly leading to the loss of passenger connections. Alternatively, in the case of overpredicted departure delays, excessive fuel is carried, leading to unnecessary fuel burn due to increased aircraft mass.

Predicting these delays is not trivial, as they stem from diverse causes such as weather conditions, airspace capacity, airport congestion, and airline resource allocation. Conventional approaches often rely on historical averages of flight occurrences, which yield suboptimal outcomes. However, given the dynamic nature of airspace operations and the escalating density of air traffic, historical data frequently fails to accurately reflect present delays.

Research has focused on developing prediction methods that go beyond averaging historical data, resorting to statistical and stochastic methods (Mueller and Chatterji, 2002; Abdel-Aty et al., 2007; Tu et al., 2008). However, existing studies predominantly focus on the determination of a single delay value at a specific moment prior to flight departure. Nevertheless, the accuracy of the prediction models varies considerably over time with the assimilation of new data as departure

* Corresponding author.
E-mail address: m.j.ribeiro@tudelft.nl (M. Ribeiro).

https://doi.org/10.1016/j.jairtraman.2025.102788
Received 19 June 2024; Received in revised form 18 March 2025; Accepted 19 March 2025
Available online 11 April 2025

time approaches. An understanding of the level of uncertainty at a certain point in time, and of how new input values affect the final delay, can provide valuable guidance for decision-making. Additionally, often results are constrained by the lack of airline data (Dalmau et al., 2021). However, airline decisions directly influence flight prioritization and resource allocation, directly shaping the final departure scenario. Finally, there is still wide concern on how to effectively translate the outcomes of these prediction models into practical tools for airline operations.

This study aims to cover the previous research gaps by extending current departure delay prediction models towards a dynamic and probabilistic domain. Thus, forecasting departure delay probability distributions for individual flights at different moments relative to their scheduled departure times. Simultaneously, this study aims to further minimize prediction errors by exploring novel day-of-operations flight characteristics, available from an international airline flight dataset. Finally, given the importance of the decisions that follow from the model predictions, this study places strong emphasis on explainability by outlining the reasoning behind the predictions of the model. Recommendations are made on how to present the outcomes of the model to flight dispatchers.

The remainder of the paper is structured as follows. Related work and the contribution of this study are presented in Section 2. Thereafter, the methodology adhered to in this study is elaborated upon in Section 3. The results of the proposed models are then presented in Section 4, where one of the models is selected. Additionally, it is discussed how to present the results of the model to flight dispatchers. The model validation is performed in Section 5. Subsequently, the results are discussed in Section 6, alongside the results of shadow runs performed with real-world data and conditions at an international airline. Finally, Section 7 concludes this work.

2. Related work

Early related work focused on quantifying departure delays using statistical methods. Mueller and Chatterji (2002) found that departure delay probability distributions were best modelled using Poisson distributions and that arrival delay probability distributions better fitted Normal distributions. Furthermore, departure delays were often considered to be a sum of temporal components. Abdel-Aty et al. (2007) proposed a model building on daily, monthly, and seasonal patterns, and Tu et al. (2008) summed daily and seasonal patterns with a residual. Historical data shows that departure delays are not evenly distributed: the majority of flights experience minor delays whereas only a few flights are delayed more significantly. This unevenness, referred to as *positive* skew, was the motivation for Pérez-Rodríguez et al. (2017) to compare symmetric and asymmetric Bayesian logistic models for predicting flight delays. The skewed nature of the dataset favoured the performance of the latter model.

In recent work on departure delay prediction using supervised learning, the applicability of both tree-based models and neural network models has been studied. For tree-based models, Kalliguddi and Leboulluec (2017) and Khan (Khan et al., 2021), concluded that simple decision trees were outperformed by random forests. This is expected as the later uses an ensemble of decision trees, showing higher accuracy for datasets with high-dimensional feature spaces. At the same time, Manna et al. (2017) showed that random forests, in their turn, were outperformed by boosting models. The latter can be more accurate as, contrary to Random Forest, it trains one tree at a time, each tree correcting the errors of the previous ones. With the aim of evaluating the performance of EUROCONTROL's Enhanced Tactical Flow Management System (ETFMS) against a supervised learning model, Dalmau et al. (2021) proposed a different boosting model: Gradient-Boosted Decision Trees (GBDT). Using a large number of features (over 30), it was found that the existing system is outperformed by the GBDT model, especially for prediction horizons larger than 60 min. GBDT has the advantage of each tree in the gradient boosting correcting the errors of its predecessor. Vorage (2021) extended the departure delay prediction to the probabilistic domain. Random Forests and Mixture Density Networks were proposed to generate probability density functions for individual flights from Amsterdam Airport Schiphol. From these distributions, the probability that a forecast delay is accurate within some time-error interval could be computed. Later, this approach was used by Zoutendijk and Mitici (2021), constructing similar models to predict departure delays using Rotterdam Airport flight data, reaching a Mean Absolute Error (MAE) of around 12.5 min.

Sun et al. (2022) aimed to predict airline delays from a network perspective, testing the applicability of several neural networks including a Dynamic Spatial–Temporal Graph Attention (DST-GAT) network and a Long Short-Term Memory (LSTM) network. Whilst the network architectures differed significantly, the outcomes for both models were comparable, with Root Mean Square Error (RMSE) values between 5–10 min, differing per airport in the network. DST-GAT and LSTM networks offer the possibility of greater accuracy with large amounts of data with long-range dependencies. Finally, Birolini and Jacquillat (2023) collaborated with European airline Vueling, comparing the performance of Linear Regression, Random Forest, and Extreme Gradient boosting models for predicting the airline's flight delays. For each flight, only the primary flight delay was considered, eliminating the effect of precedent flights. Taking into account airline-specific information, such as crew rosters and aircraft availability, the Extreme Gradient boosting model reached an MAE of around 7 min, outperforming the other models.

Most reviewed literature considered temporal features, flight schedule features, and weather features. Among others, Sternberg et al. (2016) demonstrated that including weather features benefits the model performance. Other novel features such as flight de-icing status (Dalmau et al., 2021) and take-off runway (Khan et al., 2021; Alonso and Loureiro, 2015) were proposed. Moreover, Yu et al. (2019) especially focused on short-term features, including the boarding option (jet bridge or bus), closing time of cargo doors and passenger doors and the time between check-in, boarding, and gate closure. Only two papers considered passenger connection information, aggregating numbers per flight destination. Furthermore, only a few studies had access to detailed airline data. Finally, although probabilistic departure delay forecasts for individual flights were proposed (Vorage, 2021; Zoutendijk and Mitici, 2021), it has not yet been investigated how such probabilistic forecasts change over time.

In resume, this paper intends to advance the state-of-the-art and cover the previously mentioned research gaps by:

1. Considering the number of planned connection passengers for every unique inbound-to-outbound flight combination: Only two papers considered passenger connection information. Ciruelos et al. (2015) assumed monthly connecting passenger percentages, thereby not specifying the numbers per individual flight. Sismanidou et al. (2022) had access to real passenger itineraries from a Marketing Information Data Tapes (MIDT) dataset, but used this data to determine "*a proportion of connecting passengers for a specific itinerary by a specific air carrier*", therefore also averaging the connecting passenger numbers, missing out on the opportunity to use the flight-specific passenger connection data for predicting the departure delays.
2. Considering novel features of the day-of-operation of airlines: Only few studies had access to detailed airline data, through research partnerships with Peach and Vueling (Birolini and Jacquillat, 2023; Horiguchi et al., 2017). Despite these partnerships, the studies refrain from proposing detailed day-of-operations features, but instead hold on to mostly booking and schedule information. Thus, there remains a research gap for the effect of day-of-operation features, such as last-minute airframe assignments, Target Start-up Approval Time (TSAT) changes, and airport delay levels.

3. Analysing how probabilistic departure delay forecasts for individual flights change over time relative to scheduled departure, given the availability of new data. Several papers compared the performance of the model at multiple moments relative to scheduled departure times (Dalmau et al., 2021; Sun et al., 2022; Rebollo and Balakrishnan, 2014; Choi et al., 2016; Gopalakrishnan and Balakrishnan, 2017; Schösser and Schönberger, 2022). However, none provides an indication of the probability distribution of the determined delay value. Although probabilistic departure delay forecasts for individual flights were already proposed (Vorage, 2021; Zoutendijk and Mitici, 2021), it has not yet been investigated how such probabilistic forecasts change over time. The concept of dynamic probabilistic forecasts was previously used by Felder et al. (2010) in the domain of dynamic wind power forecasting. No such research has yet been performed for the prediction of departure delays of individual flights.

3. Methodology

This section aims to describe the methodology adopted for this study. First, the case study is introduced in Section 3.1. Thereafter, data preprocessing is discussed in Section 3.2. The model development process is then elaborated upon in Section 3.3. Finally, the feature engineering processes are explained in Section 3.4 and the model training and result processing are discussed in Section 3.5.

3.1. Case study

Several departure delay prediction methods are proposed and tested through a case study on data provided by a major international airline. The case study was conducted for the period between 1 January 2020 and 1 August 2023. It should be noted that this period includes both the COVID period and the period just after, where logistical problems resulted in above-average departure delays. Most of the logistical problems were solved from November 2022 onwards, which translated into traffic and delay levels almost back at pre-COVID standards, still impacted by seasonality. Including the data of these periods with significantly different dynamics turns out to be beneficial as it allows the model to be trained on a wider variety of historical data. The presence of more flights towards the extremes of the departure delay spectrum simultaneously improves the sampling practices.

The case study involves the use of supervised learning algorithms to improve on an existing statistical model, currently in use at the airline. For predicting departure delays, the statistical model solely uses historical data of other flights within the same flight series. The supervised learning algorithms are trained to draw patterns based on all historical flights. Next to further minimizing the prediction error, the case study also aims to better explain the predicted delays by predicting departure delays as probability distributions over time and by explaining what inputs contribute to the predicted output. The study is split into multiple stages, illustrated in Fig. 1. These stages are elaborated upon in more detail in Section 3.2 to Section 3.5.

3.2. Data preprocessing

Through the collaboration with the airline, detailed flight data and basic traffic data of the hub airport are available. Additionally, historical weather data is available from the Iowa State University Environmental Mesonet (Iowa State University, 2023). The following describes how these three datasets were processed:

- **Airline Data:** After eliminating cancelled flights, flights with unpredictable delays,[1] ferry flights, test flights and flights without recorded departure delays, the airline dataset consists of 286,000 unique flights. For each flight, (anonymized) passenger information and flight events, including TSAT updates and Estimated Time of Arrival (ETA) predictions of inbound flights, are known. Additionally, connection times and passenger counts of any flight arriving at the hub airport are available for over 90% of the outbound flights.
- **Airport Data:** After similar initial data cleaning as the airline data, the airport dataset consists of 562,000 flights (including those of the considered airline). From the available information on origin–destination pairs, scheduled and actual departure times, and flight numbers, the number of flights and average delays at the hub airport can be determined for any time interval within the case study period.
- **METeorological Aerodrome Report (METAR) Data:** Available from Iowa State University Environmental Mesonet, weather reports with an update frequency of 30 min are obtained for the hub airport, including air temperature, dew point, humidity, wind direction, wind speed and gust, precipitation, visibility, cloud coverage, and cloud height information (Iowa State University, 2023).

The blue bins in the left of Fig. 2 represent the number of flights for a specific departure delay value. The distribution of the bins confirms that the departure delay data for this study is positively skewed. Therefore, the data has to be balanced through undersampling. The objective of undersampling is to reduce the model's bias towards predicting delay values near zero, which is the majority class. The hypothesis is that decreasing the prevalence of these smaller delays in the training data will improve the model's accuracy in predicting delay values within the minority class of larger delay values. Undersampling was performed by selecting a maximum number of flights for each bin. This maximum number is based on the number of flights in a select bin. Selecting the correct bin to undersample the excess data introduces a trade-off between two metrics: R2 and RMSE, illustrated on the right of Fig. 2. No data sampling yields relatively low RMSE values because the model is overfitted to the part of the departure delay spectrum with most data points. As a result, the errors for flights with uncommon departure delays are relatively high, causing the fit of the model to remain low.

To guarantee a balance of global fit and optimal model performance, it is chosen to select a sampling strategy whereby the number of flights is limited to the number of flights in the 15 min bin. This yields the purple distribution on the left of Fig. 2. For this sampling strategy, the R2 value is above its trend and the RMSE value is below its trend, where additional importance is given to the RMSE value as it increases relatively faster than the R2 value. The sampling process is performed at random, because other methods such as cherry-picking may introduce biases to the model.

Finally, to improve the accuracy and reliability of the model, outliers are removed from the dataset. Because of the positive skew in the departure delay distribution, visible on the left of Fig. 2, only values towards the extreme positive end are removed, since these values lie much farther from the median value than those towards the negative end. Only 1% of outlier data is removed, thereby preventing the potential loss of valuable data patterns. After outlier removal, the remaining departure delay spectrum ranges from −20 to +97 min.

3.3. Model development

From existing studies, it is apparent that simple decision trees are nearly always outperformed by random forests and boosting models (Kalliguddi and Leboulluec, 2017; Khan et al., 2021; Manna et al., 2017). Additionally, more complex neural network structures would outperform simpler structures as long as the available training dataset is of sufficient size (Ye et al., 2020; Sun et al., 2022; Birolini and Jacquillat, 2023). When insufficient training data is available, or the predicted

[1] Delays are registered using International Air Transport Association (IATA) Delay Codes (EUROCONTROL, 2023), a classified selection of which is deemed unpredictable, e.g., technical issues.

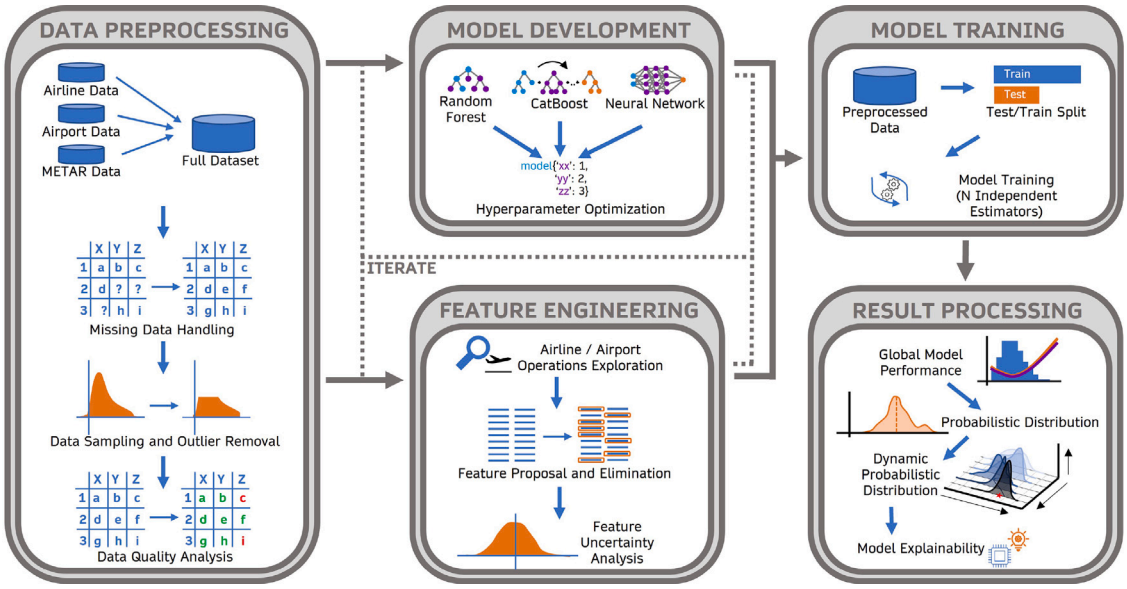

Fig. 1. Methodology for Dynamically Forecasting Airline Departure Delay Probability Distributions.

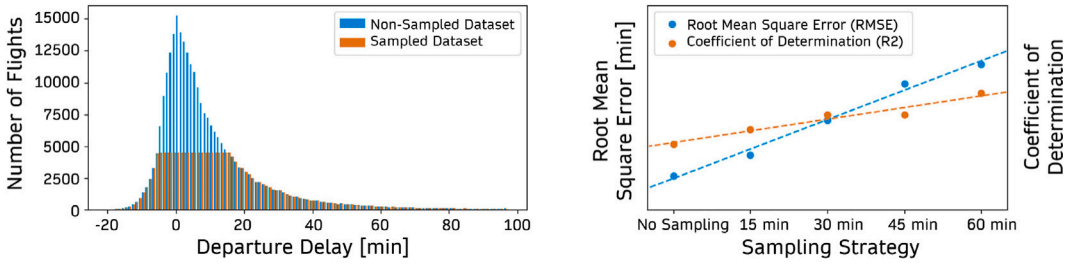

Fig. 2. Non-Sampled and Sampled Departure Delay Distributions (left) and Sampling Strategy Trade-Off (right).

Table 1
Hyperparameters of the models employed in this study.

(a) Random forest model.

Hyperparameter	Value
nr_estimators	1000
max_features	4
max_depth	10
min_samples_split	4
min_samples_leaf	2

(b) CatBoost model.

Hyperparameter	Value (CatBoostTH)	Value (CatBoostPR)
iterations	400	10
learning_rate	0.02	0.5
depth	10	10
posterior_sampling	*True*	*True*
random_state	*Random*	*Random*

(c) Deep neural network model.

Hyperparameter	Value
nr_input_neurons	15
nr_output_neurons	1
nr_hidden_neurons	16
nr_hidden_layers	4
dropout	0
leakyrelu_negative_slope	0.1
n_epochs	1000
batch_size	2048
lr_initial	1e−4
lr_increase	1.2
lr_decrease	1.2
lr_stop	1e−10

processes are too random, simpler neural networks or tree-based models may outperform more complex models. This work directly compares the usage of three different models: Random Forest, CatBoost, and Deep Neural Network models.

First, Random Forest was chosen for its simplicity and high level of explainability, as the decision-making process at each tree can be easily traced to the final outcome. Next, in theory, CatBoost promises an increase in accuracy when compared to Random Forests due to its boosting mechanics. As a member of the family of GBDT machine learning ensemble techniques, CatBoost is particularly advantageous due to its exceptional ability to handle categorical data (Hancock and Khoshgoftaar, 2020). Finally, the utilization of a Deep Neural Network allows to explore whether neural-based methods could be a viable option for this model. These have demonstrated superior performance over boosting methods in handling complex data structures, though this often comes at the expense of longer processing times and the need for large datasets. These three models are further developed and are elaborated upon in Sections 3.3.1, 3.3.2, and 3.3.3. The hyperparameters for all models are shown in Table 1. The values used align with the best practices recommended in previous studies. Multiple different hyperparameters were tested, without any significant improvements in performance.

The delays are predicted at six prediction moments: 90, 75, 60, 45, 30, and 15 min before scheduled departure time. To prevent the models from falsely propagating errors and uncertainty from one prediction moment to another, separate models are trained for each prediction

moment. For all models, hyperparameter optimization is performed empirically, as the effect of all parameters was understood through numerous model iterations in the test phase.

3.3.1. Random forest

Given its computational efficiency, robustness to outliers, and interpretability, random forests have been a popular choice for modelling stochastic processes (Kalliguddi and Leboulluec, 2017; Sun et al., 2022; Birolini and Jacquillat, 2023; Rebollo and Balakrishnan, 2014). The random forest is an ensemble method, a forest of decision trees that serve as independent predictors. Furthermore, the method relies on the concept of bagging (bootstrap-aggregating), meaning that sub-samples of the dataset are used to construct unique decision trees and that, for regression problems, the final prediction is the mean of all individual results. For a probabilistic approach, a probability distribution can be created using all individual tree results (Vorage, 2021; Zoutendijk and Mitici, 2021).

To allow for generating detailed probability distributions, the number of decision trees (*nr_estimators*) is set to 1000. The maximum number of features for the model (*max_features*) is set to 4, following the binary logarithm of the number of features. The maximum model depth (*max_depth*) is set to 10 to prevent overfitting as a result of extremely large trees. The minimum number of samples for splits (*min_samples_split*) and leaves (*min_samples_leaf*) is set to 4 and 2 respectively, because larger values may yield too simple decision trees. The hyperparameters are summarized in Table 1a.

3.3.2. CatBoost

CatBoost (Prokhorenkova et al., 2018) is an open-source gradient boosting library that allows for efficient and fast predictions. The model treats data sequentially to prevent data leakage. The use of symmetric trees makes weaker learners for the boosting process, resulting in faster computation times. Moreover, the underlying boosting scheme of CatBoost prevents overfitting and eases hyperparameter tuning (CatBoost, 2023). CatBoost is inherently less sensitive to data imbalance. CatBoost models require M iterations to reach the final prediction. To allow probabilistic modelling, a total of N independent CatBoost models are created, where N is set to 1000 to match the number of estimators in the Random Forest model.

Two different CatBoost models are proposed, one with a smooth iteration scheme ($iterations$ = 400 and $learning\_rate$ = 0.02), in the remainder referred to as *CatBoostTH* for its theoretical application, and one with a rougher iteration scheme ($iterations$ = 10 and $learning\_rate$ = 0.5), in the remainder referred to as *CatBoostPR*. A larger learning rate typically results in bigger differences between the predictions of consecutive trees. This often leads to greater variance in the final prediction compared to a model with a lower learning rate. The increased variance in the final predictions may provide a more reliable measure of the uncertainty in the model output.

For both models, the model depth ($depth$) is set to 10, complying with that of the Random Forest model, allowing for their results to be compared directly. Note that the value of the model depth is a balanced trade-off, a deeper tree can fit the training data better, but it can also lead to overfitting. The latter is especially a problem with gradient boosting methods. A max depth of 10 was empirically found to be a good value. During training of the model, different depth values were compared. None of the other depths tested resulted in significant differences in performance.

Finally, posterior sampling ($posterior\_sampling$) is enabled to *"obtain uncertainty predictions with good theoretical properties"* (Prokhorenkova et al., 2018). To ensure that all N independent models are unique, the unique state ($random\_state$) is set randomly for every model estimator. The hyperparameters are summarized in Table 1b.

3.3.3. Deep neural network

Neural networks are machine learning models consisting of interconnected nodes that are activated by activation functions. These models can handle nonlinear feature relationships and are trained using error back-propagation, a feedback loop that tunes the internal model parameters to achieve the optimal performance (Svozil et al., 1997). For adequate training, a substantial amount (10^5 to 10^6) of historical flights is required to achieve meaningful results (Thiagarajan et al., 2017). Compared to the other models, neural networks present more challenges in terms of model explainability.

For the neural network structure, the number of input neurons ($nr\_input\_neurons$) is the number of input features. Since the departure delay is the only output, the number of output neurons ($nr\_output\_neurons$) is equal to 1. There is more flexibility in determining the number of hidden neurons ($nr\_hidden\_neurons$), which is set to 16 for the most optimal result. Controlling the depth of the model, the number of hidden layers (nr_hidden_layers), is set to 4, to prevent an overly complex model from forming. Since preliminary results yielded comparable results for in-sample and out-of-sample data, the model was not overfit. Therefore, dropout ($dropout$) is not required and is set to 0. Finally, for the Leaky ReLU (Rectified Linear Unit) activation functions in each of the layers of the model, the negative slope ($leakyrelu\_negative\_slope$) is set to 0.1, to allow for negative inputs.

The hyperparameters are summarized in Table 1c. The number of epochs (n_epochs) is set to 1000, to facilitate enough learning iterations. For the regression algorithm, the MSE loss function is used. Batch sizes ($batch\_size$) of 2048 datapoints are used, resulting in sufficient batches considering that the full dataset is two orders of magnitude larger. An adaptive learning rate is used to ease the search for global optima. The initial learning rate ($lr\_initial$) is set to 1e-4. After every epoch, the learning rate is increased by an increase factor ($lr\_increase$) of 1.2 if the current epoch prediction is better than the current-best prediction and decreased by a decrease factor ($lr\_decrease$) of 1.2 otherwise. When the learning rate becomes smaller than the stop criteria learning rate (lr_stop) of 1e-10, the model is deemed to have converged. These parameter values were found by empirically tuning baseline literature values. This guarantees the model's ability to generalize whilst keeping training times acceptable.

3.4. Features

The selected features are presented in Section 3.4.1 and their correlation is discussed in Section 3.4.2. Finally, the potential uncertainty related to some of the features is evaluated in Section 3.4.3.

3.4.1. Selected features

The selected features from Table 2 are described in more detail below. Note that for the Deep Neural Network, there are trigonometric variations to the month of the year and hour of the day features.

- **Month of Year:** The numeric.[2] month of the flight, following from the Scheduled Off-Blocks Time (SOBT)[3]
- **Hour of Day:** The departure hour of the flight, following from the SOBT3.
- **Passenger Load Factor:** The number of passengers booked compared to the number of seats available on the aircraft. For a fully booked flight, the passenger load factor is 1. Taking a ratio rather than an absolute number of passengers allows to consider flights operated on aircraft with different seating capacities in the same model.
- **Baggage Load Factor:** The number of booked pieces of baggage relative to the number of booked passengers. Similarly to the

[2] e.g. January → 1, February → 2, etc.
[3] In Coordinated Universal Time (UTC).

Table 2
Selected features after feature elimination.

Feature name	Unit	Numeric	Dynamic	Example
Month of Year	[–]	✓		*4*
Hour of Day	[–]	✓		*13*
Passenger Load Factor	[–]	✓		*0.73*
Baggage Load Factor	[–]	✓		*1.14*
Transfer Passenger Percentage	[–]	✓		*67*
Number of Passengers Reduced Mobility	[–]	✓		*2*
Total Passengers Day in Membership Program	[–]	✓		*35000*
Median Delay of Flight Number	[min]	✓		*7*
Effective Delay Previous Flight	[min]	✓		*12*
Current Number of Flights at Hub Airport[a]	[–]	✓	✓	*20*
Current Average Delays at Hub Airport[a]	[min]	✓	✓	*32*
Current TSAT Delay[b]	[min]	✓	✓	*4*
Last Aircraft Tail Swap[b]	[min]	✓	✓	*1500*
Wind Speed Longitudinal Direction	[kts]	✓		*8.32*
Wind Speed Latitudinal Direction	[kts]	✓		*−4.25*

[a] In the 30 min interval before prediction moment.
[b] At prediction moment.

passenger load factor, taking a ratio rather than an absolute number of baggage pieces allows to consider flights operated on aircraft with different seating capacities in the same model.

- **Transfer Passenger Percentage:** The percentage of passengers booked for an outbound flight that connects from any inbound flight at the hub airport.
- **Number of Passengers Reduced Mobility:** The number of booked passengers with wheelchair assistance.
- **Total Passengers Day in Membership Program:** The total number of daily passengers subscribed to the membership program.
- **Median Delay of Flight Number:** The median departure delay of all flights for a given flight series. For flight series with fewer than 25 recordings, the median delay calculation is considered to be too random because of the small number of data points. For these uncommon flights, a zero median delay is registered.

- **Effective Delay Previous Flight:** The effective arrival delay of the previous flight that propagates to the outbound flight for the same aircraft. Airlines incorporate slack times into their schedules to mitigate potential delays (Birolini and Jacquillat, 2023; Lan et al., 2006). For inbound flight i and outbound flight j, operated by airline k using aircraft type l at airport m, the Scheduled Turn-Around Time ($STAT_{i,j}$) is the time difference between the Scheduled Off-Blocks Time of the outbound flight ($SOBT_j$) and the Scheduled In-Blocks Time of the inbound flight ($SIBT_i$), see Eq. (1). The slack time ($\rho_{i,j,k,l,m}$) is the time difference between the Minimum Turn-Around Time[4] ($MTAT_{k,l,m}$) and the Scheduled Turn-Around Time, see Eq. (2). MTAT values may differ per airline, aircraft type, and airport. The arrival delay of the inbound flight (δ_{arr_i}) is the time difference between the Actual In-Blocks Time ($AIBT_i$) and Scheduled In-Blocks Time of the inbound flight, see Eq. (3). Finally, the effective arrival delay of the inbound flight is the arrival delay minus the slack time, see Eq. (4). Since negative slack times do not exist, the effective arrival delay can never exceed the arrival delay itself. All negative effective arrival delays are set to 0 as there is enough available time to turn around the aircraft, regardless of how early the previous flight arrived. Fig. 3 illustrates the effect of slack times on effective arrival delays.

$$STAT_{i,j} = SOBT_j - SIBT_i \tag{1}$$

$$\rho_{i,j,k,l,m} = STAT_{i,j} - MTAT_{k,l,m} \tag{2}$$

$$\delta_{arr_i} = AIBT_i - SIBT_i \tag{3}$$

$$\delta_{arr,eff_{i,j,k,l,m}} = \delta_{arr_i} - \rho_{i,j,k,l,m} \tag{4}$$

- **Current Number of Flights at Hub Airport:** The total number of flights departing from the hub airport in a 30 min time interval before the prediction moment. This feature is dynamic because for each prediction moment, the 30 min time interval is different, possibly resulting in a different total number of flights.
- **Current Average Delays at Hub Airport:** The average delay of all flights departing from the hub airport in a 30 min time interval before the prediction moment. This feature is dynamic because for each prediction moment, the 30 min time interval is different, possibly resulting in a different average delay. Mathematically, for every prediction moment t, the average departure delay (δ_{dep,avg_t}) is the sum of individual departure delays (δ_{dep_j}), divided by the total number of flights in the 30 min time interval (N_t), see Eq. (5).

$$\delta_{dep,avg_t} = \frac{\Sigma_{j=1}^{N_t} \delta_{dep_j}}{N_t} \tag{5}$$

- **Current TSAT Delay:** The latest available Target Start-up Approval Time delay update. This delay, imposed by Air Traffic Control, is dynamic because it may update between prediction moments.
- **Last Aircraft Tail Swap** The time difference between the last aircraft tail swap (new airframe allocation) and scheduled departure time. This feature is dynamic because tail swaps may occur between prediction moments.
- **Wind Speed Longitudinal/Latitudinal Direction:** The longitudinal (East-West) and latitudinal (North-South) components of the wind speed at the departure airport. Including the wind direction and wind speed by themselves may train the model to believe some wind direction would favour delays, even if the wind speed is almost zero. This can be avoided by combining wind speed (V_w) and wind direction (Γ_w) into longitudinal and latitudinal wind components, as presented in Eqs. (6) and (7).

$$V_{w_{lon}} = -V_w \cdot \cos(\Gamma_w - 90°) \tag{6}$$

$$V_{w_{lat}} = V_w \cdot \sin(\Gamma_w - 90°) \tag{7}$$

3.4.2. Feature correlation

The statistical correlation between all features at the 90 min prediction horizon is presented in Fig. 4. The feature describing the

[4] Airline-issued times that indicate the minimum number of minutes required between arrival and departure of two consecutive flights.

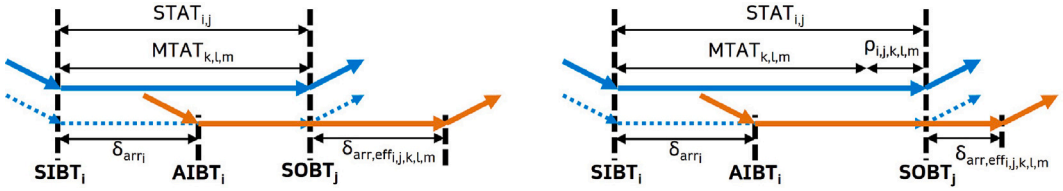

Fig. 3. Aircraft Turn-Around without Slack Times (left) and with Slack Times (right).

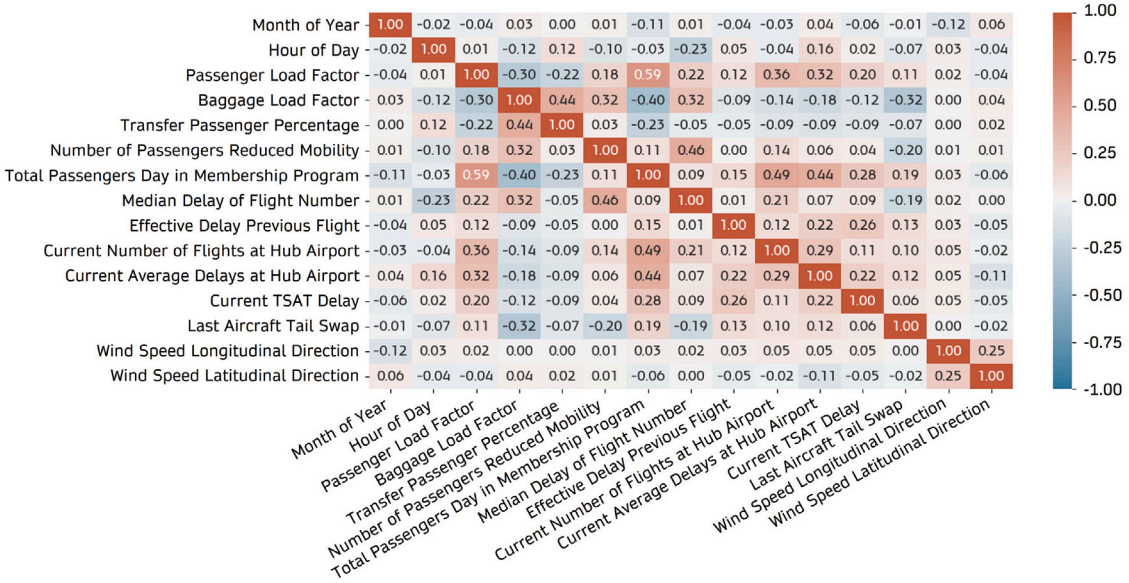

Fig. 4. Feature Correlation Matrix for the 90-Minute Prediction Horizon.

total number of daily passengers in the membership program shows relatively strong correlations with three other features: the passenger load factor (0.59), the current number of flights at hub airport (0.49), and the current average delays at hub airport (0.44). The latter two correlations follow from the coupling of traffic levels and associated delays. The former correlation follows from the fact that during busy periods, the number of daily passengers increases faster than the number of flights, thus resulting in higher passenger load factors. This also explains the negative correlation with the baggage load factor (−0.40), as it appears that for busy periods, the number of passengers increases faster than the pieces of baggage that are carried along.

Furthermore, the median delay in flight number shows a relatively strong correlation with the number of passengers with reduced mobility. Data reveals that the 500 flights with the highest number of passengers with reduced mobility were operated under 11 unique flight numbers only. The correlation is evident because the number of passengers with reduced mobility heavily impacts the turnaround process and potential departure delays.

The feature correlations between the non-dynamic features remain the same at the 15 min prediction moment. The correlations between the dynamic features (e.g., current average delays at hub airport and current TSAT delay) become significantly larger. This can be explained by the fact that the updated dynamic features are closer to the actual values. Finally, for the trigonometric features used for the Deep Neural Network, the cosine of the hour of day has a relatively strong negative correlation (−0.49) with the current number of flights at hub airport. This makes sense as the cosine value of the hour of day is high for the early morning and late evening, but gradually decreases for the middle of the day. The number of scheduled flights develops in the opposite manner.

3.4.3. Feature uncertainty

For flights yet to be predicted, not all input features may be exactly known at each of the different prediction moments. Since the historical training dataset consists of *actual* values, the use of predicted input values may introduce noise and biases. To guarantee the accuracy of the model, such potential uncertainty needs to be evaluated. Most of the features mentioned in Table 2 are either constant throughout the prediction horizon of a flight (e.g., month of year, hour of day, median delay of flight number), can be exactly computed for all prediction moments (e.g., current number of flights, delays, and TSAT delays) or undergo only very minor changes over the prediction horizon of a flight (e.g., passenger and baggage numbers). For a check on a subset of the flight data, large differences in actual and booking passenger and baggage numbers were observed only very rarely, leading to the assumption of using constant passenger and baggage numbers during the prediction horizon. For the effective delay of previous flights and the wind speed features, however, the uncertainty is higher.

Effective Delay Previous Flight Uncertainty: Flights operated by the airline on single-aisle aircraft have minimum turn-around times of 35–50 min and thus may still be airborne 90 min prior to departure of the outbound flight. Flights operated on twin-aisle aircraft have minimum turn-around times of over 120 min. However, these flights may experience delays on the inbound flight, impeding the aircraft from being at the gate 90 min prior to departure of the outbound flight. Throughout the studied period, 42% of flights arrived more than 90 min before departure of the outbound flight; for these flights, there is no feature uncertainty. For the remaining flights, the effective delay of the previous flight can be computed from the ETA of the previous flight.

For the flights that were not yet at the hub airport 90 min prior to the outbound flight, the error distributions between actual and

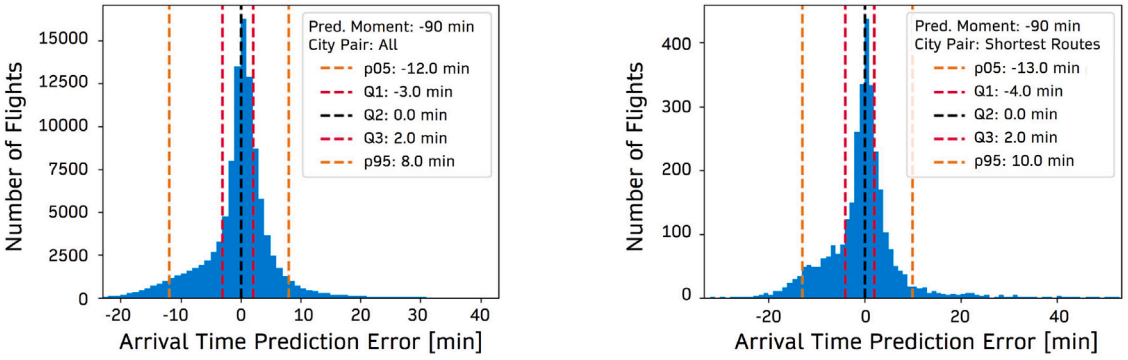

Fig. 5. Previous Flight Arrival Time Prediction Error Distributions for All Flights (left) and the 3 Shortest Flights (right).

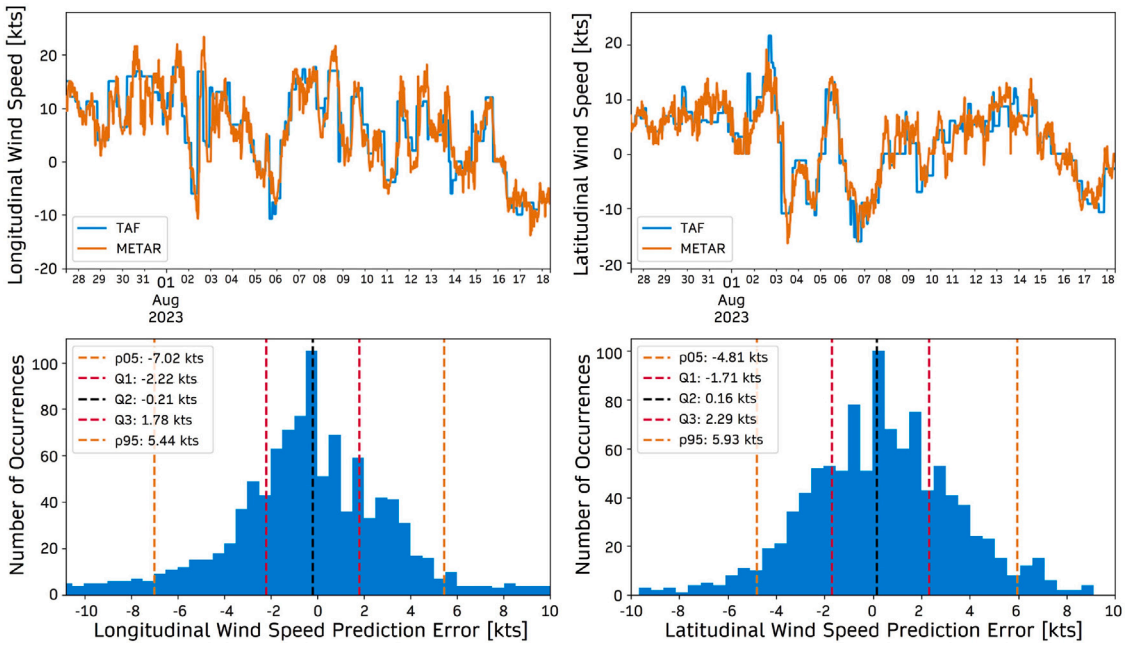

Fig. 6. Longitudinal and Latitudinal Wind Speed Developments (upper left, upper right) and Prediction Error Distributions (lower left, lower right).

predicted arrival times are presented in Fig. 5, differentiating the total set of flights and that of the three shortest routes that the airline operates. Inbound flights on these three routes are most likely to still be on-ground at the origin airport 90 min before departure of the outbound flight from the Hub Airport, affecting the uncertainty of ETA predictions more than whilst airborne. As a result of using the predicted ETA, noise is added to the model. The Inter-Quartile Range (IQR) is at most 6 min. Although the ETA prediction model seems slightly conservative, predicting a large number of flights to arrive later than they did in reality, there are no clear biases. Prior to this analysis, one might hypothesize that the error in this variable would be greater for shorter inbound flights that have not yet departed. In such cases, the final delay prediction would need to account for both departure and en-route delays. Conversely, for longer inbound flights, at −90 min before the departure of the outbound flight, these flights are already in the enroute phase, and the departure delay is already known. Nevertheless, this proves that this is not the case.

Wind Speed Longitudinal/Lateral Direction Uncertainty: The wind speed features contain uncertainty because only METAR data is available to train the model. For flights yet to be predicted, the METAR at the scheduled departure time is still unknown at the prediction moment. Therefore, Terminal Aerodrome Forecasts (TAF) are used for predicting new flights instead. The unavailability of open-source TAF reports restricts the training of the model on TAF data.

To validate the use of METAR data for the hub airport, for 22 days (1056 recordings) with various wind conditions,[5] the predicted wind (TAF) is compared to the actual wind (METAR). The developments over time for the longitudinal and lateral wind components are presented in the upper figures of Fig. 6. In general, the TAF is capable of adequately predicting long-term weather developments. Some of the prediction errors can be attributed to the TAF reports describing how the weather is expected to change over a longer period of time. This is represented by the horizontal sections on the blue lines in the graphs. METARs have higher update frequencies and vary more heavily, leading to other small prediction errors. The bottom figures in Fig. 6 present the wind speed prediction error distributions for longitudinal and lateral directions. It can be concluded that the IQR for both directions never exceeds 4 knots. Similar to the feature describing the effective delay of the previous flight, the use of predicted weather data introduces noise to the model, but no bias as both error distributions are symmetric and the medians are near-zero.

[5] Wind speeds ranging from 0 to 27 kts, from every direction (rounded to 10 degrees) at least once.

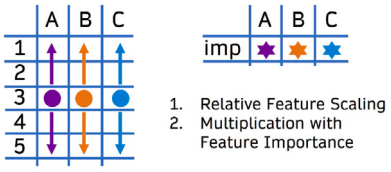

Fig. 7. Explainability Through Feature Scaling Method.

3.5. Model training and result processing

To train the models, the full dataset is randomly split into a training dataset (80% of data) and a test dataset (20% of data). The departure delays in the training and test datasets are similarly distributed as in the full dataset. To minimize data leakage, the split was made per day instead of per flight. This ensures that when testing the model, it has no prior knowledge about the dynamics on the day of the flight.

To facilitate probabilistic departure delay forecasting, the models use all independent predictions to create probability distributions. This is preferred over using majority voting or computing the mean of all independent predictions since probability distributions indicate the likelihood that a delay value is predicted. To ensure high granularity, 1000 unique independent estimators are considered. To create dynamic probabilistic departure delay predictions, the predictions for all prediction moments are combined. The dynamic probabilistic departure delay predictions not only show how the predicted delay value changes over time, but also the evolution of the associated probability density and certainty.

The costs and emissions associated with the decisions made using the departure delay prediction model make explainability an important aspect of this study. For that reason, the probabilistic model performance is one of the considerations for the model selection. Therefore, a method is introduced that explains the predictions based on the relative scaling of the features, as illustrated in Fig. 7.

First, the feature values of all flights in the test dataset are scaled by fitting a StandardScaler()[6] on the dataset. The size of the dataset should be at least order of magnitude 1000 to obtain meaningful scaled feature values.[7] The scaling is performed for each feature individually and returns the number of standard deviations a feature value differs from the mean of all flights. Large scaled feature values indicate that a flight stands out from others in the respective feature. Simply scaling the feature values does not explain the model prediction; it only considers the model inputs, not what the model is doing with this input data. For that reason, the scaled values are weighted by the feature importances, thereby including the importance assessment of the model. Although this method does not explain the exact decision-making of the model, it indicates how the model has treated the underlying data and how this affects the prediction for a certain flight. Finally, for usability purposes, thresholds are determined for classifying the scaled weighted feature values towards large, moderate, and small impacts.[8] Following these impacts, an explainability message is constructed and provided to the flight dispatchers.

4. Results

This section aims to present and analyse the results from the four proposed models. The global model performance is presented in Section 4.1, where the most suitable model is selected. The dynamic prediction behaviour of this model is then presented in Section 4.2. The model explainability results are discussed in Section 4.3.

[6] From the scikit-learn module.

[7] To have enough data points to generate feature distributions of sufficient granularity.

[8] $X \geq 0.5 \rightarrow$ large impact; $0.3 \geq X > 0.5 \rightarrow$ moderate impact; $0.1 \geq X > 0.3 \rightarrow$ small impact, for scaled weighted feature value X.

4.1. Global model performance and selection

The performance metrics used for the model selection are outlined in Section 4.1.1. The global and probabilistic model performances are presented in Section 4.1.2 and Section 4.1.3, respectively. Finally, the model performance per departure delay bin is discussed in Section 4.1.4.

4.1.1. Performance metrics

Five performance metrics are proposed for evaluating the four proposed models, covering the model error, model fit, and probabilistic performance. Each metric is briefly elaborated upon in Table 3. Whereas the first three metrics are commonly used, the final two metrics are introduced in this study, especially to quantify the probabilistic performance of the models. The *ActInDistr* metric represents the percentage of flights for which the actual departure delay is within the predicted departure probability distribution. The *AvgIQR* metric represents the average inter-quartile range of predicted flights, a measure of the model confidence. Ideally, models score high on *ActInDistr* and low on *AvgIQR*, resulting in confident and correct predictions. Overconfident models predict narrow probability distributions and thus score low for both metrics, while underconfident models predict wide probability distributions and thus score high for both metrics.

4.1.2. Global model performance

For each model, distinct sub-models are developed for every prediction moment. For each respective model, the MAE, RMSE, and R2 are tabulated in Table 4 and graphically presented in Fig. 8. For all models, the errors decrease for shorter prediction horizons, whilst the R2-values increase. This follows logically from the perceived updates on the dynamic features. For these features, the change in correlation with the departure delay is illustrated in Fig. 9. As expected, the current average delays at hub airport and current TSAT delays yield much higher correlations over time. Fig. 8 shows that between 30 and 15 min before scheduled departure, the model improves most significantly. This can be explained by the fact that a large share of delays occur just before departure. Including more short-term features would further strengthen this effect, but adding such parameters is invaluable at larger prediction horizons, which is the main focus of this study.

In terms of MAE and RMSE errors, the CatBoostTH model slightly outperforms CatBoostPR. Absolute differences in MAE, RMSE, and R2 between models are small and decrease with time, as shown in Table 4. When expressed as percentages relative to the best model, the differences remain constant over time (3.1% for the MAE, 2.1% for the RMSE). These results suggest that, for this particular dataset, a smooth iteration scheme is preferable, as it may more effectively reduce the average error in the loss function. However, this approach could, in some cases, lead to overfitting or model instability. Therefore, empirical testing is essential for each dataset to ensure optimal performance.

The CatBoostTH model has a quasi-constant advantage over the other models because only for the Deep Neural Network at the 30 min prediction horizon, the relative differences are larger. This may be explained by the model reaching the maximum number of epochs, whereas for other prediction horizons, it converged earlier.

The Deep Neural Network does not outperform the other models. Next to that, the model imposes difficulties for providing probabilistic and explainable results. Thus, it was decided to eliminate the model from further selection in this work. The prediction accuracy of the other models is illustrated in Fig. 10, where the actual and predicted delays for a 90 min prediction horizon are plotted as a heatmap. In the ideal situation, the data points follow the diagonal dotted line, where the predicted delay equals the actual delay. For all models, the distributions follow this line to some extent, albeit with notable noise. All models tend to overpredict for flights with small delays and underpredict for flights with larger delays. This behaviour is partially caused by the splitting nature of the structure of the models as well as the absence

Table 3
Performance metrics.

Performance metric	Explanation
Mean Absolute Error (*MAE*)	Absolute error between predicted and actual values
Root Mean Square Error (*RMSE*)	Standard deviation of errors between predicted and actual values
Coefficient of Determination (*R2*)	Proportion of variation in dependent variable predictable from independent variable
Actual in Distribution Percentage (*ActInDistr*)	Percentage of flights for which actual value is in predicted probability distribution
Average Inter-Quartile Range (*AvgIQR*)	Average inter-quantile range for predicted flights

Table 4
Global model performance for out-of-sample data.

		Prediction moment [min]					
		−90	−75	−60	−45	−30	−15
Random Forest	MAE [min]	8.46	8.31	8.11	7.90	7.67	7.37
	RMSE [min]	11.91	11.69	11.42	11.13	10.81	10.44
	R2 [–]	0.55	0.57	0.59	0.61	0.63	0.65
CatBoostTH	MAE [min]	8.20	8.06	7.88	7.69	7.47	7.15
	RMSE [min]	11.67	11.46	11.20	10.93	10.62	10.23
	R2 [–]	0.57	0.58	0.60	0.62	0.64	0.67
CatBoostPR	MAE [min]	8.26	8.11	7.93	7.74	7.52	7.20
	RMSE [min]	11.72	11.51	11.25	10.98	10.67	10.28
	R2 [–]	0.56	0.58	0.60	0.62	0.64	0.67
Deep Neural Network	MAE [min]	8.35	8.26	8.07	7.91	7.82	7.31
	RMSE [min]	11.81	11.64	11.39	11.16	10.92	10.42
	R2 [–]	0.56	0.57	0.59	0.61	0.62	0.66

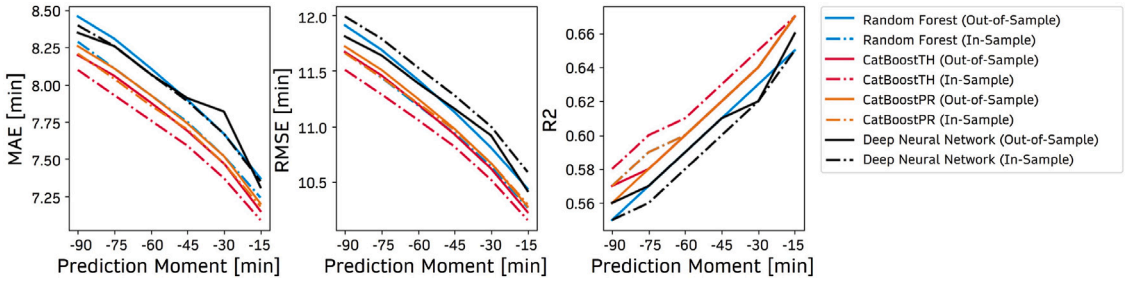

Fig. 8. Global Model Performance over Time in terms of MAE, RMSE, and R2.

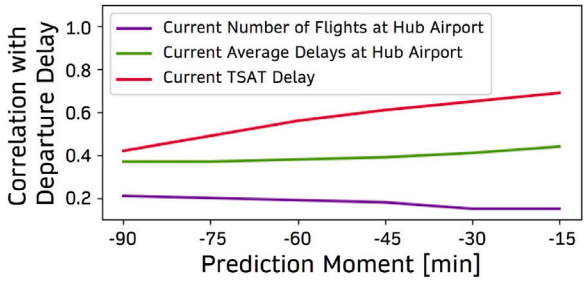

Fig. 9. Dynamic Feature Correlation over Time.

of possibly valuable information and the uncertainty associated with large prediction horizons.

For the Random Forest model in Fig. 10, the distribution of predicted delays shows a valley around 0 min and a peak around −5 min. This behaviour is caused by the data from the COVID-19 period; excluding this data removes the peak. For these flights, the total daily passenger numbers are significantly lower than regular operations. The valley can be explained by the absence of flights with total daily passenger numbers that lie in between the COVID-19 period and regular operations. The CatBoost models better correct for this data anomaly. It should be noted that excluding flights from the COVID-19 period was tested, but resulted in poorer prediction results because of a reduction in the dataset size

4.1.3. Probabilistic model performance

Based on the three global performance metrics, it may seem straightforward to select CatBoostTH for further use. Given the emphasis on explainability in this study, however, the probabilistic performance also needs to be evaluated. For the three remaining models, the ActInDistr and AvgIQR are tabulated in Table 5.

Evaluating the probabilistic performance is a trade-off between two metrics. Ideally, a model has high correctness (i.e. high ActInDistr) and high confidence (i.e. low AvgIQR). Models with high ActInDistr and above-average AvgIQR are preferred over models with low ActInDistr and below-average AvgIQR, as the latter are confident yet incorrect models. Based on the results of Table 5, the Random Forest tends to be less confident on its prediction with a higher variance. This leads to fewer instances where none of the individual tree predictions align with the actual delay. In contrast, CatBoost is highly confident in its predictions, displaying lower variance. However, this confidence increases the likelihood of scenarios where the predicted delay fails to account for the actual observed delay. The Random Forest model yields generally higher prediction variation, which is desired as it prevents the model being confidently wrong, as can be the case for the Catboost models when it fails to grasp the delay cause.

Fig. 11 shows that for around 78% of flights, the actual delays are within the predicted probability distributions for the Random Forest model, greatly outperforming the other two models (8% and 44%). Although the AvgIQR is higher for the Random Forest, the model refrains from predicting wide delay probability distributions; the model is still able to distinguish high and low likelihoods for different delay values.ß

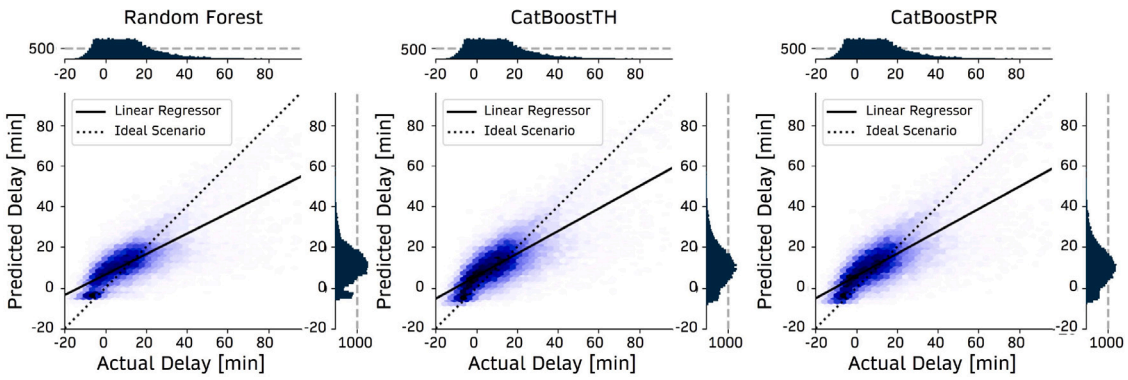

Fig. 10. Prediction Accuracies for Random Forest, CatBoostTH and CatBoostPR models for the 90-Minute Prediction Horizon.

Table 5
Probabilistic model performance for out-of-sample data.

		Prediction moment [min]					
		−90	−75	−60	−45	−30	−15
Random Forest	ActInDistr [%]	76.09	76.82	77.51	78.60	79.17	80.00
	AvgIQR [min]	5.29	5.28	5.24	5.20	5.10	4.99
CatBoostTH	ActInDistr [%]	7.43	7.67	7.61	7.93	7.87	7.89
	AvgIQR [min]	0.39	0.39	0.39	0.38	0.38	0.36
CatBoostPR	ActInDistr [%]	42.23	42.71	43.30	43.44	43.99	45.20
	AvgIQR [min]	2.35	2.35	2.33	2.29	2.26	2.21

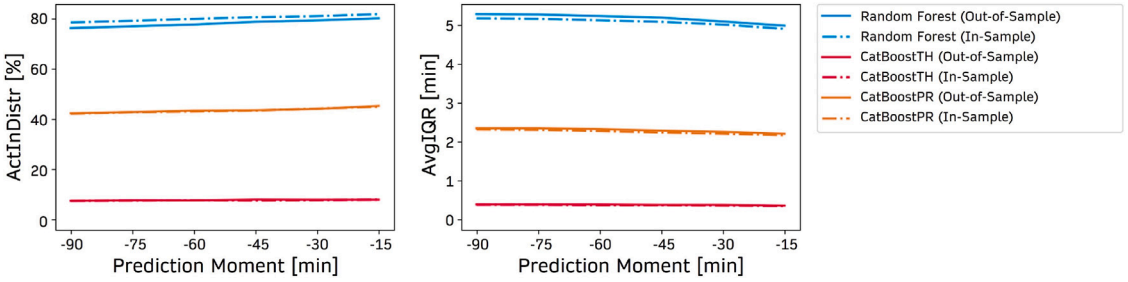

Fig. 11. Probabilistic Model Performance over Time in terms of ActInDistr and AvgIQR.

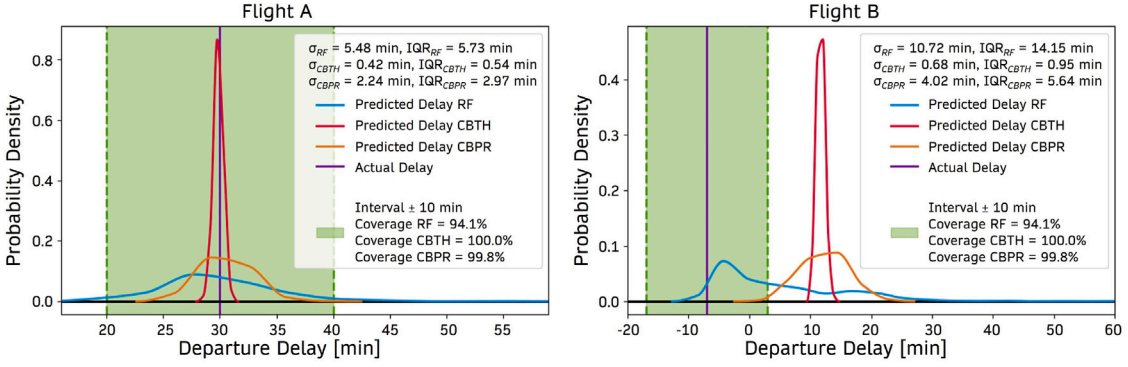

Fig. 12. Probabilistic Departure Delay Predictions for Flight A (left) and Flight B (right) by Random Forest, CatBoostTH and CatBoostPR models (abbreviated RF, CBTH and CBPR, respectively) for the 90-Minute Prediction Horizon.

Fig. 12 illustrates the probabilistic advantage of the Random Forest method over the two CatBoost models. To quantify the accuracy of the probabilistic prediction, a coverage is calculated. For the example flights, the percentage of predicted probability density within an interval of ±10 min around the actual delay is determined. For flights with small prediction errors (such as Flight A), all models achieve high coverages, especially the confident models. For flights with larger prediction errors (such as Flight B), the Random Forest reaches much higher coverages than both CatBoost models. The Random Forest model, although less confident, thus provides a better probabilistic prediction. This can be attributed to Random Forest retaining weaker trees, which may better represent the noise or outliers in the data, providing a more nuanced representation. In contrast, boosting methods, such as the Catboost, prioritize minimizing errors by reducing the effect of these weaker trees, which can lead to overly confident predictions. This is positive from a user perspective - models are desired

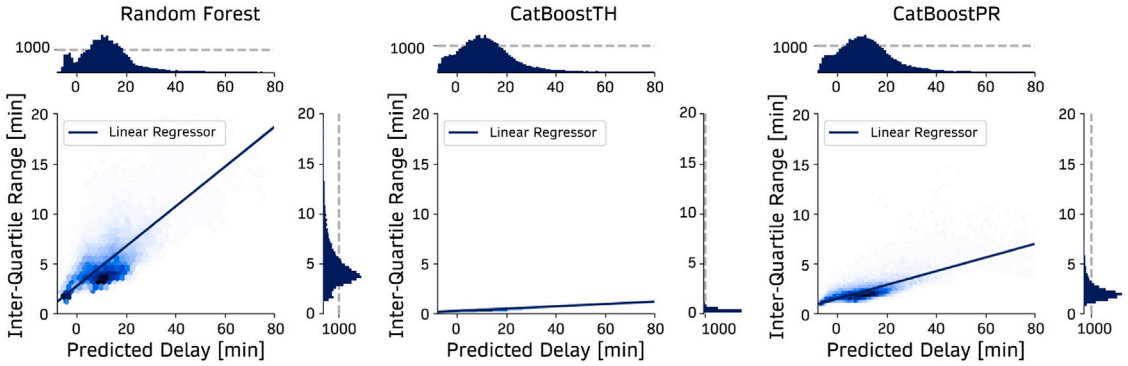

Fig. 13. Relation between Uncertainties and Predicted Delays for Random Forest, CatBoostTH and CatBoostPR models for the 90-Minute Prediction Horizon.

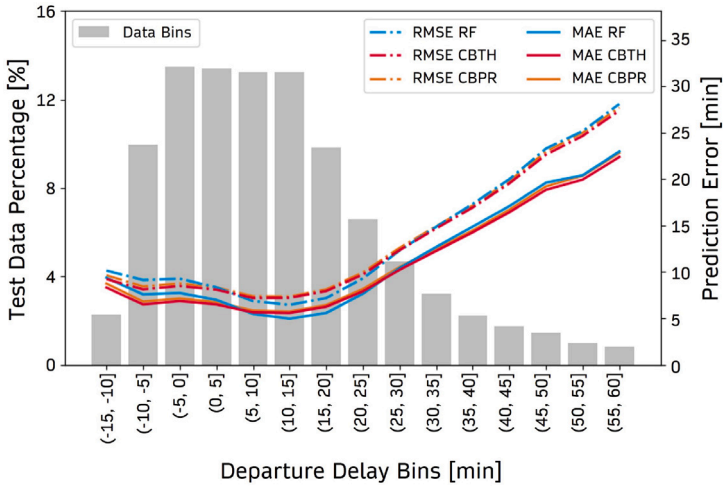

Fig. 14. Binned Performance for Random Forest, CatBoostTH and CatBoostPR models (abbreviated RF, CBTH, and CBPR respectively) for the 90-Minute Prediction Horizon, measured in terms of MAE and RMSE.

to indicate uncertainty associated to possibly incorrect predictions. If a model fails to do so, the user is likely to lose faith in the model over time.

Finally, the in ActInDistr and AvgIQR only marginally improve over time. Although the AvgIQR is expected to decrease over time as a result of the updated dynamic features, this is only very slightly the case. Since flight delays commonly arise close to departure, the average predicted delays are higher for shorter prediction horizons. It is empirically observed that some relation exists between the magnitude of the predicted delay and the certainty of the model, as illustrated in Fig. 13. As the Random Forest and CatBoostPR models predict higher delays, their predictions become less certain. This can be explained by the sampled dataset presented in Fig. 2, which contains relatively more flights with small delays than flights with large delays. Statistically, the probability of the model being trained on comparable flights is higher for flights with small delays, which results in more confident predictions for such flights. For the extremely confident CatBoostTH model, this pattern is hardly visible, as the IQR remains small for almost all predictions.

4.1.4. Binned model performance

The model performance per departure delay bin should be considered for the model selection. There is a physical limit[9] to the number of minutes a flight dispatcher can speed up a flight to compensate for departure delays.

Flights with departure delays smaller than 25 min are particularly interesting for flight dispatchers to slow down or speed up. For that reason, the sampling strategy outlined in Section 3.2 was adopted to optimize the model performance in this part of the departure delay spectrum. Fig. 14 presents the model errors for departure delay bins of 5 min, where the vertical bars represent the share of data in the bins. The distributions are comparable, particularly those of the CatBoostTH and CatBoostPR models and for delays larger than 25 min. For the flights particularly interesting for en-route speed optimization, i.e. departure delays of 5 to 25 min, the Random Forest model yields the smallest prediction errors. For that reason, also considering the superior probabilistic performance of this model, the Random Forest model is selected for the remainder of the study.

4.2. Dynamic model prediction behaviour

The changes in prediction error and IQR between every two consecutive prediction moments are presented in Table 6. Between the 90 and 75 min prediction moments already, the prediction error and IQR decrease for more flights than for which they increase. These ratios only improve for shorter prediction horizons. Between the 30 and 15 min prediction horizons, for over 60% of flights, the prediction confidence still improves. Although, at shorter prediction horizons, a significant number of flights is predicted accurately and with greater certainty, this is not the case for all flights. A few examples are discussed next.

For flight C in Fig. 15, the prediction error decreases as the certainty increases. Although the initial prediction error is relatively large, the model corrects for it as the dynamic features are updated. Alternatively,

[9] Dependent on many factors, e.g. flight distance, weather, aircraft weight, and ATFM.

Table 6
Percentage of test flights for which prediction error and IQR increase or decrease between prediction moments.

	Prediction interval [min]				
	−90 to −75	−75 to −60	−60 to −45	−45 to −30	−30 to −15
Decreasing Error Decreasing IQR	29.06%	30.34%	30.44%	31.42%	33.96%
Decreasing Error Increasing IQR	24.47%	24.03%	24.20%	23.23%	20.37%
Increasing Error Decreasing IQR	22.14%	22.64%	22.52%	23.34%	26.26%
Increasing Error Increasing IQR	24.33%	22.99%	22.84%	22.01%	19.41%

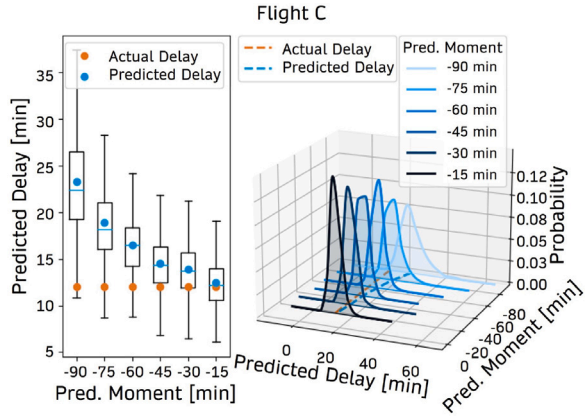

Fig. 15. Dynamic Probabilistic Departure Delay Prediction for Flight C.

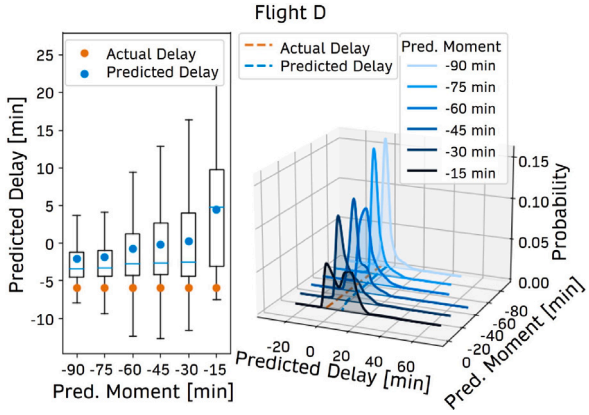

Fig. 16. Dynamic Probabilistic Departure Delay Prediction for Flight D.

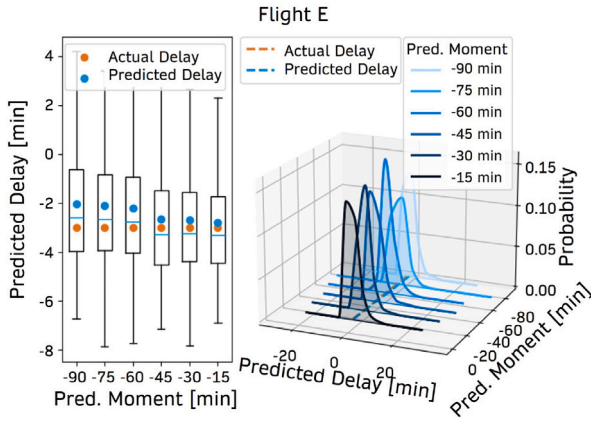

Fig. 17. Dynamic Probabilistic Departure Delay Prediction for Flight E.

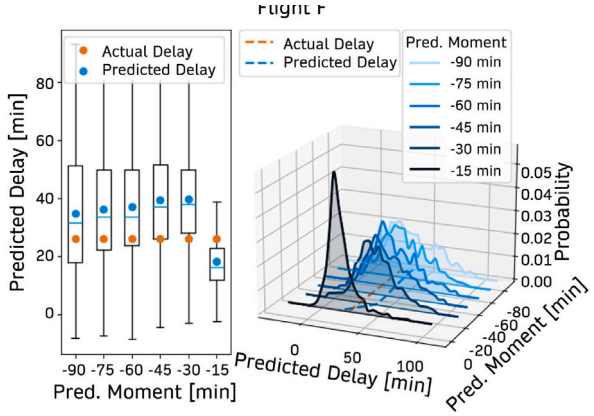

Fig. 18. Dynamic Probabilistic Departure Delay Prediction for Flight F.

for Flight D in Fig. 16, the opposite is true. The model diverges from its initial prediction because, for smaller prediction horizons, it considers the TSAT delay (which is 11 min for all prediction moments) to be more important. Therefore, the predicted delay slightly increases over time. Just one minute after the final prediction moment, the TSAT delay was updated to −5 min, which explains that the actual delay is much lower than predicted.

For a large share of flights, the prediction error and IQR hardly vary over time. For 54% of flights, the prediction error changes for all prediction intervals combined is smaller than 5 min. Similarly, for 61% of flights, the IQR changes for all prediction intervals combined is smaller than 3 min. For these flights, the model is either able to accurately predict the delays at the first prediction moment already or it is unable to improve its initial prediction. Flight E in Fig. 17, is an example for which the prediction error and IQR are small for all prediction moments. If instead, the error would be constantly large, this would result in a translation of the distribution with respect to the *Predicted Delay*-axis. Alternatively, changes in certainty would flatten or steepen the distribution curves.

For a number of flights, the prediction certainty decreases as the prediction horizon becomes smaller. This is mostly caused by the model predicting higher delays, which results in additional uncertainty, previously explained in Section 4.1.3. Moreover, increasing uncertainty can be caused by contradicting features, for example flights with large effective delays of previous flights where the TSAT delay has not yet been updated. Fortunately, in many cases, the certainty increases over the prediction horizon, in line with Table 6. Flight F, illustrated in Fig. 18, is an example for which the TSAT delay was large (45 min) for the first five prediction horizons and changed to just 12 min at the final prediction horizon, allowing the model to make a final prediction with higher confidence.

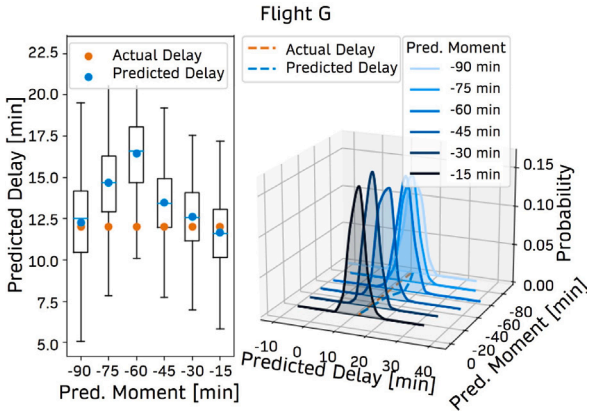

Fig. 19. Dynamic Probabilistic Departure Delay Prediction for Flight G.

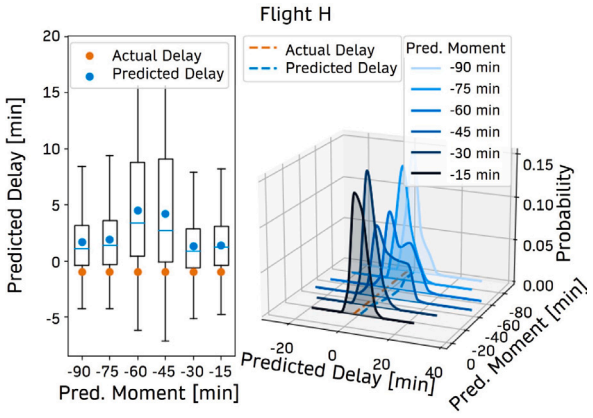

Fig. 20. Dynamic Probabilistic Departure Delay Prediction for Flight H.

For some flights, the prediction error and IQR temporarily increase for some parts of the prediction horizon. Temporary changes in prediction error are almost always the cause of temporary changes in prediction certainty. Flight G in Fig. 19, is an example for which the prediction error is temporarily higher, caused by a temporarily larger TSAT delay value. This causes the probability distribution to temporarily translate with respect to the *Predicted Delay*-axis. Flight H in Fig. 20, is an example for which temporary high average delays at the hub airport cause the prediction error to increase, thereby temporarily decreasing the prediction certainty. The coupling between the prediction error and uncertainty causes the probability distribution to be translated and flattened simultaneously.

4.3. Model explainability results

The feature importances associated with the Random Forest model are presented in Section 4.3.1 and the results of the relative feature scaling method are presented in Section 4.3.2, with two examples of explainability messages outputted to the flight dispatcher.

4.3.1. Feature importance

Given the dynamic feature updates and their improving correlations with departure delays, presented in Fig. 9, the feature importances are expected to change over time as well. The feature importances for the 90 min prediction horizon as well as their propagation over time are presented in Fig. 21. The ten least important features do not change significantly over the full prediction horizon. The shifts noticed in the five most important features are more interesting to evaluate. Whereas at longer prediction horizons, the model considers passenger information more important (ranked 3rd and 4th), for the shorter prediction horizons, these importances drop to 4th and 5th place respectively, almost halving in magnitude. Dynamic features describing the current TSAT delay and average delays at the hub airport become more important for shorter prediction horizons (1st and 3rd compared to 2nd and 5th place).

Table 7
Sensitivity flight baseline parameters.

Feature	Baseline	Change
Month of Year	4	±3
Hour of Day	14	±4
Passenger Load Factor	0.75	±0.1
Baggage Load Factor	0.85	±0.1
Transfer Passenger Percentage	60	±10
Number of Passengers Reduced Mobility	5	±4
Total Passengers Day In Membership Program	42 000	±5000
Median Delay of Flight Number [min]	10	±10
Effective Delay Previous Flight [min]	15	±10
Current Number of Flights at Hub Airport [min]	15	±5
Current Average Delays at Hub Airport [min]	10	±10
Current TSAT Delay [min]	20	±10
Last Aircraft Tail Swap [min]	−900	±300
Wind Speed Longitudinal Direction [kts]	10	±5
Wind Speed Latitudinal Direction [kts]	10	±5

4.3.2. Explainability through feature scaling

Following the relative feature scaling method presented in Section 3.5, explainability messages are provided to the flight dispatchers, indicating the significance of a given feature for this flight relative to all others. Including this information in the explainability message was one of the wishes of the flight dispatchers. The explainability message for Flight E, previously discussed in Section 4.2, is presented below in Fig. 22 for both the 90 min and 15 min prediction horizon. From the messages, the user is informed that the lower delay prediction at the 15 min prediction horizon is caused by the decrease in current TSAT delay and current average delays at hub airport. Finally, the impact of delays is excluded for flights with predicted delays of less than 15 min because these delays typically occur due to *quasi-random* operational factors close to departure, not due to the major delay causes the model was trained for.

5. Model validation

This section aims to discuss the efficiency and applicability of the model. First, the sensitivity study is discussed in Section 5.1. Thereafter, the conclusions drawn from two shadow runs are presented in Section 5.2. Finally, an error analysis is performed in Section 5.3, evaluating the flights that were most difficult for the model to predict.

5.1. Sensitivity analysis

This section focuses on the effect of altering a single input feature, one at a time. For this, a baseline flight was set up, the values and respective changes are tabulated in Table 7. The baseline values and changes were chosen such that hypothetically, outputs are most likely to lead to meaningful changes,[10]. The sensitivity of the mean predicted departure delay is shown in Fig. 23 for each of the six prediction moments separately. Furthermore, the probabilistic sensitivity of the model is illustrated in Fig. 24, for the prediction moment 90 min before scheduled departure.

[10] By choosing baseline values and changes that are not towards the extremes of the feature value range, e.g. if the feature ranges from 0.4 to 1, changes are expected to be more meaningful when comparing 0.65, 0.75 and 0.85 instead of 0.4, 0.41 and 0.42 or 0.98, 0.99 and 1.

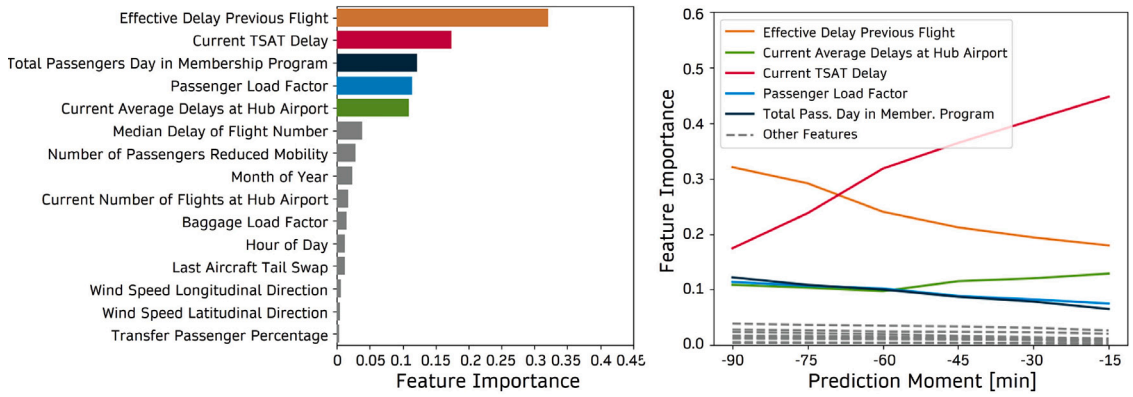

Fig. 21. Feature Importance for the Random Forest Model for the 90-Minute Prediction Horizon (left) and the Feature Importance Developments over the Full Prediction Horizon (right).

Flight Date:	YYYY-MM-DD
Flight:	XX****
Departure Airport:	ZZZ
Prediction Moment:	-90 min
Predicted Delay:	34.65 min
Large impact from:	Current TSAT Delay (90 min. before Scheduled Departure) [min]: 45.0 (average = 6.7)
Moderate impact from:	Current Average Delays at Hub Airport (120 to 90 min. before Scheduled Departure) [min]: 72.44 (average = 14.1)

Flight Date:	YYYY-MM-DD
Flight:	XX****
Departure Airport:	ZZZ
Prediction Moment:	-15 min
Predicted Delay:	18.25 min
Small impact from:	Current Average Delays at Hub Airport (45 to 15 min. before Scheduled Departure) [min]: 43.31 (average = 14.12)
Small impact from:	Current TSAT Delay (15 min. before Scheduled Departure) [min]: 12.0 (average = 8.42)

Fig. 22. Explainability messages are provided to the flight dispatchers.

In Figs. 23 and 24, the predicted departure delay hardly varies for changes on the hour of day, baggage load factor, transfer passenger percentage, passengers with reduced mobility, current number of flights at hub airport, last aircraft tail swap and wind speeds features. This is explained by their relatively low importances in Section 4.3.1. Although among the six most important features in Section 4.3.1, the model is hardly sensitive to changes in the median delay associated with the flight number. The effect of changes in other parameters is better visible. Compared to the baseline month of April, departure delays are predicted to be slightly lower if the flight was to be scheduled in January. Similar relations are obtained for the passenger load factor, current average delays at hub airport, and the total daily passengers in the membership program. These small differences can be explained by seasonality, as historically the delays are smaller in months with fewer passengers (such as January).

The effective delay of previous flight is the most sensitive feature. At the 90 min prediction horizon, an input change of 10 min results in an output change of almost the same size. At closer prediction moments, complying with the decreasing feature importance from Section 4.3.1, the effective delay of previous flight becomes slightly less sensitive. From the probabilistic sensitivity in Fig. 24, it is visible that the model is less confident in predicting larger departure delays. The same behaviour was previously observed in Section 4.1.3.

The model becomes more sensitive to the TSAT delay feature for smaller prediction horizons. It was previously found that this feature becomes more important for smaller prediction horizons. Interestingly, the model is more sensitive to TSAT delay increases than TSAT delay decreases. From historical data, it is observed that TSAT delays increase more frequently than they decrease. Once a new TSAT delay is issued, operations are often centred around accommodating this new time. As such, it rarely happens that a TSAT delay decreases after it has previously increased. Even in the case of decreases, it is often observed that a new increase is issued later in the process. The model has thus successfully learned that a TSAT delay increase is a stronger indicator for higher delays than a TSAT delay decrease is for smaller delays.

5.2. Shadow runs

Two shadow runs were conducted: one for European flights and one for intercontinental flights. Table 8 lists the flights considered during both shadow runs. Because the input data was not available in real-time, the model performance could only be evaluated as soon as the data became available. Some of the flights were delayed for reasons the model was not trained for, these cases are elaborated upon in the next paragraphs. Although the model can predict delays caused by untrained factors to some extent, it cannot predict all such delays. The model strongly benefits from the current TSAT delay feature,[11] since it covers a wide range of operational delay causes.

- **ATFM Slot:** Flight M in Fig. 25, received an ATFM slot of 27 min just after the first prediction moment. This caused the TSAT delay of the flight to get extended by the same amount. To avoid the slot, the flight dispatcher searched for an alternative route to avoid the overcrowded sector. After finding a suitable route, the slot time and TSAT delay returned to their previous values, as is visible in the departure delay prediction, where temporarily higher delays with larger uncertainty are predicted. The flight dispatcher thus is part of the loop, as his/her actions influence new predicted delays.
- **High-priority Flight:** Flight I in Fig. 26, was a high-priority flight because of an important part delivery. Because of a delayed previous flight, the model severely overpredicts for this flight. Given the high priority, every effort was made to keep the turn-around time as small as possible. The model is unable to capture the high priority, as it is an exceptional circumstance.

- **Late Fuelling:** Flight *T* in Fig. 27, was delayed because the fuelling team arrived later than planned. The model captures this effect, as the predicted delay increases at smaller prediction horizons, as

[11] TSAT delays often follow from Target Off-Blocks Time (TOBT) delays, which are issued by the airline itself.

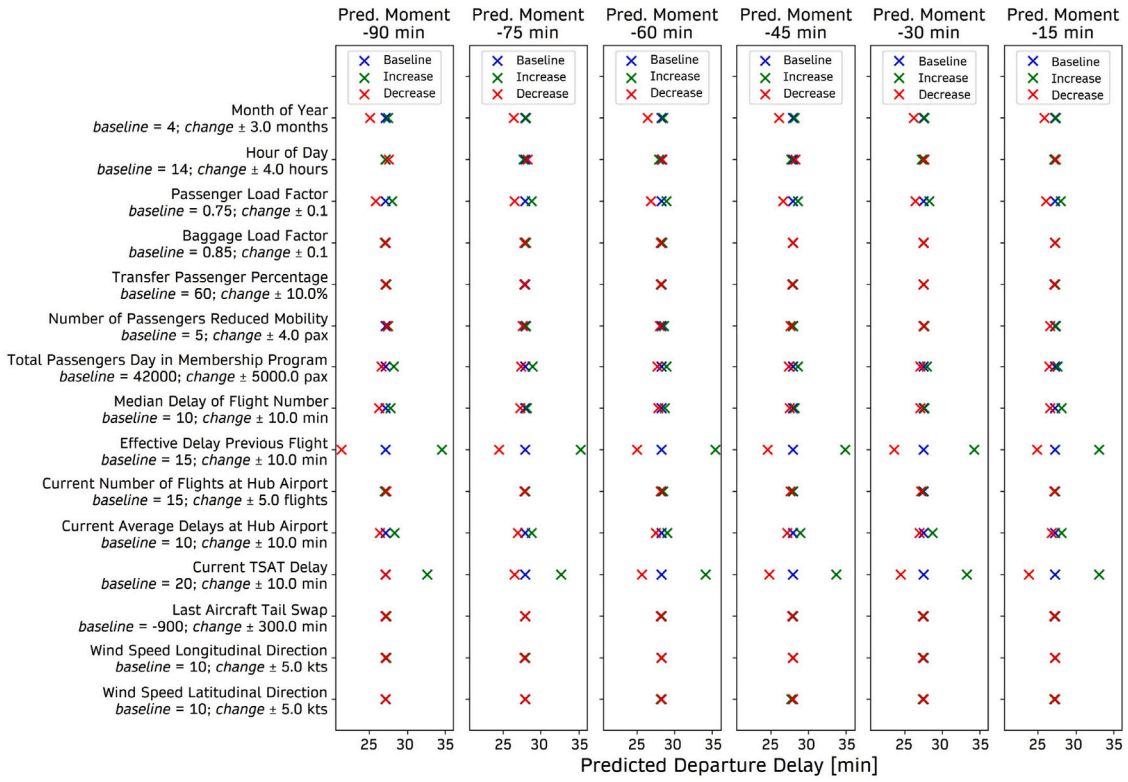

Fig. 23. Sensitivity Analysis for Random Forest Model using Baseline and Change Values from Table 7.

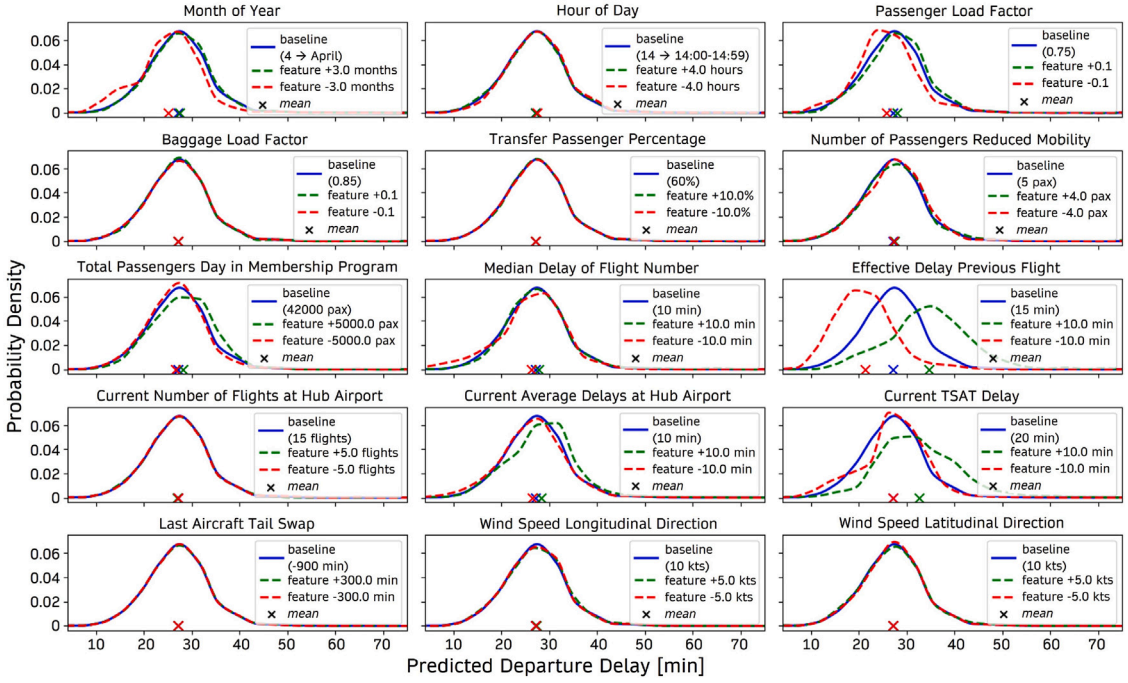

Fig. 24. Probabilistic Sensitivity Analysis for Random Forest Model at 90-Minute Prediction Horizon using Baseline and Change Values from Table 7.

a result of increasing TSAT delays. The TSAT delay feature thus covers the fuelling delay.

- **Late Arrival Baggage:** Flight R, illustrated in Fig. 28, was delayed because the baggage carts arrived later than planned. The model is unable to capture the late arrival of baggage carts, despite small increases in TSAT delays. Instead, the model most likely overpredicts because of a fully booked flight.
- **Delayed Maintenance:** Flight K was delayed for 32 min because of a late return from maintenance. Since the model is not trained with maintenance data, the effective delay of the previous flight is thought to be 0 min, as the aircraft had arrived the day before already. Although the model does not understand the maintenance delay directly, it correctly predicts the departure delay because the maintenance delay was already known 125 min before scheduled

Table 8
Shadow run flights.

Date	Flight number	Actual delay [min]	Predicted delay at the 90 min horizon [min]
2023–10–10	Flight I	14.00	29.36
	Flight J	48.00	59.91
	Flight K	32.00	32.41
	Flight L	27.00	27.22
	Flight M	29.00	28.32
	Flight N	5.00	15.07
2023–10–27	Flight O	9.00	8.41
	Flight P	32.00	32.96
	Flight Q	7.00	12.87
	Flight R	12.00	15.09
	Flight S	−4.00	15.31
	Flight T	15.00	18.59
	Flight U	16.00	15.34
	Flight V	15.00	17.25
	Flight W	5.00	16.04

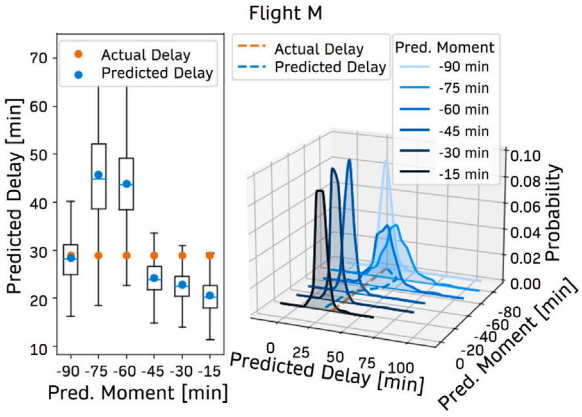

Fig. 25. Dynamic Probabilistic Departure Delay Prediction for Flight M.

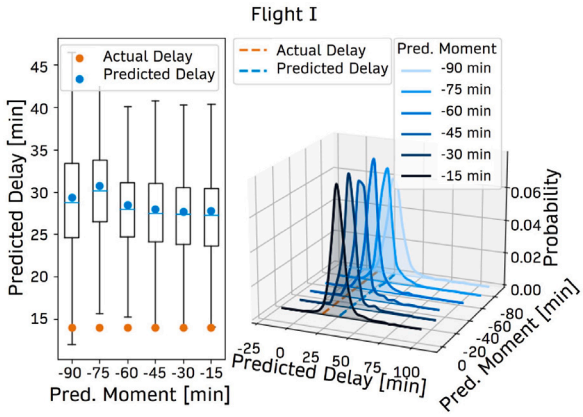

Fig. 26. Dynamic Probabilistic Departure Delay Prediction for Flight I.

departure. For that reason, the TSAT was already updated before the first prediction moment. The TSAT delay feature thus covers the maintenance delay.

5.3. Error analysis

For 179 out of the 33532 flights in the test dataset, the model predicts with an error of over 45 min. It is worth investigating the reason for such errors. Following IATA guidelines (EUROCONTROL, 2023), primary (and secondary) delay codes are issued for delayed flights. These delay codes explain the cause of the flight delay and are thus useful for explaining high prediction errors. It should be noted that ambiguity may exist for the issued delay cause, as different stakeholders have different interests for the delay code issuing.

For the 179 flights, an overview of issued delay codes is presented in Fig. 29. Primary and secondary delays are represented by Delay Code 1 and Delay Code 2, respectively. 75.4% of flights with large prediction errors are caused by just 17 delay codes. For the majority of flights with large prediction errors, the model is not trained for the underlying delay causes because the data is simply unavailable (e.g. loading, fuelling, ATFM delay, and crew rotations) or unpredictable (e.g. security/immigration, missing passengers, and flight deck crew request).

Given that primary delay codes contribute the most to flight delays and that the model accounts for the number of passengers with reduced mobility and the effective delay of previous flights, it is worth exploring why these factors were responsible for the main delay of some flights. For the three flights with delays caused by passengers with reduced mobility, the number of such passengers was 2, 7, and *unknown* respectively. Seeing that most other flights with these numbers of passengers

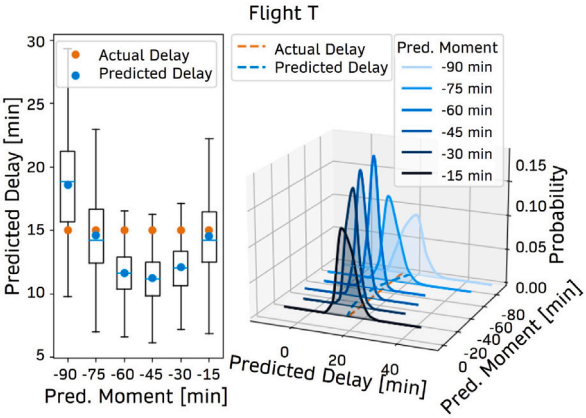

Fig. 27. Dynamic Probabilistic Departure Delay Prediction for Flight T.

with reduced mobility are only little delayed, it is no surprise that the model underpredicted the delay for these three flights. For the one flight with a delay due to aircraft rotations, the inbound flight effectively arrived 26 min late, however, the departure delay was much larger (91 min). Another reason must have caused the remainder of the delay. Thus, the model is not able to correctly predict the delay for this flight just based on the effective delay of the previous flight.

In conclusion, the model is robust to changes in input features and has demonstrated its capability of predicting delays caused by untrained factors. Due to the limited time frame of this study and the fact that before a shadow run, it is unknown what will cause the delays for the upcoming flights, only a number of these untrained factors

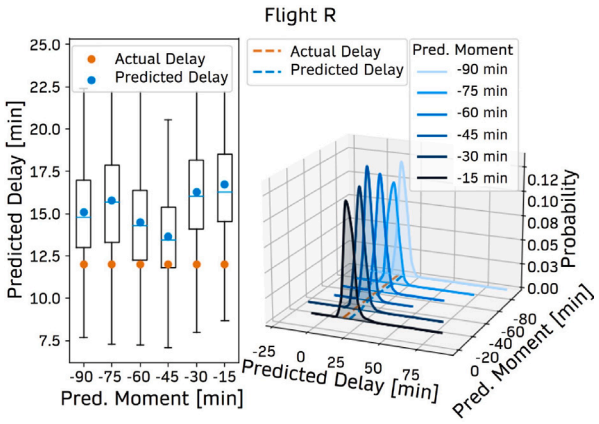

Fig. 28. Dynamic Probabilistic Departure Delay Prediction for Flight R.

could be tested. Further shadow runs are necessary to guarantee the applicability of the model in a broader sense.

6. Discussion

This section aims to further discuss some of the results. The feature elimination process is reflected upon in Section 6.1. After, Section 6.2 discusses the tendency of the model to overpredict small delays. Finally, the model performance is compared to a baseline statistical model in Section 6.3.

6.1. Selected features

The list of selected features was presented in Table 2. In line with the observed feature importance in Fig. 21, for the vast majority of trees in the random forest, the top-level split decision is based on the *Total Passengers Day in Membership Program*, *Current TSAT Delay* or *Effective Delay Previous Flight* feature. These features improved the accuracy of the final model.

The final list of selected features went through several rounds of evaluation, where some interesting features were deemed not beneficial to the model. For example, visibility was an important feature in previous work by Sternberg et al. (2016), it was either covered by one of the other features, reducing the need for a separate feature, or the number of recordings with extreme fog was too low to make the models understand its effect. Other examples are the total number of seats and aircraft size, which showed a high correlation with two other features: total passengers and median delay of flight number. As a result, the total number of seats and aircraft size were removed from the model's feature set. The correlation with total passengers can be attributed to the aircraft's fill rate, which tends to be consistent. The correlation with median delay can be explained by the fact that the fleet consists of three distinct aircraft size groups, each typically experiencing different departure delays. Additionally, the number of seats was not found to be a significant factor in departure delay, as the turnaround process already accounts for longer boarding times in larger aircraft. Therefore, the differentiator in the boarding process is not the actual number of passengers but the relative number, which is captured by the Passenger Load Factor (and Baggage Load Factor) features.

Additionally, flight-specific passenger connection data was not used. Instead, the *percentage* of transfer passengers, was selected. Although this feature does not consider every inbound-outbound flight combination, it still stands out from other research as the feature is unique per outbound flight. This does imply, however, that the flight-specific passenger connection data was not valuable enough to predict departure delays.

Finally, incorporating additional features related to operational decisions from airports, other airlines, or airspace constraints could significantly enhance the model's ability to identify external causes of delays. However, such data is often scarce and confidential. Future work should prioritize the collection and analysis of this data.

6.2. Overpredicting for small delays

The overpredicting behaviour for small delays was previously evaluated in Fig. 10. At the 90 min prediction horizon, flights with actual delays up to 15 min are overpredicted by 5.17 min on average. Alternatively, flights with actual delays of over 15 min are underpredicted by 9.68 min on average. The overall overprediction distribution for the Random Forest model is presented in Fig. 30, for each of the six distinct prediction horizons. Naturally, overprediction leads to unnecessary fuel burn and is undesirable.

The overpredicting behaviour for smaller delays is attributed to two things. Firstly, almost all feature-target relations are positive,[12] see Fig. 31. As a result, when some feature values are above average, the model may already be inclined to predict higher delays since there are no features that impact the predicted delays negatively. Secondly, given the randomness involved with the turn-around process, the model is not always able to accurately predict the delay (R2 is *only* 0.55 at the 90 min prediction horizon). The overpredicting behaviour may also partially be caused by the fact that not all required information is already known at early prediction moments. This is visible in Fig. 30, where the overpredicting behaviour reduces for shorter prediction horizons, but does not disappear completely.

For future work, several other logical follow-up steps can be taken to further improve dynamic probabilistic airline departure delay forecasting. Firstly, one can test the impact of incorporating other novel features. It is recommended to explore the effect of push-back truck availability and crew rotations data as both are critical in the turnaround process. Secondly, it is recommended to consider actual departures rather than planned departures for the current average delays at hub airport, in order to always have complete data. Thirdly, the granularity of the prediction horizons can be increased to 5 or 10 min to allow for quicker incorporation of dynamic feature updates. Finally, it is recommended to perform further validation shadow runs to guarantee the applicability of the model to a wider range of untrained delay causes.

6.3. Improvement compared to statistical baseline model

The developed model is compared to an existing statistical model that is currently used in operations. The latter analyzes past flights, recording the most significant causes of delay, while excluding unpredictable factors like random technical difficulties. This statistical model calculates the moving average of past flights, adjusted for these excluded causes. Note that specific details about this model's functioning are herein intentionally omitted due to confidentiality. However, its use in operational settings suggests that it has an acceptable level of accuracy.

When considering the same case study period, the existing model reaches an MAE of 9.51 min, an RMSE of 18.62 min, and an R2 of 0.13. Since this model does not consider multiple prediction moments, these performance metrics values are the same for all prediction horizons. The prediction performance of both models at the 90 min prediction moment is presented in Fig. 32, considering just a single flight series (upper figure) and a single day at the Hub airport (lower figure). Although the existing model captures the global dynamics, judging from the upper plot of Fig. 32, it is unable to predict large delays. The proposed Random Forest model is much better capable of

[12] The feature-target relations that are not positive are near zero.

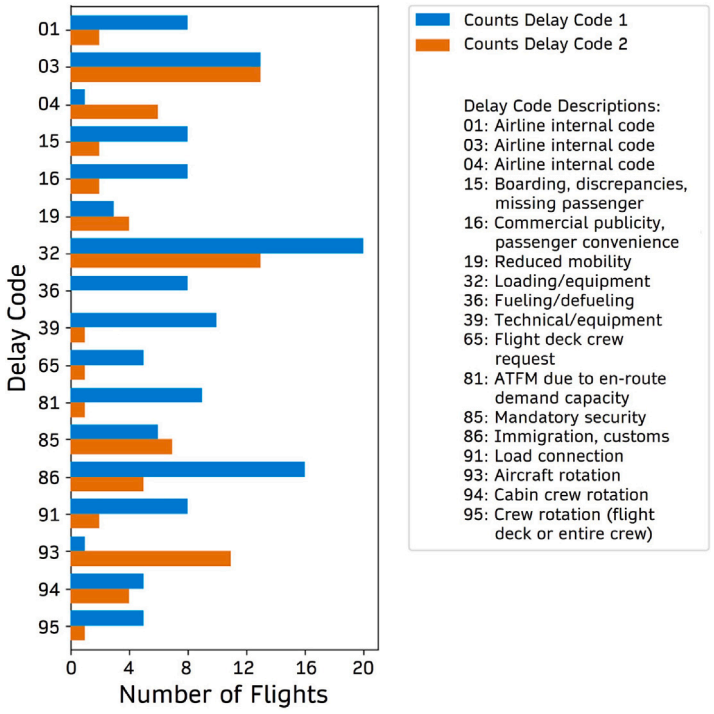

Fig. 29. Delay Code Issued for Flights with Highest Model Prediction Errors.

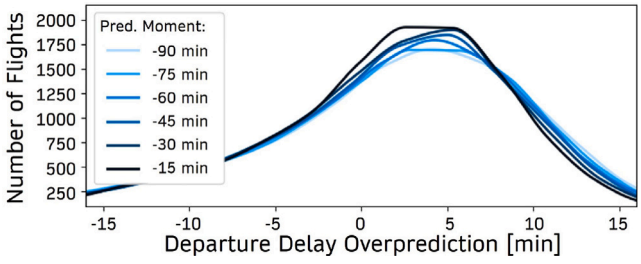

Fig. 30. Random Forest Departure Delay Overprediction.

predicting such large delays. Whereas the existing model overpredicts less for smaller delays, it is useless for days with disrupted operations (see lower figure in Fig. 32). The proposed Random Forest model is particularly of added value for the prediction of severely delayed flights.

6.4. Comparison with other studies

Comparing the results of this model to previous studies presents challenges due to differences in datasets, airlines, airports, as well as differences in the timeframes for predicting departure delays. Generally, studies predicting delays longer than 90 min before departure have reported an MAE of 5 to 15 min across various European airports (Dalmau et al., 2021; Birolini and Jacquillat, 2023). For shorter look-ahead times, the MAE typically decreases to around 3.8 to 7.7 min (Sun et al., 2022). In the United States, studies tend to report lower MAEs, ranging from 2.48 to 3.6 min (Wang et al., 2022). In contrast, the random forest model in this study exhibits an MAE of 8.46 min around 90 min before departure, which decreases to 7.37 min at 15 min prior to departure (see Table 4). This indicates that our model's performance aligns with recent optimal values identified in literature. Nevertheless, direct comparison is impossible as different datasets, fleets, airlines, and airports are used across all studies.

A key advantage of our model is its probabilistic nature, which evolves over time as more information becomes available. The parameter 'ActInDist', introduced in this study, measures how close predictions align with the final delay, a metric not addressed in other works. Two other studies focused on probabilistic departure delay forecasts for individual flights, namely Vorage (2021) and Zoutendijk (Zoutendijk and Mitici, 2021). These achieved an MAE between 12.51 to 13.23 min several days before flight. Zoutendijk (Zoutendijk and Mitici, 2021) achieved a Continuous Ranked Probability Score (CRPS) (Matheson and Winkler, 1976), measuring the deviation of the estimated delay from the actual value, of 8.86 min. Our model seems to have improved from this value, likely due to the reduced training dataset size used by Zoutendijk (Zoutendijk and Mitici, 2021).

Currently, there is a lack of direct comparisons among studies focused on departure delays. This is partly due to the use of data confidential to airlines and airports. Future efforts should prioritize the release of data suitable for comparative research. Given that elements related to airline, passengers, and airport technical resources are crucial for model performance, as demonstrated in this study, collaboration among various industry stakeholders is essential for such efforts.

7. Conclusion

Hub-and-spoke airlines generally adjust their operations to guarantee passenger connections. For that reason, punctuality is one of the key performance indicators of such airlines. To ensure on-time arrivals, flights that were delayed upon departure need to compensate for the lost time whilst airborne. For adequate fuelling, flight dispatchers use departure delay predictions. The goal of this study was to propose an explainable supervised learning model that improves on an existing departure delay prediction model, as there was room for improvement.

A Random Forest model was selected as it outperformed the other models for the flights most suitable for en-route speed optimization and demonstrated superior probabilistic performance. The dynamic probabilistic model performance analysis then indicated that for shorter prediction horizons, the model was able to improve on initial predictions for a large number of flights, both in terms of correctness

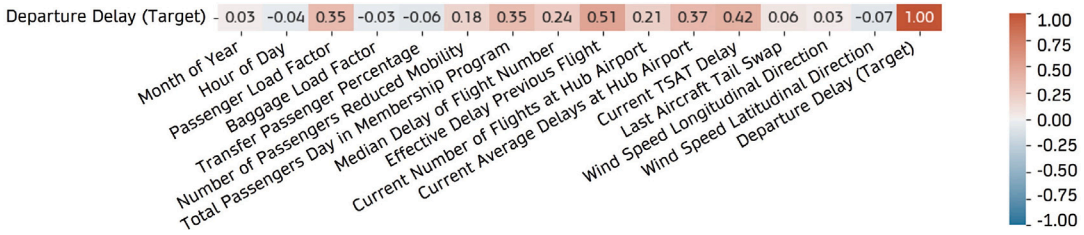

Fig. 31. Target Correlation Matrix for the 90-Minute Prediction Horizon.

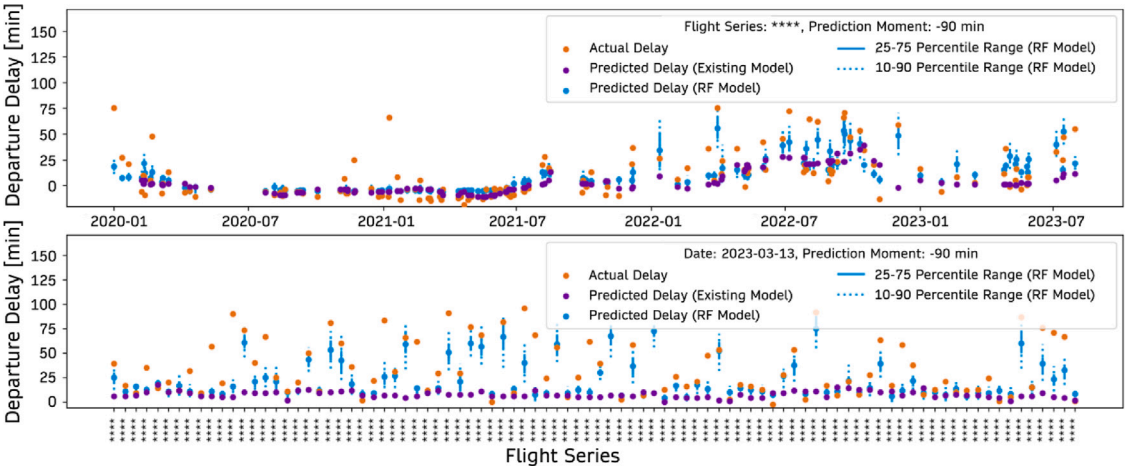

Fig. 32. Comparison with baseline statistical model at 90-Minute Prediction Horizon.

and certainty. At the default 90 min prediction horizon, the model reaches an MAE of 8.46 min, an RMSE of 11.91 min, and an R2 of 0.55. At the 15 min prediction horizon, these values improve to 7.37 min, 10.44 min, and 0.65, respectively. At all prediction moments, for around 78% of flights, the actual delay was within the predicted departure delay probability distribution. For the flights that are most suitable for en-route speed optimization, the Random Forest model reached MAE values of around 5 min.

Finally, the model was validated in two shadow runs, proving the robustness of the model in real-life operational scenarios. Future work may focus on further testing the applicability of the model in different use cases. Further data collection and analysis may be performed to cover scenarios where the model lacks efficiency due to unknown missing data. Additionally, the granularity of the prediction horizons may be increased to improve the quality of decision support to airlines. Finally, adding features related to airspace constraints has the potential to improve the accuracy of the model.

CRediT authorship contribution statement

Maarten Beltman: Writing – original draft, Validation, Software, Data curation. **Marta Ribeiro:** Writing – review & editing, Supervision. **Jasper de Wilde:** Writing – review & editing, Supervision. **Junzi Sun:** Writing – review & editing, Supervision.

References

Abdel-Aty, Mohamed, Lee, Chris, Bai, Yuqiong, Li, Xin, Michalak, Martin, 2007. Detecting periodic patterns of arrival delay. J. Air Transp. Manag. 13 (6), 355–361.

Alonso, Hugo, Loureiro, António, 2015. Predicting flight departure delay at porto airport: A preliminary study. In: 2015 7th International Joint Conference on Computational Intelligence. IJCCI, 3, IEEE, pp. 93–98.

Birolini, Sebastian, Jacquillat, Alexandre, 2023. Day-ahead aircraft routing with data-driven primary delay predictions. European J. Oper. Res..

CatBoost, 2023. CatBoost. https://catboost.ai/ (Accessed on 10 October 2023).

Choi, Sun, Kim, Young Jin, Briceno, Simon, Mavris, Dimitri, 2016. Prediction of weather-induced airline delays based on machine learning algorithms. In: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–6.

Ciruelos, C., Arranz, A., Etxebarria, I., Peces, S., Campanelli, B., Fleurquin, P., Eguiluz, V.M., Ramasco, J.J., 2015. Modelling delay propagation trees for scheduled flights. In: Proceedings of the 11th USA/EUROPE Air Traffic Management R&D Seminar, Lisbon, Portugal. pp. 23–26.

Dalmau, Ramon, Ballerini, Franck, Naessens, Herbert, Belkoura, Seddik, Wangnick, Sebastian, 2021. An explainable machine learning approach to improve take-off time predictions. J. Air Transp. Manag. 95, 102090.

EUROCONTROL, 2023. All-causes delays to air transport in Europe annual 2022. CODA Dig..

Felder, Martin, Kaifel, Anton, Graves, Alex, 2010. Wind power prediction using mixture density recurrent neural networks. In: Poster presentation Gehalten auf der European wind energy conference.

Gopalakrishnan, Karthik, Balakrishnan, Hamsa, 2017. A comparative analysis of models for predicting delays in air traffic networks. In: Twelfth USA/Europe Air Traffic Management Research and Development Seminar. ATM Seminar.

Hancock, John, Khoshgoftaar, Taghi, 2020. CatBoost for big data: an interdisciplinary review. http://dx.doi.org/10.21203/rs.3.rs-54646/v1,

Horiguchi, Yuji, Baba, Yukino, Kashima, Hisashi, Suzuki, Masahito, Kayahara, Hiroki, Maeno, Jun, 2017. Predicting fuel consumption and flight delays for low-cost airlines. In: Proceedings of the AAAI Conference on Artificial Intelligence. 31, (2), pp. 4686–4693.

Iowa State University, 2023. Iowa environmental mesonet. Data retrieved from Iowa State University Environmental Mesonet, https://mesonet.agron.iastate.edu/request/download.phtml (Accessed on 22 March 2023).

Kalliguddi, Anish M., Leboulluec, Aera K., 2017. Predictive modeling of aircraft flight delay. Univers. J. Manag. 5 (10), 485–491.

Khan, Waqar Ahmed, Ma, Hoi-Lam, Chung, Sai-Ho, Wen, Xin, 2021. Hierarchical integrated machine learning model for predicting flight departure delays and duration in series. Transp. Res. Part C Emerg. Technol. 129, 103225.

Lan, Shan, Clarke, John-Paul, Barnhart, Cynthia, 2006. Planning for robust airline operations: Optimizing aircraft routings and flight departure times to minimize passenger disruptions. Transp. Sci. 40 (1), 15–28.

Manna, Suvojit, Biswas, Sanket, Kundu, Riyanka, Rakshit, Somnath, Gupta, Priti, Barman, Subhas, 2017. A statistical approach to predict flight delay using gradient boosted decision tree. In: 2017 International Conference on Computational Intelligence in Data Science. ICCIDS, IEEE, pp. 1–5.

Matheson, James E., Winkler, Robert L., 1976. Scoring rules for continuous probability distributions. Manag. Sci. 22 (10), 1087–1096.

Mueller, Eric, Chatterji, Gano, 2002. Analysis of aircraft arrival and departure delay characteristics. In: AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum. p. 5866.

Pérez-Rodríguez, Jorge Vicente, Pérez-Sánchez, José María, Gómez-Déniz, Emilio, 2017. Modelling the asymmetric probabilistic delay of aircraft arrival. J. Air Transp. Manag. 62, 90–98.

Prokhorenkova, Liudmila, Gusev, Gleb, Vorobev, Aleksandr, Dorogush, Anna Veronika, Gulin, Andrey, 2018. CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 31.

Rebollo, Juan Jose, Balakrishnan, Hamsa, 2014. Characterization and prediction of air traffic delays. Transp. Res. Part C Emerg. Technol. 44, 231–241.

Schösser, Delia, Schönberger, Jörn, 2022. On the performance of machine learning based flight delay prediction–Investigating the impact of short-term features. Promet- Traffic& Transp. 34 (6), 825–838.

Sismanidou, Athina, Tarradellas, Joan, Suau-Sanchez, Pere, 2022. The uneven geography of US air traffic delays: Quantifying the impact of connecting passengers on delay propagation. J. Transp. Geogr. 98, 103260.

Sternberg, Alice, Carvalho, Diego, Murta, Leonardo, Soares, Jorge, Ogasawara, Eduardo, 2016. An analysis of Brazilian flight delays based on frequent patterns. Transp. Res. Part E Logist. Transp. Rev. 95, 282–298.

Sun, Junzi, Dijkstra, Tristan, Aristodemou, Constantinos, Buzetelu, Vlad, Falat, Theo, Hogenelst, Tim, Prins, Niels, Slijper, Benjamin, 2022. Designing recurrent and graph neural networks to predict airport and air traffic network delays. In: 10th International Conference for Research in Air Transportation. FAA & Eurocontrol.

Svozil, Daniel, Kvasnicka, Vladimir, Pospichal, Jiri, 1997. Introduction to multi-layer feed-forward neural networks. Chemometr. Intell. Lab. Syst. 39 (1), 43–62.

Thiagarajan, Balasubramanian, Srinivasan, Lakshminarasimhan, Sharma, Aditya Vikram, Sreekanthan, Dinesh, Vijayaraghavan, Vineeth, 2017. A machine learning approach for prediction of on-time performance of flights. In: 2017 IEEE/AIAA 36th Digital Avionics Systems Conference. DASC, IEEE, pp. 1–6.

Tu, Yufeng, Ball, Michael O., Jank, Wolfgang S., 2008. Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. J. Amer. Statist. Assoc. 103 (481), 112–125.

Vorage, Laurence, 2021. Predicting Probabilistic Flight Delay for Individual Flights using Machine Learning Models. MSc Thesis. Delft University of Technology.

Wang, Liya, Tien, Alex, Chou, Jason, 2022. Multi-airport delay prediction with transformers. In: AIAA AVIATION 2022 Forum. p. 3707.

Ye, Bojia, Liu, Bo, Tian, Yong, Wan, Lili, 2020. A methodology for predicting aggregate flight departure delays in airports based on supervised learning. Sustainability 12 (7), 2749.

Yu, Bin, Guo, Zhen, Asian, Sobhan, Wang, Huaizhu, Chen, Gang, 2019. Flight delay prediction for commercial air transport: A deep learning approach. Transp. Res. Part E Logist. Transp. Rev. 125, 203–221.

Zoutendijk, Micha, Mitici, Mihaela, 2021. Probabilistic flight delay predictions using machine learning and applications to the flight-to-gate assignment problem. Aerospace 8 (6), 152.