Journal of the Air Transport Research Society 4 (2025) 100062

Contents lists available at ScienceDirect

Journal of the Air Transport Research Society

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

Enhancing aircraft arrival transit time prediction: A two-stage gradient boosting approach with weather and trajectory features

Go Nam Lui [a,1], Chris HC Nguyen [a], Ka Yiu Hui [a,2], Kai Kwong Hon [b], Rhea P. Liem [a,*]

[a] The Hong Kong University of Science and Technology (HKUST), Hong Kong Special Administrative Region
[b] Hong Kong Observatory, Hong Kong Special Administrative Region

ARTICLE INFO

Keywords:
Aviation weather
Arrival transit time prediction
Machine learning
Air traffic management

ABSTRACT

Accurate aircraft arrival transit time predictions are critical for reliable, efficient airport traffic management. This task is made more challenging by the different airspace characteristics across airports. While recent data-driven models show promise, two key limitations remain, namely the exclusion of tactical arrival operations and inadequate weather consideration. In this study, we develop a two-stage gradient boosting framework for aircraft arrival transit time prediction, incorporating new weather and trajectory features. The framework decomposes the prediction problem into holding pattern classification and transit time regression, explicitly modeling operational decision-making processes. Specifically, we perform a case study on 58,378 arrival flights in 2018 at the Hong Kong International Airport (HKIA). We introduce several new features including Bayesian weather-induced traffic features, route-specific rainfall intensity metrics, and trajectory-based identifiers for Standard Terminal Arrival (STAR) assignments. Our results show that the proposed framework with these features significantly improves predictive accuracy, particularly under adverse weather conditions. The two-stage gradient-boosting framework achieves a 6.09 percentage point reduction in mean absolute percentage error (MAPE) under extreme weather scenarios. Our Bayesian weather-induced traffic features outperform the established ATMAP weather metric, demonstrating superior capability in capturing weather impacts on arrival times. This new framework provides a more comprehensive understanding of airspace characteristics. The use of data types that are commonly available in almost all airports in the feature derivation makes it possible to apply the same approach in other airports.

1. Introduction

Accurate estimated time of arrival (ETA) prediction can provide air traffic control officers (ATCO) and airport operators with essential information to optimize landing sequences and reduce the overall aircraft's airborne time. Such an optimization will improve arrival efficiency by reducing flight delays and fuel consumption, thereby bringing numerous benefits to stakeholders. ETA prediction requires estimating the *arrival transit time*, which is the time an aircraft spends at the terminal airspace prior to landing. The benefit of predicting arrival transit time accurately on improving efficiency was previously demonstrated by Jun et al. (2022), who developed a strategy for ATCO to reduce overall delays by shifting holding from the terminal area to *en route* airspace. However, arrival transit time prediction can be challenging as it involves randomness and complexity pertaining to the converging air traffic in the terminal airspace. These characteristics and patterns (which need to be realistically modeled for an accurate prediction) vary from airport to airport, depending on flight and weather patterns in that particular airspace.

Despite the growing body of literature, gaps remain in accurately predicting aircraft arrival transit time within the terminal airspace, mainly owing to the challenges in comprehensively modeling airspace characteristics as mentioned above. In particular, limitations still exist in accounting for tactical arrival operation (such as holding patterns) and weather variations in the prediction models. Indeed, the final approach phase has not been the main focus in flight-time or delay prediction researches. Most of existing studies have been focused more on the *en route* stage or the entire flight with origin–destination (OD) pairs, including delay propagation and air traffic network effect (Rebollo & Balakrishnan, 2014; Wang et al., 2022; Yu et al., 2019; Zhu & Li, 2021). Considering these limitations and the importance of accurate arrival

* Correspondence to: Department of Aeronautics, City and Guilds Building, Exhibition Road, Imperial College London, London SW7 2AZ, United Kingdom.
E-mail addresses: g.n.lui@lancaster.ac.uk (G.N. Lui), hcnguyenaa@connect.ust.hk (C.H. Nguyen), kyhuiaf@connect.ust.hk (K.Y. Hui), kkhon@hko.gov.hk (K.K. Hon), r.liem@imperial.ac.uk (R.P. Liem).
[1] Present address: Lancaster University Management School, Bailrigg, Lancaster LA1 4YX, United Kingdom.
[2] Present address: NATS, Southampton, Hampshire SO31 7AY, United Kingdom.

https://doi.org/10.1016/j.jatrs.2025.100062
Received 22 November 2024; Received in revised form 27 January 2025; Accepted 2 February 2025
Available online 10 February 2025
2941-198X/© 2025 The Author(s). Published by Elsevier Inc. on behalf of Air Transport Research Society. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

transit time prediction, we realize the need to improve the prediction model capability, particularly in capturing airspace-specific flight and weather characteristics.

This paper makes several contributions to the field of aircraft arrival transit time prediction by introducing new features and adapting established machine learning techniques to the unique requirements of air traffic management. Firstly, we conduct a comprehensive, data-driven feature investigation for arrival transit time prediction at the Hong Kong International Airport (HKIA), providing new insights into air traffic operations in Hong Kong. Secondly, based on a Bayesian weather impact model (Lui et al., 2022), we derive two new features that outperform the commonly used air traffic management airport performance (ATMAP) weather metric. These features quantify the weather-induced total number of delayed flights and the amount of delays in an hour, as will be further elaborated in Section 3.1.2. Thirdly, we analyze radar rainfall images along the aircraft's assigned route and derive suitable metrics to quantify the impact of heavy precipitation on arrival transit times, providing a more detailed and accurate representation of weather conditions affecting each flight. In addition to these weather-related features, we introduce trajectory-based features, specifically the identification of holding patterns and assigned Standard Terminal Arrival (STAR), which offer valuable insights into the actual flight paths and operational conditions experienced by aircraft during the arrival phase.

These comprehensive features are effectively used in the arrival transit time prediction with a newly developed two-stage gradient boosting framework. This framework decomposes the complex prediction problem into two sequential learning tasks, namely (1) a specialized holding pattern classifier that captures the binary decision process in air traffic management, and (2) a transit time regressor that incorporates both the predicted holding probability and other input features. This decomposition approach not only explicitly models the significant impact of holding patterns on arrival times but also allows for differential feature importance between holding decisions and transit time estimation. By leveraging the probability output from the holding pattern classifier as an additional feature in transit time prediction, our framework captures the uncertainty in operational decisions while maintaining the sequential nature of air traffic management processes, resulting in a more nuanced and operationally relevant prediction system.

This paper is structured as follows. In Section 2, we provide an overview of the current state-of-the-art of arrival transit time prediction model and their limitations. In Section 3, we introduce our methods, with a brief description of our new framework, feature engineering procedures, and the prediction models used. Section 4 introduces our case study pertaining to Hong Kong, including data description, data analysis, and case implementation. Our results are presented and discussed in Section 5, and Section 6 concludes our work.

2. Overview of the current state-of-the-art

Arrival transit time is defined as the duration between an aircraft's entry into terminal airspace and its touchdown on the runway. Specifically, we measure from the timestamp when the aircraft crosses the terminal maneuvering area (TMA) boundary until the timestamp of its actual landing on the designated runway. This measurement captures the complete terminal phase of flight, including any holding patterns, vectoring, or delay absorption procedures implemented during the arrival sequence. Traditionally, predicting aircraft arrival transit time inside the terminal airspace relies on deterministic physics-based models that incorporate aircraft type, wind conditions, and separation requirements (Nedell et al., 1990). However, these models do not account for uncertainties that are inherent in real operations, which can lead to inaccurate predictions. With the increasing availability of aviation data from real-world operation, data-driven methodologies have been widely applied in aircraft ETA/arrival transit time prediction inside the terminal airspace to better estimate the arrival time. Owing to this increased availability of data, improving arrival transit time prediction has been made possible by applying advanced machine learning and statistical technique (Sternberg et al., 2016). Regression model is one of the common methods to predict ETA and arrival transit time, by utilizing each aircraft's operation data before it enters into the terminal airspace (Dhief et al., 2020; Hong & Lee, 2015; Wang et al., 2018; Zhang et al., 2022). Some notable examples are discussed below.

Glina et al. (2012) used quantile regression forests (QRF) for the prediction and uncertainty quantification of aircraft landing times, with a case study at Dallas/Fort Worth International Airport. They presented the machine learning approaches' capability for ETA prediction in terms of predictive accuracy and computational performance. Kern et al. (2015) presented a method for enhancing aircraft ETA predictions with random forest (RF), with input features including general flight, weather, and air traffic information. Jie et al. (2019) investigated the potential of spatiotemporal clustering of ADS-B trajectory data in ETA prediction, based on nearly 3,000 flights within three minutes in the terminal area; the study was performed for multiple Chinese airports. Wang et al. (2018) proposed a hybrid machine learning method for ETA prediction. The model's first layer used principal component analysis and clustering techniques to cluster the flights. The second layer applied a multi-linear regression and multi-cell neural network for ETA prediction based on 8,677 flights at Beijing Capital International Airport (BCIA). Following this work, they further improved the model performance by including more prediction models and different preprocessing settings based on more data (12,775 flights) from the same airport (Wang et al., 2020). One of the most recent works on BCIA was presented by Ma et al. (2022), which offered a spatiotemporal neural network model for ETA that considers both trajectory patterns and time prediction. Other researchers constructed an ETA prediction study at Guangzhou Baiyun International Airport concerning wind profile and arrival sequence pressure (Gui et al., 2021; Zhang et al., 2022). Dhief et al. (2020) predicted the aircraft landing time at Singapore Changi Airport and assessed the feature importance of each factor via gradient boosting machine (GBM), RF, and extra trees. In their follow-up work, they developed an arrival flight delay mitigation strategy with the holding and TMA delay prediction using CatBoost (Jun et al., 2022). Recently, Wang et al. (2023) expanded the scope of their ETA prediction study from single airport to a multi-airport system, which shows the potential of further scope improvement in aircraft ETA prediction.

Despite the demonstrated effectiveness of data-driven arrival transit time prediction development, there are still rooms for improvement. Specifically, although arrival flights with maneuvers such as go-around (Dai et al., 2021; Dhief et al., 2022) and holding patterns (Jun et al., 2022; Lui, Klein, & Liem, 2020) have been observed and studied, they are often removed from the dataset as outlier flights prior to training the prediction model in ETA prediction studies (Dhief et al., 2020; Jie et al., 2019; Liu et al., 2023; Wang et al., 2018, 2020). Whilst this approach is reasonable for airports with few tactical arrival operations such as holding patterns, it may not suffice when the arrival traffic conditions are more complex. Rather than excluding them, incorporating these maneuvers into model training can enable more realistic forecasts of terminal air traffic flows.

Furthermore, the limited inclusion of weather factors in current terminal airspace arrival time prediction studies represents a major gap. This is evident in our results that will be presented in Section 5.4, whereby more comprehensive representation of meteorological factors positively affects the forecast accuracy. Some past studies have incorporated wind profiles (Gui et al., 2021; Zhang et al., 2022), surface winds (Dhief et al., 2020), and precipitation (Kern et al., 2015). However, convection, which is a key driver of disruptions in terminal airspace, is frequently excluded (Dhief et al., 2020; Wang et al., 2018, 2020; Zhang et al., 2022) or only partially accounted for (Kern et al., 2015; Ma et al., 2022) in arrival time models. This

Table 1
Comparison of different works in terms of additional weather and trajectory consideration and dataset for the arrival transit time prediction task. (In the table, METAR refers to the meteorological aerodrome reports and ATMAP is as previously defined).

Literature	Additional dataset		Additional Consideration		
	METAR	Radar image	Weather	Standard arrival route	Tactical arrival operations
Glina et al. (2012), Jie et al. (2019)	No	No	No	No	No
Kern et al. (2015)	Yes	No	No dangerous phenomenon	No	No
Wang et al. (2018, 2020)	No	No	No	No	Removed as outliers
Dhief et al. (2020)	Yes	No	Surface wind only	No	Removed as outliers
Gui et al. (2021), Zhang et al. (2022)	No	No	Wind profile only	No	No
Ma et al. (2022)	No	Yes	Radar echo layer only	No	No
Jun et al. (2022)	Yes	No	ATMAP	Yes	Holding
This work	Yes	Yes	Weather-induced & radar-based features	Yes	Holding

deficiency constitutes a critical gap because convective weather conditions such as thunderstorms frequently necessitate major deviations from scheduled flight plans. A recent study by Jun et al. (2022) used ATMAP weather score developed by Eurocontrol (Eurocontrol, 2011) as a weather input. However, as an expert-based algorithm developed for European airports, ATMAP's generic weather scoring approach does not effectively capture localized meteorological impacts on arrival efficiency at specific airports like HKIA, as it lacks airport-specific tuning and validation (Lui et al., 2022). A more systematic, comprehensive approach is needed to integrate relevant weather conditions into arrival transit time prediction models.

Realizing the limitations of existing models encourages us to develop a framework for arrival transit time prediction that can comprehensively incorporate more features related to weather and trajectory and is aligned with ATCO's decision making process. For clearer comparison, Table 1 summarizes the key consideration in this work, along with those of some important representative literature. In addition, we also seek to investigate the values and benefits of including them in the regression model training to properly account for weather conditions and tactical arrival flight operations. Our methods, particularly on the new feature derivation, will be described in the next section.

3. Methodology

The derivation of our methodology is primarily driven by the realization that while holding patterns can significantly impact arrival transit times, it is challenging to model the relationship, which is complex and nonlinear. As such, rather than treating holding pattern occurrence as a simple binary feature (which is impractical since holding cannot be known at the time an aircraft enters the terminal airspace), we propose to model its probability instead, and includes it in the prediction task via a two-stage gradient boosting framework. In particular, this framework first predicts the likelihood of holding patterns using a dedicated classifier, then incorporates this probabilistic prediction alongside other features to estimate arrival transit times. This decomposition allows our model to capture both the direct impact of operational decisions (through holding pattern prediction) and their uncertainty (through holding probabilities), while maintaining the sequential nature of air traffic management processes.

Fig. 1 illustrates the new two-stage gradient boosting framework for arrival transit time prediction, along with the complete data processing pipeline. The framework begins with multiple raw data sources (shown in database icons on the left), including METAR data, flight information, radar images, ADS-B trajectory data, and STAR configuration. These raw data undergo three parallel feature extraction processes, namely (1) a weather impact model that derives weather-induced traffic features including the number of delayed flights and summation of arrival delay time, (2) radar image feature extraction that computes mean critical rainfall amount, and (3) trajectory feature extraction that processes STAR information. Additionally, baseline input features (shown within the blue-framed box under the "input features" column) are also considered; these include standard entry state parameters such as latitude, longitude, altitude, heading, groundspeed, descent rate, and hour, which have been commonly used in previous studies (Dhief et al., 2020; Gui et al., 2021; Wang et al., 2018, 2020; Zhang et al., 2022).

The extracted features are then fed into the two-stage prediction framework. Upon completing data preprocessing, a gradient boosting classifier predicts the likelihood of holding patterns for each flight in Stage 1 (Classification). This probabilistic prediction, alongside other features, then serves as input to Stage 2 (Regression), where a gradient boosting regressor predicts the final arrival transit time. This sequential approach allows the model to first assess the probability of significant operational events (holding patterns) before making the final transit time prediction, mirroring the actual decision-making process in air traffic management.

3.1. Feature extraction

In this section, we introduce the feature extraction procedure for the trajectory and weather features in our prediction framework.

3.1.1. Trajectory features

For this part, we wish to detect the corresponding arrival route (based on STAR) and holding pattern operations from data. To do so, we first trim raw arrival flight data to contain only points that are within the local area, such that all flight data are within the same geometric boundary. To improve computational efficiency, it is necessary to reduce the number of points that represent a flight trajectory, while preserving the geometric characteristics of the flight path. We achieve this by using the Ramer–Douglas–Peucker (RDP) algorithm (Douglas & Peucker, 1973). Based on these simplified data, the required detection procedures are performed, as summarized in Fig. 2.

Standard Terminal Arrival (STAR). Our STAR detection procedure leverages the dynamic time warping (DTW) algorithm to measure the similarity between flight trajectories and STAR routes. For two temporal sequences $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_n)$, DTW computes an optimal alignment via a minimization problem, as shown below:

$$DTW(X, Y) = \min_{\phi} \sum_{i=1}^{T} d\left(x_{\phi_x(i)}, y_{\phi_y(i)}\right), \tag{1}$$

where $\phi = (\phi_x, \phi_y)$ represents the warping path and $d(\cdot, \cdot)$ is the Euclidean distance between points. For each flight trajectory F, we determine its corresponding STAR s^* through:

$$s^* = \arg\min_{s \in S} DTW(F, s), \tag{2}$$

where S represents the set of available STAR routes. The efficiency of the STAR detection is enhanced through the following process. First, trajectory points are filtered using the standard altimeter setting region to reduce the dataset size. Second, we identify the closest entry points between the filtered flight trajectory and STAR routes, providing the most suitable starting points for the DTW algorithm. This dual approach notably reduces both the number of points to be compared and the warping path search space.

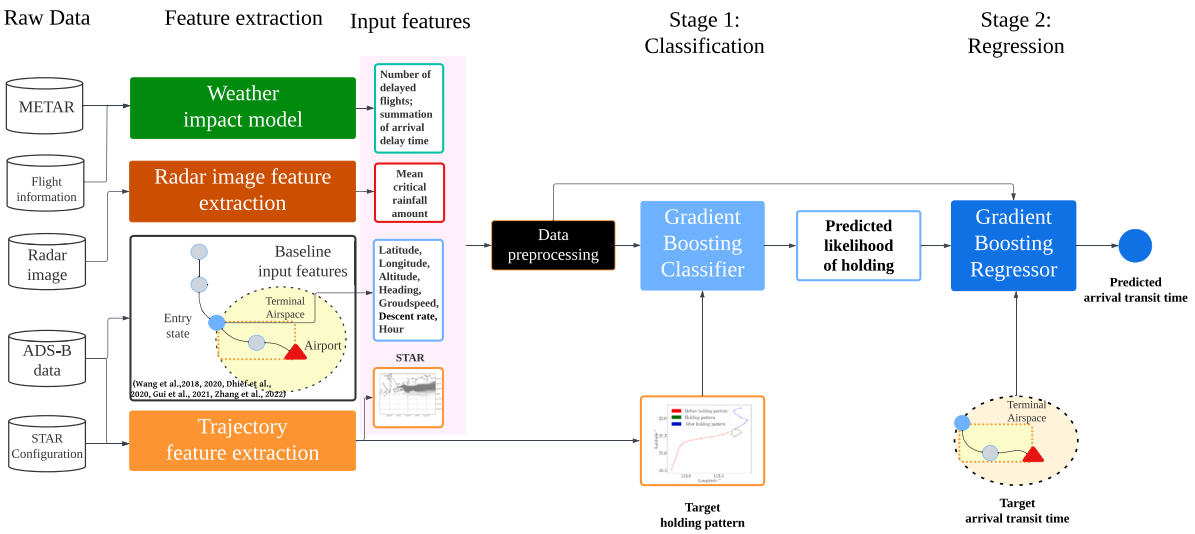


Fig. 1. The two-stage gradient boosting framework incorporating multi-source data preprocessing (METAR, flight information, radar images, ADS-B data, and STAR configuration), feature extraction processes (weather impact model, radar image analysis, and trajectory features), holding pattern likelihood prediction (Stage 1: Classification), and final arrival transit time estimation (Stage 2: Regression).

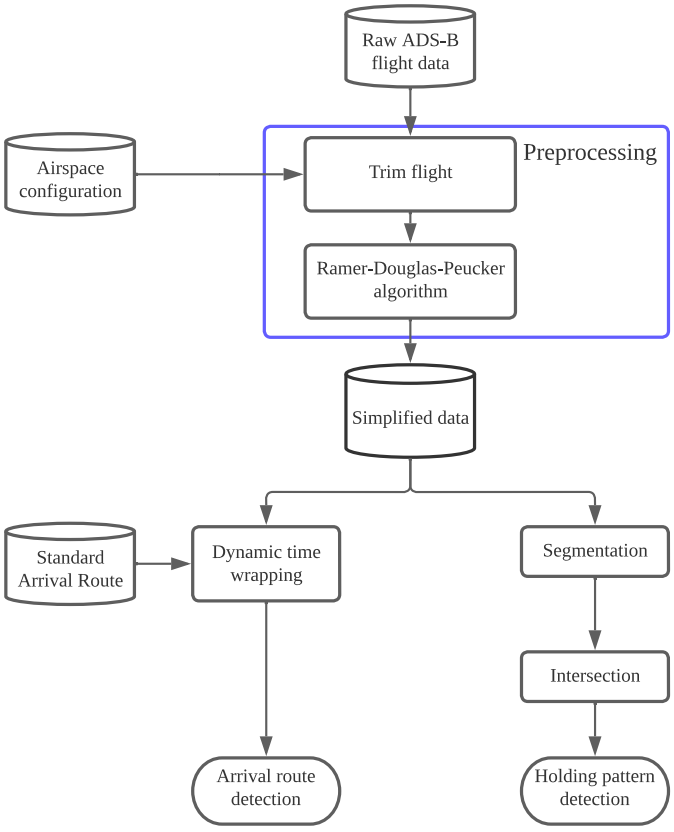


Fig. 2. Trajectory-based feature extraction procedure for STAR detection and HP detection.

Holding pattern (HP). The holding pattern detection procedure used in this study employs a geometric self-intersection analysis on trajectories that are already simplified using the RDP algorithm (based on a work by Lui, Klein, and Liem (2020)). Our method processes the original trajectory $T = \{v_1, \ldots, v_n\}$ to generate line segments $L_i = \overline{v_i v_{i+1}}$, represented as the lines connecting two adjacent red points in Fig. 3(a). The detection algorithm evaluates all pairwise segment combinations $(L_i, L_j) \in L \times L$. For each valid pair, spatial intersection is detected via the counter-clockwise orientation test, which evaluates whether points from one line segment lie on the opposite side of the other line segment, and vice versa.

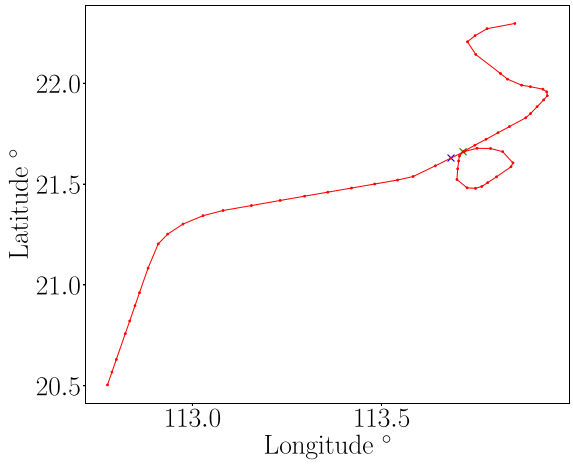
The intersection points, if found, are labeled and used to identify the starting and ending points of an HP, as shown by the blue and green crosses in Fig. 3(a). Once the starting and ending points are identified, we can separate parts of the flight trajectory corresponding to pre-holding, holding, and post-holding phases, as shown in Fig. 3(b). This procedure extracts a binary holding feature (1 = holding, 0 = otherwise) that will serve as the target variable for our first-stage prediction model, which will be further elaborated in Section 3.2.1.

3.1.2. *Weather features*

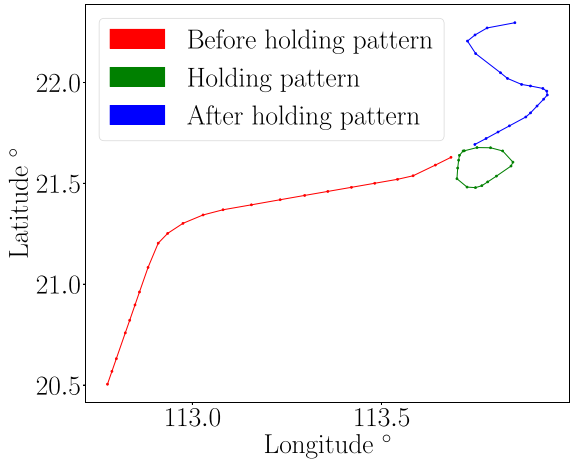
In this section, we explain the procedures to obtain weather-based features. In particular, the weather-induced traffic features described are derived upon a Bayesian-based weather impact model developed by Lui et al. (2022). In addition, we also introduce features based on radar rainfall images.

Weather-induced traffic features. Weather-induced traffic features considered in this study are derived based on meteorological aerodrome reports (METAR) data and flight information data. METAR provides hourly weather reports for an area enclosed within a 16 km radius around an airport. For ATM purposes, raw METAR data offer a lengthy string containing a variety of weather components, such as visibility, moisture, wind, etc. Flight information data used here refer to records of aircraft's actual and scheduled operations. ATMAP weather score is a conventional metric to quantify the adverse level of weather from METAR data (Eurocontrol, 2011). ATMAP was created by Eurocontrol in 2009 with the main purpose of evaluating airport performance. There are five components of the ATMAP weather score, namely wind, visibility, precipitation, freeze condition, and dangerous phenomenon. Yet, the impact of dangerous phenomenon on traffic performance is hard to quantify due to the inherent ATMAP scoring mechanism.

To account for dangerous phenomena specified in METAR data, separate models are derived for data containing dangerous phenomena, depending on the target airport. For instance, we need to include thunderstorms, shower, and cumulonimbus to fit the Hong Kong situation. Different airports might have different critical dangerous phenomena, such as volcanic ash, which needs further examination. However, it is important to note that convective weather (including thunderstorms, showers, and cumulonimbus clouds) is a predominant hazard for the majority of airports worldwide, albeit with different scales. As such, the



(a) Intersection detection.



(b) Holding pattern segmentation.

Fig. 3. Automatic holding pattern detection based on segmentation and intersection.

case study of HKIA presented herein serves as a representative example of the challenges faced by most airports in terms of weather impact on air traffic management. For an effective implementation of our model in other airports, information specific to the airport of interest should be considered and applied.

Using a Bayesian-based weather impact model (Lui et al., 2022), we can obtain the corresponding functions for normalized mean arrival delay per hour ($\tilde{\mu}$) and delay rate per hour (RT_d) as functions of the weather score x (by excluding the dangerous phenomena), for a given set of ATMAP weather scores and flight information data,

$$\tilde{\mu} \sim \mathcal{N}\left(\mathcal{F}\left(x\middle|\boldsymbol{\theta}_{\mu}\right), \sigma_{\mu}^{2}\right), \tag{3}$$

$$RT_d \sim \mathcal{N}\left(\mathcal{F}\left(x\middle|\boldsymbol{\theta}_{d}\right), \sigma_{d}^{2}\right), \tag{4}$$

where $\mathcal{F}$ refers to the mean trend function and θ presents the associated local parameters. The Gompertz function is used as the mean trend function $\mathcal{F}$, which was found to be the most accurate model describing the air traffic situation in Hong Kong in terms of the expected log pointwise predictive density (ELPD). The analysis, which was presented in the original work by Lui et al. (2022), was performed by comparing five trend functions (logistic, Gompertz, power, quadratic, and linear). For other airports, an evaluation based on ELPD is required for the mean trend function selection.

Actual values of $\tilde{\mu}$ and RT_d used to derive the models can be obtained from historical flight information data. To obtain $\tilde{\mu}$, we first calculate the average aircraft arrival delay (μ_{AD}) based on N flights, as shown below,

$$\mu_{AD} = \frac{1}{N}\sum_{f=1}^{N}(AAT_f - SAT_f)^{+}. \tag{5}$$

Each occurrence of *aircraft arrival delay* is defined as the difference between *scheduled arrival time (SAT)* and *actual arrival time (AAT)* for each flight. The notation f indicates the flight index and the superscript $^{+}$ denotes the non-negativity of the metric. Once μ_{AD} values are calculated, $\tilde{\mu}$ can then be obtained by normalizing μ_{AD},

$$\tilde{\mu} = \frac{\mu_{AD} - \min(\boldsymbol{\mu_{AD}})}{\max(\boldsymbol{\mu_{AD}})}. \tag{6}$$

Meanwhile, RT_d is obtained by calculating the total number of delayed flights,

$$RT_d = \frac{1}{N}\sum_{f=1}^{N} D_f, \text{ where } D_f = \begin{cases} 1 & \text{when flight } f \text{ is delayed;} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

To determine the delay indicator D_f, a flight is considered delayed when it cannot arrive within 15 minutes of the scheduled time (Mueller & Chatterji, 2002), following the documentation from the Federal Aviation Administration (FAA). Upon deriving the functions shown in Eq. (6) and Eq. (7), we can obtain parametric expressions of $\tilde{\mu}$ and RT_d as functions of ATMAP weather score, which will in turn be used to derive weather-induced traffic features used in the arrival transit time regression models. Recall that we also need to derive an additional model for each dangerous phenomenon detected in the target airport. Hence, we can obtain the corresponding $\tilde{\mu}$ and RT_d values for any weather conditions, with or without dangerous phenomena.

Both $\tilde{\mu}$ and RT_d are dimensionless quantities. To produce suitable input features for arrival transit time prediction models (*i.e.*, the present work), we need to associate the current hourly flight information data to generate a meaningful indicator. For the t-th hour, there are N flights arriving at the terminal airspace. Based on the definition of RT_d, we can infer *the number of weather-dependent delayed flights* N_d for the t-th hour as the product of N and RT_d:

$$N_d^t = \left(N \times RT_d\right)_t. \tag{8}$$

This is the first weather-induced traffic feature we can obtain from the weather impact model. Within the same hour, *the summation of weather-induced arrival delays* (S_{AD}) for all flights is used as the second weather-induced feature. Technically, this value should be obtained by taking the summation of delay time of all arrival flights in the t-th hour (δ_f^t, $f = 1, \ldots, N$). To simplify the calculation, we approximate S_{AD} by multiplying N_d^t (which is obtained using Eq. (8)) and μ_{AD}^t (from Eq. (5)), instead of using the actual delay time for each flight (which requires extra processing). This calculation is shown below,

$$S_{AD}^t = \sum_{f=1}^{N}\delta_f^t \approx N_d^t \times \mu_{AD}^t. \tag{9}$$

The derivation of N_d and S_{AD} and their usage in arrival transit time prediction models constitute some original contributions of the present work. The importance of deriving these two weather-induced traffic features (N_d^t and S_{AD}^t), instead of using ATMAP scores of METAR data directly in regression models, will be discussed and presented in Section 5.4.

Radar rainfall image feature. To include the impact of heavy precipitation in the arrival transit time prediction model, we introduce a feature based on rainfall intensity, namely the mean critical rainfall amount μ_R. This quantity is obtained from the radar rainfall image and the STAR information.

Before explaining the derivation of μ_R, we define the relevant radar image parameters that will be used in the description. Some of these

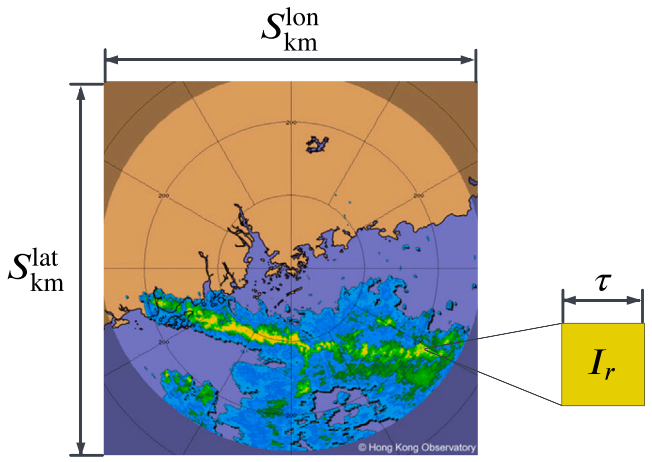


Fig. 4. Image parameters for μ_R derivation, shown on an example radar rainfall image from the Hong Kong Observatory (HKO).

parameters are illustrated in Fig. 4. First, we define the image's area as $S^{lon}_{km} \times S^{lat}_{km}$, where S^{lon}_{km} and S^{lat}_{km} correspond to the covered distance (in km) in the longitudinal and latitudinal direction, respectively. The corresponding image resolution is given as $N^{lon}_{pix} \times N^{lat}_{pix}$, indicating the number of pixels in the image, where N^{lon}_{pix} and N^{lat}_{pix} refer to the number of pixels in the longitudinal and latitudinal direction, respectively. Additionally, we refer to the rainfall intensity at individual pixels as I_r, as shown in Fig. 4.

For square radar rainfall images, where $S_{km} = S^{lon}_{km} = S^{lat}_{km}$, we also have the same number of pixels on both longitudinal and latitudinal directions,

$$N_{pix} = N^{lon}_{pix} = N^{lat}_{pix}. \tag{10}$$

Based on these parameters, we can define τ as the length in kilometer per pixel,

$$\tau = \frac{S_{km}}{N_{pix}}. \tag{11}$$

To obtain μ_R, we then define the *critical area* by setting a distance R_{km} around each waypoint (along the STAR route) to frame a 2-D square geospatial area. The framed area around a waypoint corresponds to the critical area for that particular waypoint. Converting the information contained within radar images into geospatial data is highly dependent on radar resolution. Based on the radar image resolution τ, we convert a radius in kilometers (R_{km}) to its equivalent in pixels:

$$R_{pix} = \lfloor \frac{R_{km}}{\tau} \rceil, \tag{12}$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, as the number of pixels must be a whole number. Assume that for arrival flight f, STAR_m with n waypoints are detected based on the procedure described in Section 3.1.1. Including the airport, there are $n+1$ essential waypoints for this particular flight. For an individual waypoint, we assign a pixel based on its coordinate. Since our critical area is defined as a square, the total number of critical pixels around each waypoint can then be computed as:

$$P = \left[2R_{pix} + 1\right]^2. \tag{13}$$

The side length of one critical area is $2R_{pix} + 1$, based on the definition given in Eq. (12) and includes the center pixel. For all waypoints, P is the same since we assume the same R_{km} and R_{pix}. Thus, the mean critical rainfall amount μ_R can be computed by averaging the rainfall intensity in all critical areas of all waypoints,

$$\mu_R = \frac{1}{P(n+1)} \left[\sum_{k=1}^{n+1} \sum_{j=1}^{P} \left(I_r\right)_{j,k} \right], \tag{14}$$

where $\left(I_r\right)_{j,k}$ refers to the rainfall intensity for the j-th pixel of the k-th waypoint.

3.2. Two-stage gradient boosting framework

Using the derived input features presented above, we now describe the new two-stage gradient boosting framework that can estimate the arrival transit time (in Stage 2) after predicting the holding pattern likelihood (in Stage 1). This approach is devised after some careful consideration. A single-stage model that directly predicts transit time would fail to explicitly account for the binary nature of holding pattern decisions. Similarly, treating holding patterns as a simple binary feature is impractical since holding cannot be known at the time an aircraft enters the terminal airspace, as mentioned in the opening of Section 3. Our proposed framework addresses these limitations by modeling holding patterns probabilistically and incorporating this uncertainty into the final transit time prediction.

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote our training feature matrix with n samples and d features, combining entry state features $\mathbf{X}_{\text{entry}}$ (latitude, longitude, altitude, heading, groundspeed, descent rate, and hour), operational features $\mathbf{X}_{\text{op}}$ (STAR), and weather features $\mathbf{X}_{\text{weather}}$ (weather score and mean rainfall):

$$\mathbf{X} = \left[\mathbf{X}_{\text{entry}}, \mathbf{X}_{\text{op}}, \mathbf{X}_{\text{weather}}\right]. \tag{15}$$

Our training dataset comprises 80% of full dataset, which accounts for 46,702 flights recorded for overall data, as will be further discussed in Section 4.1.2. Prior to the training process, the preprocessing pipeline handles mixed data types through one-hot encoding for categorical variables and standard scaling for numerical features.

3.2.1. Stage 1: Holding pattern likelihood prediction

The first stage employs a gradient boosting classification method to estimate holding pattern probabilities. This method is selected over alternatives (such as random forests or neural networks) due to its superior performance in handling imbalanced datasets and ability to capture complex feature interactions while maintaining interpretability.

Given a flight's feature vector $\mathbf{x}_f \in \mathbb{R}^d$ (*i.e.*, the transpose of the f-th row in the matrix $\mathbf{X}$ given in Eq. (15)), we define $y^h_f \in 0, 1$ to denote the binary holding indicator where 1 represents the presence of a holding pattern for flight f, and 0 when there is none. The gradient boosting classifier $M_h : \mathbb{R}^d \to [0, 1]$ estimates the probability of a holding pattern. For any flight feature vector $\mathbf{x}_f$, the model outputs:

$$M_h\left(\mathbf{x}_f\right) = p\left(y^h_f = 1 \middle| \mathbf{x}_f\right). \tag{16}$$

This model learns to minimize the binary cross-entropy loss function L_h, which is particularly suitable for capturing prediction uncertainty in binary classification tasks:

$$L_h = -\frac{1}{n} \sum_{f=1}^{n} \left[y^h_f \log\left(M_h\left(\mathbf{x}_f\right)\right) + \left(1 - y^h_f\right) \log\left(1 - M_h\left(\mathbf{x}_f\right)\right) \right]. \tag{17}$$

3.2.2. Stage 2: Arrival transit time prediction

The second stage leverages both the original features and the predicted holding pattern probability to improve transit time prediction accuracy. We construct an augmented feature matrix $\mathbf{X}_{\text{aug}} \in \mathbb{R}^{n \times (d+1)}$ by concatenating the original features (from Eq. (15)) with the holding probability predictions:

$$\mathbf{X}_{\text{aug}} = \left[\mathbf{X}, \hat{\mathbf{p}}\right], \tag{18}$$

where $\hat{\mathbf{p}} = \left[M_h\left(\mathbf{x}_1\right), \ldots, M_h\left(\mathbf{x}_n\right)\right]^\top \in \mathbb{R}^n$ represents the vector of predicted holding probabilities from Stage 1. Given a flight's augmented feature vector $\mathbf{x}_{\text{aug},f} \in \mathbb{R}^{d+1}$, let $y^t_f \in \mathbb{R}^+$ denote the arrival transit time for flight f. The gradient boosting regressor M_t predicts the arrival transit time:

$$\hat{y}^t_f = M_t\left(\mathbf{x}_{\text{aug},f}\right). \tag{19}$$

The model learns to minimize the mean squared error loss L_t:

$$L_t = \frac{1}{n}\sum_{f=1}^{n}\left(y_f^t - M_t\left(\mathbf{x}_{\text{aug},f}\right)\right)^2. \tag{20}$$

This two-stage architecture provides several advantages. First, it explicitly models the uncertainty in holding pattern decisions through probability estimates rather than binary predictions. Second, by incorporating these probabilities into the transit time prediction, the model can learn complex relationships between holding likelihood and transit time. Finally, this approach maintains the sequential nature of air traffic management decisions, where holding pattern assessments inform subsequent transit time estimates.

3.2.3. Model selection

To find the appropriate models for our framework, we evaluate and implement two state-of-the-art gradient boosting algorithms, namely XGBoost and LightGBM for both stages. Both algorithms extend the traditional gradient-boosted decision tree (GBDT) methodology, which iteratively constructs an ensemble of weak learners to create a robust predictive model (Friedman, 2001).

XGBoost employs an additive strategy where each new tree focuses on correcting the residual errors of previous iterations (Chen & Guestrin, 2016). This method optimizes both model performance and computational efficiency through second-order gradient statistics and advanced regularization techniques. For our holding pattern classification task, XGBoost's capacity to handle class imbalance through weighted objective functions is particularly advantageous.

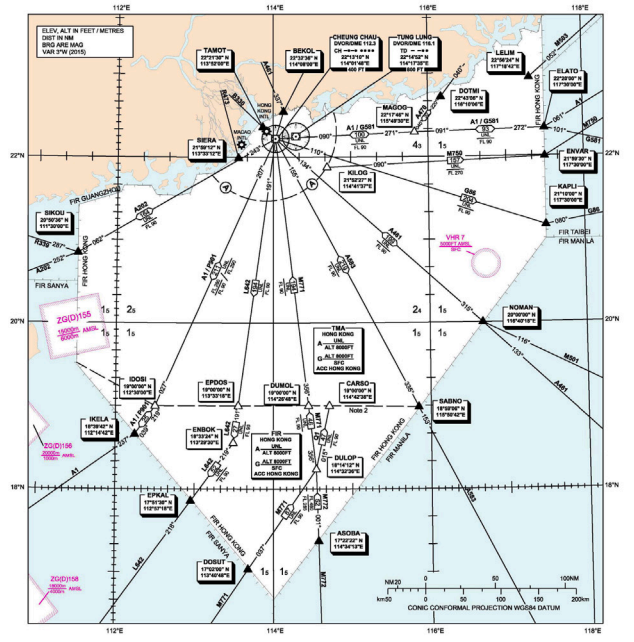
LightGBM offers complementary strengths with its leaf-wise tree growth strategy and gradient-based one-side sampling (Ke et al., 2017). These innovations result in reduced memory consumption and accelerated training speeds while maintaining competitive accuracy. The framework's efficient parallel training capabilities are especially valuable for our large-scale arrival management system.

4. Case study at the Hong Kong International Airport (HKIA)

To investigate the capability of our proposed additional features, we use HKIA as a case study. HKIA is one of the largest passenger hubs, gateways for various destinations, and transshipment centers in Asia and worldwide.[3]

Besides its high air traffic demand (Hon, 2021; Ng et al., 2017) and complex traffic mix (Hon et al., 2022), the unique airspace configuration surrounding HKIA makes it an interesting case study to test the generalizability of features in arrival transit time prediction models. Hong Kong airspace is located between the airspaces of Guangzhou, Macao, Zhuhai, and Shenzhen, in a way that HKIA is positioned at the north of the Hong Kong airspace. Fig. 5 shows the comparison between Hong Kong airspace and San Francisco airspace; the latter exhibits a circular airspace around the airport, which is more common. Note that the San Francisco airspace is shown here only for illustration and airspace shape comparison purposes; the analyses presented in this paper will be focused solely on the air traffic pertaining to HKIA, to maintain the focus and scope of the current paper.

Details of the case study at HKIA will be presented in this section. First, the data used in this study are described in Section 4.1, which include weather and flight data. The corresponding data analyses are presented in Section 4.2, to describe the characteristics of air traffic activities around HKIA. Next, Section 4.3 presents our feature importance analysis, which validates both our proposed framework and feature selection approach. Section 4.4 then details our experimental methodology, including comparative model evaluation and implementation specifications.



(a) Hong Kong airspace.

(b) San Francisco airspace.

Fig. 5. Airspace geometry comparison.

4.1. Data description

For our investigation, we use flight and weather data, which are described below. The STAR and airspace configurations used in this study are obtained from the Aeronautical Information Publication (AIP) Hong Kong 2019.[4] It is important to note that not all STAR procedures are active during a given day, and the active STAR may vary based on factors such as runway configuration, weather conditions, and air traffic control decisions. Given the large dataset spanning six months and the computational complexity involved in the analyses, the daily variation of STAR is not considered in the present study.

[3] HKIA at a glance. https://www.hongkongairport.com/iwov-resources/file/the-airport/hkia-at-a-glance/facts-figures/2021TC.pdf (in Chinese, last accessed on 22 November 2024).

[4] Hong Kong Aeronautical Information Services. https://www.ais.gov.hk/index.html (last accessed on 22 November 2024).

4.1.1. Weather data

We use two weather data sources in this study, which are METAR data at HKIA and 256 km high-resolution radar rainfall images collected and provided by the Hong Kong Observatory (HKO). Over 55,000 local METAR data (from 2017–2018) are used in this study for the weather feature derivation, which are obtained from https://navlost.eu. The observed dangerous phenomena at HKIA are thunderstorms, cumulonimbus, and shower rainfall, which will be included in the weather-induced traffic feature derivation.

To further evaluate the model performance under extreme weather, we use a weather score greater than or equal to 2 as the threshold for extreme weather scenarios. The selection of 2 refers to previous literature, in which 1.5 is commonly used as the standard for bad weather day (Eurocontrol, 2011; Murça et al., 2018; Schultz et al., 2018).

The ground weather radar images provided by HKO cover an approximately circular area of 256 km around Hong Kong. The resulting radar reflectivity values can serve as an estimate of the instantaneous rainfall rate. This information is freely available online.[5] This data source has been commonly used in several recent research on HKIA, due to its high resolution and availability (Liu et al., 2023; Lui, Liem, & Hon, 2020). In particular, 4,378 radar rainfall images are used in this study.

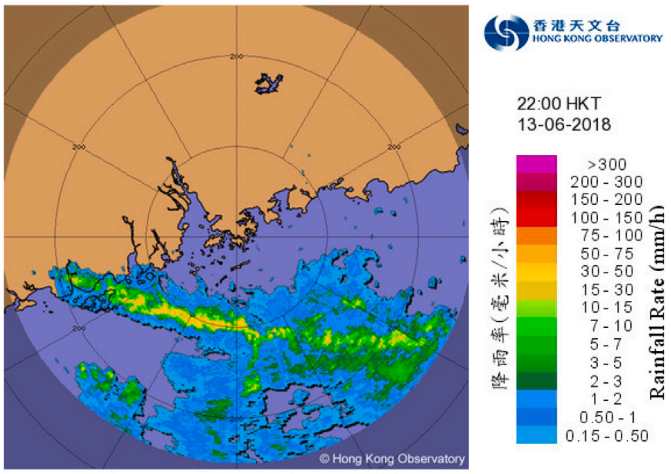
Fig. 6 shows an example of radar image and the corresponding extracted values, following the procedure described below. In order to obtain rainfall intensity values from the image, we start by associating the right color bar's rainfall intensity information with the corresponding RGB value (as shown in Fig. 6(a)). We then proceed to examine each rainfall image pixel individually to find the corresponding rainfall intensity value based on its RGB value. We achieve this by computing the square of the Euclidean distance of color vectors (in terms of their RGB values) between the pixel and the reference color bar to find the rainfall intensity value that corresponds to the minimum distance. The square of the Euclidean distance between RGB values is a standard metric for quantifying color difference. With this "color mapping" procedure, we can assign the corresponding rainfall intensity value for each pixel in the radar rainfall image.

4.1.2. Flight data

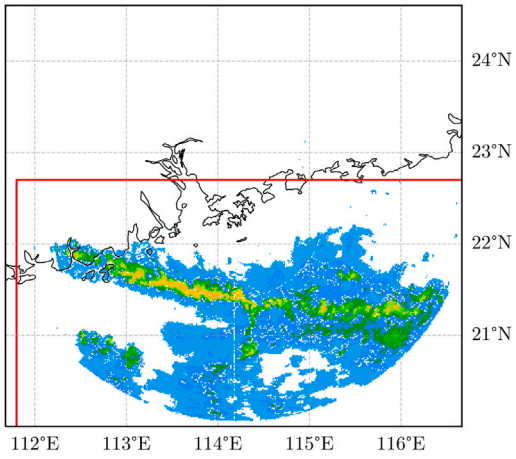
In this study, we use the automatic dependent surveillance-broadcast (ADS-B) technology from the *Flightradar24* platform (Flightradar24, 2022) as flight trajectory data. The period under review is from January 2018 to July 2018, which includes 58,378 arrival flights' ADS-B data. Among these flights, 2,348 (which account for 4.1% of the overall flight data) are under extreme weather based on the threshold described above. The real-time operation states and geographical location states are recorded in the ADS-B data. Flight information data, which contain the actual and scheduled arrival time of flights, are collected by the HKIA and are publicly available.[6] 433,680 arrival flight information are used for the weather impact model derivation.

4.2. Preliminary data analysis

Using the data described above, we perform a preliminary data analysis to characterize air traffic movements within the HKIA terminal airspace. The main purpose of performing these analyses is to identify tactical arrival operations and weather conditions pertaining to HKIA that will affect arrival transit time. In our study, we adhere to the operational patterns in HKIA airspace as closely as possible, to ensure



(a) Original radar image.



(b) Extracted value illustration.

Fig. 6. Illustration of the hourly rainfall intensity based on a radar image (22:00, 13th June 2018).

that our results reflect the actual operations in HKIA. This includes the selection of entry waypoints and the proportion of flights – including those entering holding patterns – at each waypoint.

Based on the number of flights observed in the flight trajectory data, Fig. 7 shows the proportion of arrival flights with holding patterns for four main entry waypoints to the HKIA STAR, including ABBEY, BETTY, CANTO, and SIERA. These entry points are illustrated in Fig. 8. The vertical bar charts shown in Fig. 7 are ordered from the most commonly used entry point to the least frequently used, and the proportion of flights with holding patterns is shown in each bar chart. The horizontal bar chart on top shows the proportion of arrival flights entering from each waypoint.

From Fig. 7, we can observe that most arrival flights enter the STAR through ABBEY, which is located east of the HKIA, while the least-used entry is CANTO. Operation-wise, CANTO can be treated as a "backup" option for regular operation entry. Most of the time, CANTO serves as one of the waypoints in the SIERA-series STAR. One interesting observation is that the proportions of holding pattern at the two least-used entries (BETTY and CANTO) can reach around 37%, which is higher compared to the two more frequently used entries. In general,

[5] Hong Kong Observatory Radar Image (256 km). https://www.hko.gov.hk/en/wxinfo/radars/radar.htm (last accessed on 22 November 2024).

[6] Flight schedule information of Hong Kong International Airport (Historical). https://data.gov.hk/en-data/dataset/aahk-team1-flight-info (last accessed on 22 November 2024).

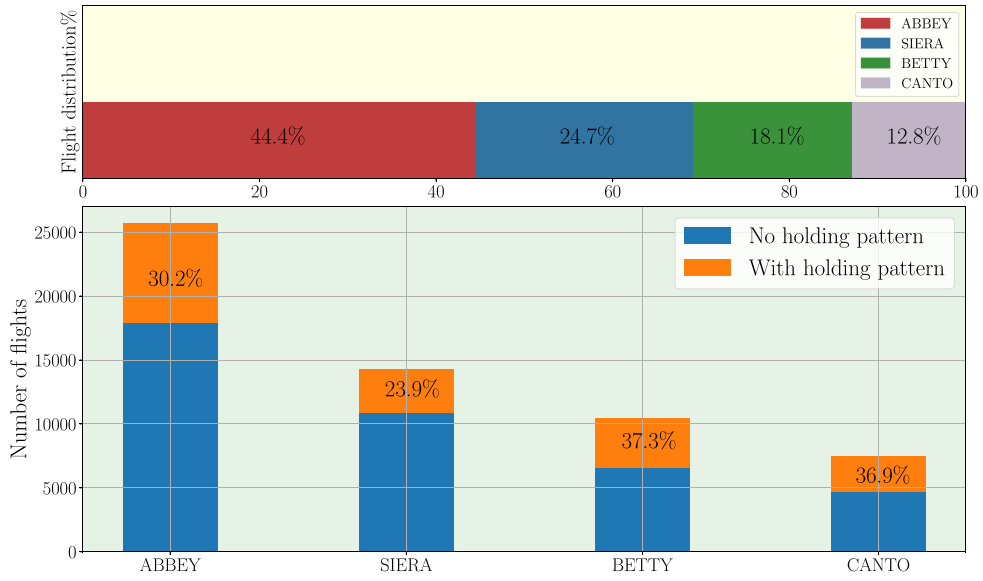


Fig. 7. Holding-pattern counts for different entries of HKIA STAR and their relative proportions.

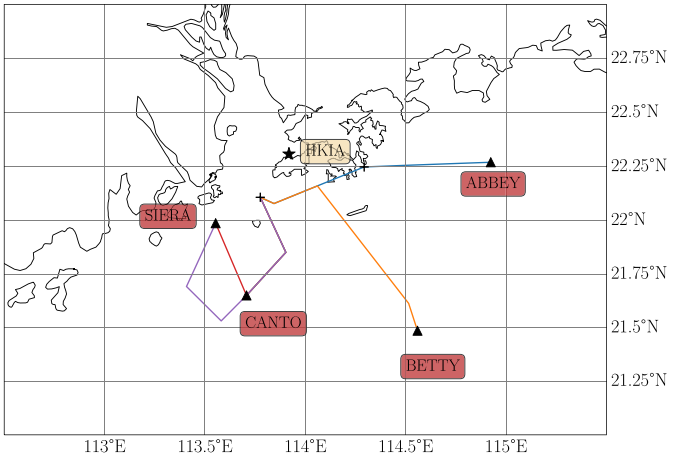


Fig. 8. Illustration of the STAR entries of Hong Kong airspace.

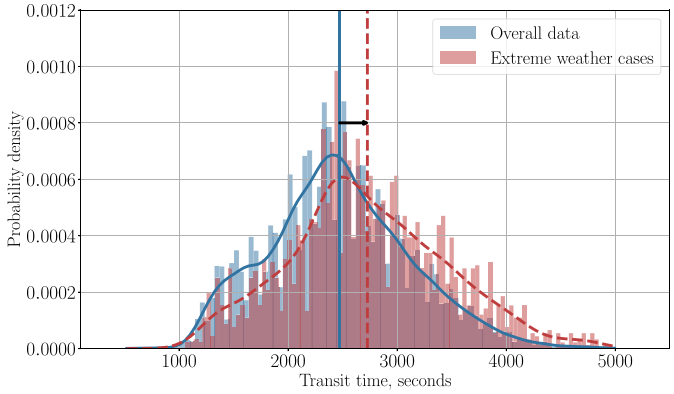


Fig. 9. The distribution of aircraft arrival transit time for overall data and data under extreme weather, where the mean shift in transit time due to extreme weather is indicated by the black arrow.

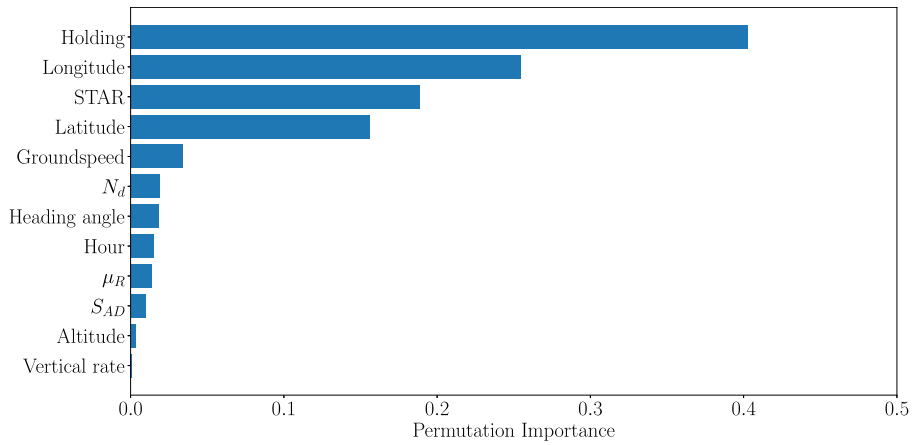
holding patterns are common occurrences within the HKIA terminal airspace, regardless of the entry point.

Next, we examine the relationship between extreme weather conditions and aircraft arrival transit time. Fig. 9 presents the distribution of arrival transit times for both the complete dataset (blue) and the subset of flights operating under extreme weather conditions (red). The x-axis represents transit time in seconds, whereas the y-axis shows the corresponding probability density. Our analysis reveals that on average, flights experience longer transit times under extreme weather conditions, with a mean of 2,726.76 s compared to 2,469.26 s for the overall dataset—a difference of approximately four minutes. Additionally, transit time variability increases under extreme weather, with the standard deviation rising from 25.58 to 26.90 s. These quantitative differences, despite the seemingly similar shapes of the distributions, underscore the importance of incorporating weather information in arrival transit time prediction models.

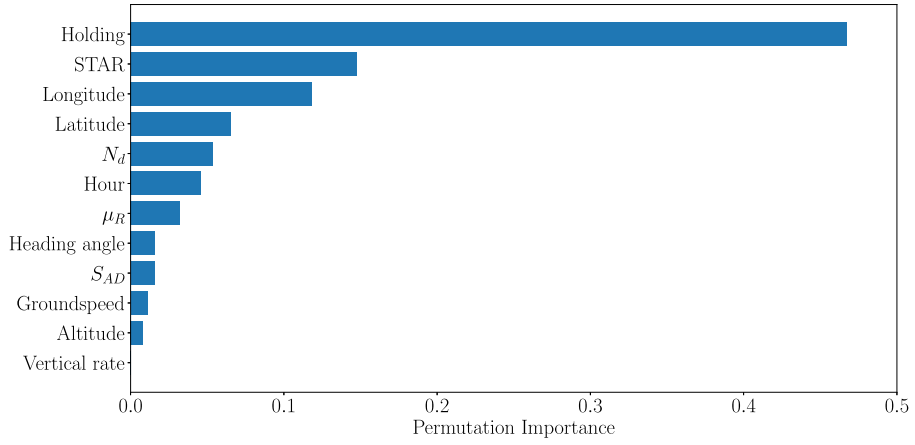
4.3. Feature importance analysis

In addition to preliminary data analysis, we also perform a feature importance analysis to gain insight into the factors that influence arrival traffic. In particular, we employ the *permutation feature importance*, which is a widely used method for evaluating the importance of features in tabular data (Fisher et al., 2019). This approach involves evaluating the effect of feature permutation on the prediction error, which enables the quantification of the relative importance of each input feature. We perform separate analyses for both overall and extreme-weather datasets using XGBoost, with the same 8:2 training-test split ratio used in our prediction study. Fig. 10 presents the results for both scenarios.

As Fig. 10(a) illustrates, trajectory-based features rank first and third as the most important features. The HP feature is essential among all inputs, with the highest permutation importance score of around 0.4. For the baseline features, entry longitude, latitude, and ground-speed have a more critical impact on arrival transit time for the overall



(a) Overall data



(b) Extreme weather cases

Fig. 10. The permutation feature importance through XGBoost on two studied datasets.

data. Except for vertical rate, all other input features have a positive impact. In summary, under normal circumstances, the 2-D geographical position information (STAR, longitude, latitude) and potential holding of entry aircraft dominate the prediction of arrival transit time. We observe that weather factors are not the most critical. Since most of the overall data are under good weather conditions, the weather impact is not evident in most cases. The importance of holding pattern indicator in arrival transit time prediction shown herein further confirms the need for its occurrence prediction, since holding pattern information is not known at the time of TMA entry. In our solution, this is achieved in Stage 1 of the framework, as described in Section 3.2.1.

The importance of weather-based features increases under extreme weather conditions. In particular, N_d and μ_R rank fifth and seventh in the list of the most important features, respectively (Fig. 10(b)). However, the magnitude of their impact remains limited, possibly due to the effectiveness of the strategies and operational adjustments deployed by Hong Kong ATCO. Given the frequency of convective weather in Hong Kong, ATCO is well-equipped to handle adverse weather situations by implementing measures such as rerouting, holding patterns, and adjusting aircraft separation distances, which help mitigate the impact of weather on flight operations and arrival transit times.

Another interesting change is the crucial sequence for entry groundspeed. Intuitively, a high correlation exists between mean ground speed and the aircraft's arrival transit time inside the terminal airspace. For typical cases, entry groundspeed is partially inferred to represent the speed situation inside the terminal airspace. However, groundspeed's low feature importance score drastically reduces under extreme weather. This phenomenon indicates the unstable speed profile for aircraft under adverse weather. ATCO tends to arrange tactical arrival operations for aircraft under extreme weather circumstances.

4.4. Experiment setup

To evaluate the performance of our proposed two-stage framework (using both LightGBM and XGBoost), we conduct experiments comparing it with other models. These include conventional machine learning methods such as Random Forest (Ho, 1995) and K-Nearest Neighbor (Altman, 1992), as well as deep learning algorithms including Deep Forward Networks and Long Short-Term Memory (LSTM), which were used in most recent arrival flight time prediction studies (Basturk & Cetek, 2021; Deng et al., 2023; Nguyen & Liem, 2025). Additionally, we compare our results against single-stage XGBoost and LightGBM models that do not incorporate a holding pattern prediction stage. However, it is important to note that deep neural network (DNN) may be prone to overfitting on smaller tabular datasets due to their high model complexity, which can contribute to fitting failures.

Including the scenario with only baseline inputs, we have six scenarios in total, which are illustrated in Fig. 11. Note that all models

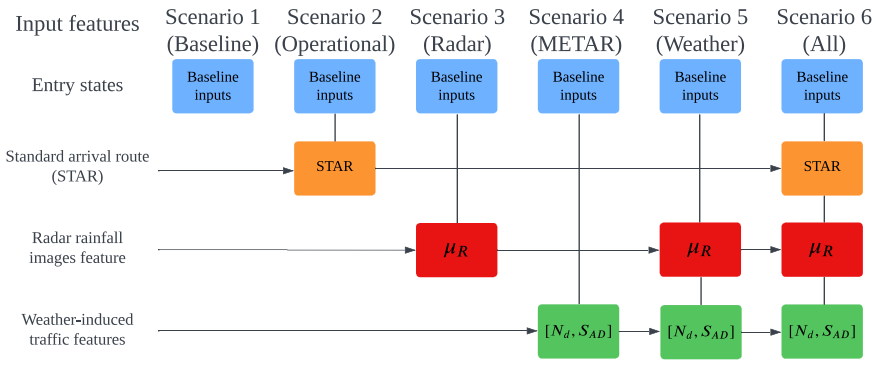


Fig. 11. Six scenarios for our model evaluation, where each scenario has a different set of input features.

mentioned above will be applied to all six scenarios. In choosing which scenarios to be included in the analyses, we perform some preliminary experiments and observations to carefully select representative scenarios that can capture the features' impact and most relevant interaction. As such, we can draw meaningful conclusions about the impact of different feature types on the model's predictive performance, without exhaustively presenting all possible feature combinations—where some might not offer meaningful results and insights.

To ensure the reliability of our results, we implement a comprehensive validation strategy. For each experiment, we first partition the data into training (80%) and test (20%) sets using different random seeds to ensure robust evaluation across multiple data splits. Within the training phase for each experiment, we apply a five-fold cross-validation for the holding pattern classifier in Stage 1, generating unbiased out-of-fold predictions that are used as features for the transit time regressor. In our machine learning pipeline, we employ separate preprocessing modules, including standardization using the z-score for numerical inputs and one-hot encoding for categorical features (STAR indicators). For hyperparameter tuning, we utilize the random search algorithm (Bergstra & Bengio, 2012), which can efficiently explore a large hyperparameter space. We perform a three-fold cross-validation and evaluate 100 different combinations to determine the optimal hyperparameters for each regression model.

The predictive accuracy of the models is evaluated using multiple complementary metrics. For holding likelihood prediction, we utilize accuracy, precision, and the area under the curve (AUC) metrics. Accuracy measures the overall correctness of predictions and is calculated as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Observations}}, \tag{21}$$

where True Positives and True Negatives represent correctly predicted positive and negative cases, respectively. Precision quantifies the proportion of correct positive predictions and is expressed as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}, \tag{22}$$

where False Positives are instances incorrectly predicted as positive. The AUC metric evaluates the model's ability to distinguish between classes and is defined as:

$$\text{AUC} = \int_0^1 \text{TPR}(s)\,\text{FPR}'(s)\,\mathrm{d}s, \tag{23}$$

where TPR represents the True Positive Rate and FPR represents the False Positive Rate at different classification thresholds s.

For the overall arrival transit time prediction assessment, we employ two standard metrics, namely the root mean squared error (RMSE) and mean absolute percentage error (MAPE). The RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{f=1}^{n}\left(y_f^t - \hat{y}_f^t\right)^2}, \tag{24}$$

where n represents the total number of observations, y_f^t denotes the actual arrival transit time value for the f-th flight, and $\hat{y}_f^t$ represents the corresponding predicted value, as introduced in Section 3.2.2. This metric effectively captures the model's prediction accuracy while giving higher weights to larger errors due to its quadratic nature. The MAPE provides a scale-independent measure of accuracy and is calculated as:

$$\text{MAPE} = \frac{1}{n}\sum_{f=1}^{n}\left|\frac{y_f^t - \hat{y}_f^t}{y_f^t}\right| \times 100\%. \tag{25}$$

This metric offers the advantage of expressing prediction errors in percentage terms, facilitating interpretation and cross-comparison across different scales. For evaluating individual flight predictions, we utilize the absolute error (AE), expressed as:

$$\text{AE}_f = \left|y_f^t - \hat{y}_f^t\right|. \tag{26}$$

This metric provides a direct measure of prediction accuracy at the individual flight level, complementing the aggregate performance metrics described above.

5. Results and discussion

In this section, we present a comprehensive evaluation of our methodology for predicting aircraft arrival transit times at HKIA. Our analyses are structured in four parts, beginning with the validation results of our machine learning models for holding likelihood prediction in Section 5.1. We then examine the general predictive performance across overall data and extreme weather cases (Section 5.2), analyze the impact of our proposed framework on individual flight predictions (Section 5.3), and finally assess the effectiveness of our weather impact model for feature extraction (Section 5.4).

5.1. Holding pattern likelihood prediction

Our initial validation focuses on comparing the predictive accuracy of different machine learning models (including DNN, LightGBM, Random Forest, and XGBoost) for holding pattern likelihood, as illustrated in Fig. 12. The evaluation metrics reveal XGBoost as the top performer, achieving an AUC of 0.827, accuracy of 0.759, and precision of 0.667. LightGBM demonstrates similarly strong performance, with an AUC of 0.823, accuracy of 0.755, and precision of 0.662, closely matching XGBoost's capabilities. While Random Forest and DNN models also perform sufficiently well, their metrics are notably lower than those of the gradient boosting approaches. Random Forest achieves an AUC of 0.798, accuracy of 0.727, and precision of 0.655, while DNN records an AUC of 0.793, accuracy of 0.733, and precision of 0.592. Despite DNN maintaining competitive accuracy and AUC scores, it has the lowest precision among all tested models.

Table 2
General predictive error comparison across eight models and six scenarios on overall data.

Model	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5		Scenario 6	
	RMSE[s]	MAPE[%]	RMSE[s]	MAPE[%]	RMSE[s]	MAPE[%]	RMSE[s]	MAPE[%]	RMSE[s]	MAPE[%]	RMSE[s]	MAPE[%]
Random Forest	408.26	13.04	375.09	11.39	399.47	12.86	397.50	12.72	394.70	12.70	361.77	11.10
kNN	418.10	13.24	370.08	10.95	409.38	13.04	403.67	12.85	396.39	12.61	367.92	11.01
XGBoost	400.39	12.70	354.80	10.53	386.37	12.37	378.50	12.17	369.96	11.98	336.00	10.21
LightGBM	402.20	12.71	354.73	10.51	392.69	12.60	379.30	12.27	372.79	12.11	338.45	10.30
DNN	455.17	14.97	405.09	12.45	453.20	14.94	418.07	13.30	415.67	13.43	371.34	11.02
LSTM	410.75	12.75	365.72	10.51	406.32	12.83	405.16	12.72	404.51	12.58	353.78	10.49
2S-LightGBM	402.63	12.74	354.30	10.52	389.71	12.45	373.66	12.00	364.76	11.77	323.15	9.81
2S-XGBoost	401.77	12.70	353.60	10.49	387.67	12.38	370.31	11.91	359.48	11.61	**318.17**	**9.66**

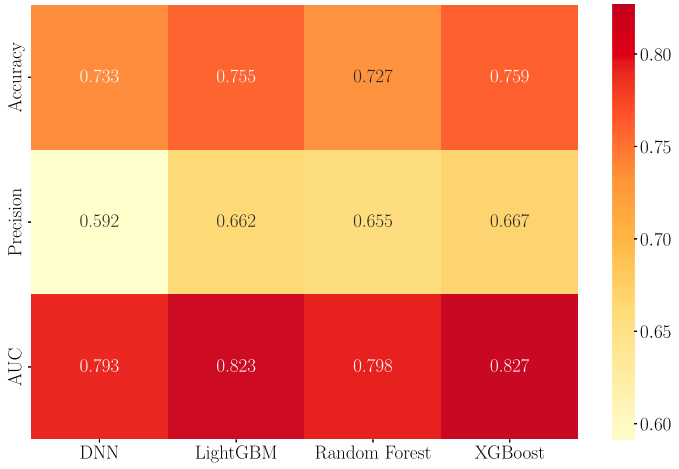


Fig. 12. The performance comparison of holding likelihood prediction.

The consistently high AUC scores across all models, ranging from 0.793 to 0.827, demonstrate robust discriminative capability in predicting holding likelihood. These results indicate that our selected features and modeling approaches (*i.e.*, the two-stage gradient boosting framework) effectively capture the underlying patterns in holding operations at HKIA, providing a solid foundation for subsequent analyses.

5.2. General predictive accuracy of arrival transit time

This section presents the general predictive accuracy for overall data and extreme weather cases for all six scenarios and eight models. As mentioned in Section 4.1.2, 2,348 arrival flights under extreme weather represent the extreme weather cases, which stand for 4.1% of the overall data. We test all models mentioned in Section 4.4, and evaluate predictive performance based on their RMSE and MAPE values for further investigation. The general predictive errors for overall data and extreme weather cases are presented in Tables 2 and 3, respectively. Our new two-stage methods are indicated by the label '2S' in the tables. The combinations of scenario and regression model that yield the minimum RMSE and MAPE values are highlighted in bold.

The proposed two-stage gradient boosting framework demonstrates remarkable advantages over traditional single-stage approaches, establishing its effectiveness across both normal and extreme weather conditions. Through the innovative approach of explicitly modeling holding pattern uncertainty via probability estimates before transit time prediction, our framework achieves consistently superior performance compared to conventional machine learning and deep learning methods. This improvement is particularly evident in scenario 6, *i.e.*, the most comprehensive testing configuration, where the 2S-XGBoost implementation achieves an outstanding RMSE of 318.17 s and MAPE of 9.66% under normal conditions, significantly outperforming all other models.

The framework's robustness becomes even more apparent when examining performance under extreme weather conditions, where it maintains exceptional prediction accuracy with an RMSE of 354.40 s and MAPE of 10.15%. This stability stands in stark contrast to other models, particularly deep learning approaches such as LSTM, which exhibit substantial performance degradation with MAPE exceeding 60% in challenging weather conditions. These findings align with recent research indicating that tree-based methods continue to represent the state-of-the-art approach for medium-sized datasets, specifically those with sample sizes in the order of $\mathcal{O}\left(10^3\right)$ (Grinsztajn et al., 2022).

The sustained performance advantage of our two-stage architecture, especially under extreme weather scenarios, demonstrates its effectiveness in capturing the complex, sequential nature of air traffic management decisions. By incorporating holding pattern probabilities into the final transit time predictions, the framework delivers more nuanced and reliable estimates. This enhanced accuracy has implications for improving operational efficiency and decision-making processes in air traffic management systems, particularly during adverse weather conditions when precise predictions are most crucial.

5.3. Predictive performance of individual flight

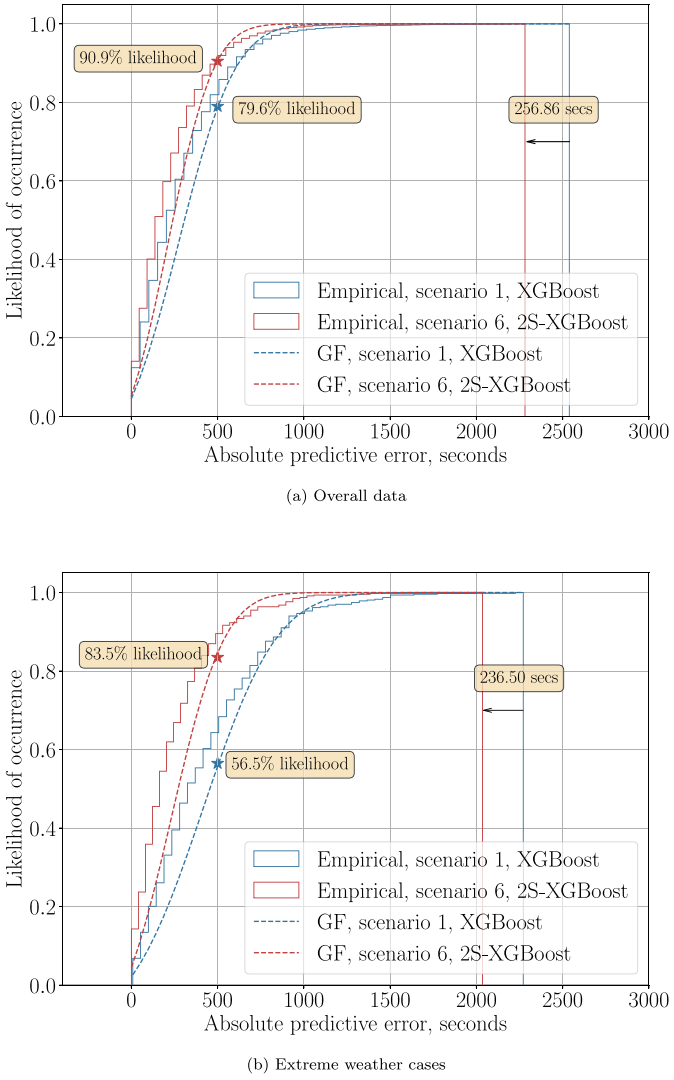
In this subsection, we aim to evaluate the predictive performance of individual flight using the cumulative distribution function for absolute error visualization. We also demonstrate the individual predictive performance of our proposed framework for both overall data and data under extreme weather. Fig. 13 shows the empirical cumulative histogram (denoted as "Empirical") and Gaussian fit (denoted as "GF") based on the statistics of the absolute errors, where the blue line represents scenario 1 with baseline inputs using XGBoost and the red line represents scenario 6 with our proposed framework.

As shown in Fig. 13, including the proposed additional input features into the new two-stage framework shows a notable predictive accuracy improvement. For overall data, using only the baseline inputs and XGBoost (the best among other single-stage model) yields a 79.6% likelihood that the absolute predictive error falls below 500 s, whereas the value can reach 90.9% with our new framework. Similarly, for the extreme weather case, the likelihood of occurrence increases from 56.5% to 83.5%.

Besides increasing the likelihood of achieving a lower predictive error, our new framework can also reduce the maximum absolute error for individual prediction. Comparing the maximum absolute error values for the overall data shows a 256.86 s reduction between scenarios 1 and 6, while the reduction is 236.5 s for the extreme weather case. There are noticeable shifts in the maximum prediction errors (in seconds) for both cases, which demonstrate the effectiveness of including the new features derived herein to reduce predictive errors for both individual flight and aggregate data, thereby highlighting the contribution and importance of the present work.

Table 3
General predictive error comparison across eight models and six scenarios for extreme weather cases.

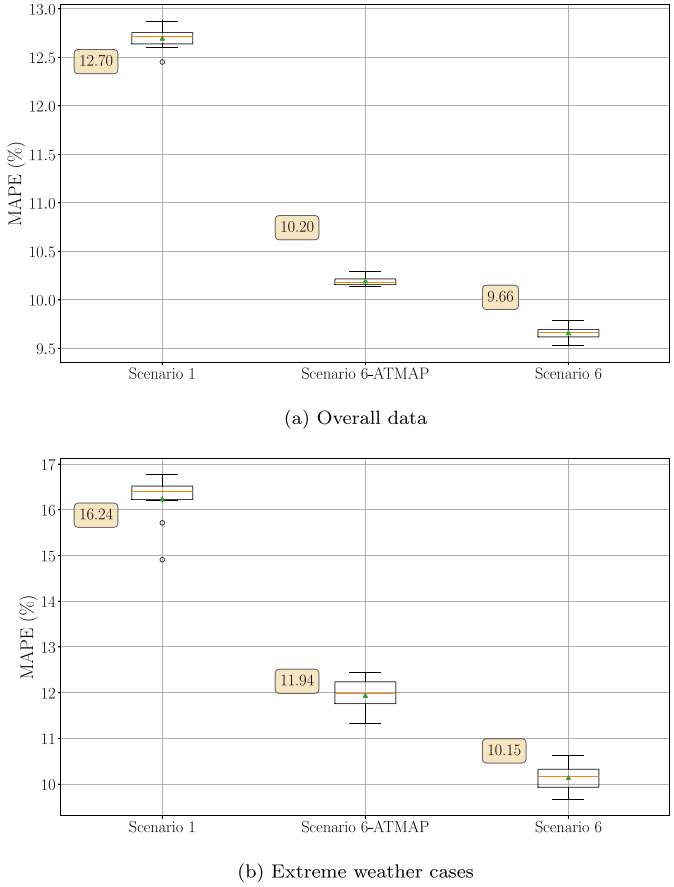
Model	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5		Scenario 6	
	RMSE[s]	MAPE[%]	RMSE[s]	MAPE[%]	RMSE[s]	MAPE[%]	RMSE[s]	MAPE[%]	RMSE[s]	MAPE[%]	RMSE[s]	MAPE[%]
Random Forest	549.68	16.85	509.68	14.67	498.07	15.36	451.91	13.88	447.36	13.80	443.23	13.31
kNN	546.84	16.91	519.58	15.04	531.82	16.09	494.14	15.26	490.55	14.83	495.91	14.38
XGBoost	509.68	14.67	479.21	13.80	414.37	12.59	369.06	12.04	367.71	11.01	376.49	10.92
LightGBM	522.97	15.88	477.28	13.91	426.30	13.02	389.36	12.10	371.75	11.01	375.09	10.90
DNN	705.93	20.15	627.79	18.00	697.95	20.11	677.44	19.59	665.83	19.63	599.75	17.52
LSTM	1961.36	64.58	1917.05	62.69	1903.46	62.07	2023.13	67.19	2108.03	70.79	1937.51	63.56
2S-LightGBM	533.49	16.18	488.02	14.24	426.60	12.70	391.87	11.76	366.76	10.81	354.85	10.27
2S-XGBoost	533.43	16.24	486.60	14.11	426.09	12.72	396.40	11.85	486.60	14.11	**354.40**	**10.15**



(a) Overall data

(b) Extreme weather cases

Fig. 13. The cumulative distribution function for individual absolute error on arrival transit time, using the single-stage XGBoost for scenario 1 (without holding pattern probability) and the two-stage XGBoost for scenario 6.



(a) Overall data

(b) Extreme weather cases

Fig. 14. The predictive error reduction of our proposed weather-induced features, the green triangles refer to the mean values.

5.4. Improvements by our weather-induced traffic features

In our study, N_d and S_{AD} are newly derived from a Bayesian-based weather impact model. However, the benefits of using these values, instead of raw weather information data, are yet to be examined. We conduct a control group study comparing three scenarios: baseline inputs only (scenario 1), all available inputs (scenario 6), and all inputs with weather-induced traffic features replaced by the ATMAP weather score (scenario 6-ATMAP). ATMAP weather score is selected here because it is a commonly used weather feature in other ATM studies (Jun et al., 2022; Murça et al., 2018; Schultz et al., 2018, 2021). Using 2S-XGBoost as the model, we perform our study on two datasets (overall and extreme weather cases). The evaluation metric is MAPE.

Results pertaining to the overall data show that, with appropriate application of the Bayesian-based weather impact model, our weather-induced traffic features perform better in increasing the arrival transit

time predictive accuracy than when using the conventional ATMAP weather score directly. As shown in Fig. 14(a), considering only the baseline inputs yields the highest MAPE, around 12.45% to 12.87%. Including ATMAP weather score, the mean MAPE reduces from 12.70% to around 10.20%. With our proposed weather-induced features, *i.e.*, N_d and S_{AD}, the predictive error of the model notably decreases the most. The approximate amplitude of error reduction (compared to the baseline case) is 3.04 percentage points.

For the extreme weather cases, the predictive error for the baseline case increases from 12.70% to 16.24%. Under this circumstance, the ATMAP weather score still performs well, reducing the MAPE by 4.3 percentage point. A more notable improvement is observed when using N_d and S_{AD} in the prediction model, which is similar to our observation of the overall data. The mean error is 10.15%, which is around 6.09 percentage point reduction from the baseline case, and the lowest bound for the boxplot can reach 9.67%.

To sum up, our proposed weather-induced features (N_d and S_{AD}) exhibit better performance in reducing the arrival transit time prediction error than the ATMAP weather score for both regular and extreme weather cases with our two-stage gradient boosting framework, thereby highlighting their importance in arrival transit time prediction. We believe that this is because the newly derived features have "translated" weather score into metrics related to aircraft delays, which are more relevant to arrival transit time prediction.

6. Conclusion

To the best of our knowledge, this study presents the first large-scale, data-driven feature investigation for arrival transit time prediction at HKIA that comprehensively addresses both general and extreme weather conditions. Our two-stage gradient boosting framework, which decomposes the prediction problem into holding pattern classification and transit time regression, demonstrates the value of modeling operational decision-making processes explicitly. The framework's effectiveness is enhanced by our newly-derived features, including the weather-induced traffic features that outperform traditional ATMAP scores, route-specific rainfall intensity metrics that capture localized weather impacts, and trajectory-based features that identify holding patterns and STAR assignments.

Our results demonstrated that the proposed framework with weather and trajectory features could notably improve arrival transit time prediction accuracy. Under extreme weather conditions, our new 2S-XGBoost approach reduced error (MAPE) by 6.09 percentage points compared to the baseline case. Additionally, the derived weather-induced traffic features were found to be more effective than the commonly used ATMAP weather score in modeling weather impact on arrival transit time within the HKIA terminal airspace. This was especially so under extreme weather conditions. The findings highlighted the importance of capturing real-world uncertainties from weather and trajectory deviations that were often neglected in previous studies.

Whilst promising, further expanding the consideration of trajectory features (such as vectoring and shortcut) in the modeling structure (including their predictions in Stage 1) and accounting for weather time-lag effects could further enhance model versatility for generalized use across airports. Implementing our proposed approach outside HKIA may require certain tuning and adjustment. Different airports have varying airspace structures, operational procedures, and weather patterns, which could influence the effectiveness of our model. The user's domain knowledge of the specific airport/airspace of interest will help ensure the effective implementation of our model. For future work, it will be interesting and insightful to demonstrate and validate our developed approach at other airports, taking into account their specific characteristics and constraints; this effort will require close collaborations with other researchers from different countries and access to suitable data. In addition, performing the feature investigation on other ETA prediction models beyond those analyzed in this study, such as new deep learning structures, can also further validate the effectiveness of the features introduced in this study. With continued refinement, we envision that these data-driven arrival time predictions will support improved traffic flow optimization through more efficient, weather-aware flight sequencing. Overall, this research provides valuable insights into leveraging airspace-specific flight patterns and meteorology information to increase prediction performance. Our approach helps move towards next-generation forecasting capabilities that consider the dynamics of the entire terminal airspace system.

Abbreviations

HKIA	Hong Kong International Airport
ATCO	Air traffic control officers
ETA	Estimated time of arrival
STAR	Standard Terminal Arrival
METAR	Meteorological aerodrome reports
N_d	The number of weather-induced delayed flights
S_{AD}	The summation of weather-induced arrival delay
ATMAP	Air traffic management airport performance
RDP	Ramer–Douglas–Peucker
DTW	Dynamic time warping
HP	Holding pattern
$\tilde{\mu}$	Normalized mean arrival delay per hour
RT_d	Delay rate per hour
$\mathcal{F}$	Mean trend function
θ	Local parameters
μ_{AD}	Mean arrival delay per hour
AAT	Actual arrival time
SAT	Scheduled arrival time
N	The number of flights per hour
δ	The amount of flight arrival delay
$S_{\text{km}}^{\text{lon}}$	Distance on the longitude side
$S_{\text{km}}^{\text{lat}}$	Distance on the latitude side
$N_{\text{pix}}^{\text{lon}}$	The number of pixels on the longitude side
$N_{\text{pix}}^{\text{lat}}$	The number of pixels on the latitude side
I_r	Rainfall intensity
τ	Length in km per pixel
R_{km}	Radius of the critical area (in km)
R_{pix}	Radius of the critical area (in the number of pixels)
P	The number of critical pixels around each waypoint
μ_R	Mean rainfall amount for critical area
n	The number of flights
d	Total number of features
$f \in 1, \ldots, n$	Flight index
$\mathbf{X} \in \mathbb{R}^{n \times d}$	Training feature matrix
$\mathbf{x}_f \in \mathbb{R}^d$	Feature vector for flight f
$y_f^h \in 0, 1$	The binary holding indicator for flight f
$y_f^t \in \mathbb{R}^+$	Arrival transit time for flight f
AUC	Area under the curve
AE	Absolute error
MAPE	Mean absolute percentage error
RMSE	Root mean squared error

CRediT authorship contribution statement

Go Nam Lui: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Chris HC Nguyen:** Software, Methodology, Formal analysis, Data curation. **Ka Yiu Hui:** Software, Methodology, Formal analysis. **Kai Kwong Hon:** Writing – review & editing, Resources, Conceptualization. **Rhea P. Liem:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Funding sources

This work is supported by the HKUST Start-Up Grant , Hong Kong Special Administrative Region (Project No. R9354) and the Innovation and Technology Commission (ITC) , Hong Kong Special Administrative Region (Project No. ITS/016/20). We would like to acknowledge the support from the University Grants Committee of the Hong Kong Special Administrative Region for providing financial support to the second author through the Hong Kong PhD Fellowship Scheme (Reference No. PF20-50039). The third author's Research Assistantship was supported by the Research Talent Hub under the Innovation and Technology Commission (ITC) , Hong Kong Special Administrative Region (Project No. InP/417/21).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are grateful for the weather data shared by the Hong Kong Observatory (HKO) and the flight data shared by Dr. Lishuai Li from City University of Hong Kong (under the data agreement signed by Dr. Lishuai Li and Dr. Rhea P. Liem). In particular, the flight data were obtained for a project funded by the Hong Kong Research Grants Council General Research Fund Grant (Project No. 11209717).

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185. http://dx.doi.org/10.2307/2685209.

Basturk, O., & Cetek, C. (2021). Prediction of aircraft estimated time of arrival using machine learning methods. *Aeronautical Journal*, *125*(1289), 1245–1259. http://dx.doi.org/10.1017/aer.2021.13.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281–305, https://dl.acm.org/doi/10.5555/2188385.2188395.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://dl.acm.org/doi/pdf/10.1145/2939672.2939785.

Dai, L., Liu, Y., & Hansen, M. (2021). Modeling go-around occurrence using principal component logistic regression. *Transportation Research Part C (Emerging Technologies)*, *129*, Article 103262. http://dx.doi.org/10.1016/j.trc.2021.103262.

Deng, W., Li, K., & Zhao, H. (2023). A flight arrival time prediction method based on cluster clustering-based modular with deep neural network. *IEEE Transactions on Intelligent Transportation Systems*, http://dx.doi.org/10.1109/TITS.2023.3338251.

Dhief, I., Alam, S., Lilith, N., & Mean, C. C. (2022). A machine learned go-around prediction model using pilot-in-the-loop simulations. *Transportation Research Part C (Emerging Technologies)*, *140*, Article 103704. http://dx.doi.org/10.1016/j.trc.2022.103704.

Dhief, I., Wang, Z., Liang, M., Alam, S., Schultz, M., & Delahaye, D. (2020). Predicting aircraft landing time in extended-TMA using machine learning methods. In *Proceedings of 9th International Conference for Research in Air Transportation*.

Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *10*(2), 112–122. http://dx.doi.org/10.1002/9780470669488.ch2.

Eurocontrol (2011). Algorithm to describe weather conditions at European airports. URL https://www.eurocontrol.int/sites/default/files/publication/files/algorithm-met-technical-note.pdf.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81, https://jmlr.org/papers/volume20/18-760/18-760.pdf.

Flightradar24 (2022). Data. URL https://www.flightradar24.com/data.com. (Accessed 3 July 2021).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, *29*, 1189–1232. http://dx.doi.org/10.1214/aos/1013203451.

Glina, Y., Jordan, R., & Ishutkina, M. (2012). A tree-based ensemble method for the prediction and uncertainty quantification of aircraft landing times. In *American Meteorological Society–10th Conference on Aritificial Intelligence Applications to Environmental Science, New Orleans, LA*.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, *35*, 507–520, https://dl.acm.org/doi/10.5555/3600270.3600307.

Gui, X., Zhang, J., Peng, Z., & Yang, C. (2021). Data-driven method for the prediction of estimated time of arrival. *Transportation Research Record*, *2675*(12), 1291–1305. http://dx.doi.org/10.1177/03611981211033295.

Ho, T. K. (1995). Random decision forests. *1*, In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (pp. 278–282). IEEE, https://dl.acm.org/doi/10.5555/844379.844681.

Hon, K. K. (2021). Artificial intelligence prediction of air traffic flow rate at the Hong Kong International Airport. *865*, In *IOP Conference Series: Earth and Environmental Science*. IOP Publishing, http://dx.doi.org/10.1088/1755-1315/865/1/012051, 012051.

Hon, K. K., Chan, P. W., Chim, K. C., De Visscher, I., Thobois, L., Rooseleer, F., & Troiville, A. (2022). Wake vortex measurements at the Hong Kong International Airport. In *AIAA scitech 2022 forum* (p. 2011). http://dx.doi.org/10.2514/6.2022-2011.

Hong, S., & Lee, K. (2015). Trajectory prediction for vectored area navigation arrivals. *Journal of Aerospace Information Systems*, *12*(7), 490–502. http://dx.doi.org/10.2514/1.i010245.

Jie, Y., Hui, C., Xingyu, L., & Xuhui, W. (2019). Research on estimated time of arrival prediction based upon ADS-B and spatiotemporal analysis. In *2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCASIT)* (pp. 630–634). IEEE, http://dx.doi.org/10.1109/iccasit48058.2019.8973117.

Jun, L. Z., Alam, S., Dhief, I., & Schultz, M. (2022). Towards a greener Extended-Arrival Manager in air traffic control: A heuristic approach for dynamic speed control using machine-learned delay prediction model. *Journal of Air Transport Management*, *103*, Article 102250. http://dx.doi.org/10.1016/j.jairtraman.2022.102250.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, https://dl.acm.org/doi/10.5555/3294996.3295074.

Kern, C. S., de Medeiros, I. P., & Yoneyama, T. (2015). Data-driven aircraft estimated time of arrival prediction. In *2015 annual IEEE Systems Conference (SysCon) Proceedings* (pp. 727–733). IEEE, http://dx.doi.org/10.1109/syscon.2015.7116837.

Liu, Y., Ng, K. K., Chu, N., Hon, K. K., & Zhang, X. (2023). Spatiotemporal image-based flight trajectory clustering model with deep convolutional autoencoder network. *Journal of Aerospace Information Systems*, 1–13. http://dx.doi.org/10.2514/1.i011194.

Lui, G. N., Hon, K. K., & Liem, R. P. (2022). Weather impact quantification on airport arrival on-time performance through a Bayesian statistics modeling approach. *Transportation Research Part C (Emerging Technologies)*, [ISSN: 0968-090X] *143*, Article 103811. http://dx.doi.org/10.1016/j.trc.2022.103811.

Lui, G. N., Klein, T., & Liem, R. P. (2020). Data-driven approach for aircraft arrival flow investigation at Terminal Maneuvering Area. In *AIAA Aviation Forum* (p. 2869).

Lui, G. N., Liem, R. P., & Hon, K. K. (2020). Towards understanding the impact of convective weather on aircraft arrival traffic at the Hong Kong International Airport. *569*, In *IOP Conference Series: Earth and Environmental Science*. IOP Publishing, Article 012067. http://dx.doi.org/10.2514/6.2020-2869.

Ma, Y., Du, W., Chen, J., Zhang, Y., Lv, Y., & Cao, X. (2022). A spatiotemporal neural network model for estimated-time-of-arrival prediction of flights in a terminal maneuvering area. *IEEE Intelligent Transportation Systems Magazine*, *15*(1), 285–299. http://dx.doi.org/10.1109/mits.2021.3132766.

Mueller, E., & Chatterji, G. (2002). Analysis of aircraft arrival and departure delay characteristics. In *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum* (p. 5866). http://dx.doi.org/10.2514/6.2002-5866.

Murça, M. C. R., Hansman, R. J., Li, L., & Ren, P. (2018). Flight trajectory data analytics for characterization of air traffic flows: A comparative analysis of terminal area operations between New York, Hong Kong and Sao Paulo. *Transportation Research Part C (Emerging Technologies)*, *97*, 324–347. http://dx.doi.org/10.1016/j.trc.2018.10.021.

Nedell, W., Erzberger, H., & Neuman, F. (1990). The traffic management advisor. In *1990 American Control Conference* (pp. 514–520). IEEE, http://dx.doi.org/10.23919/acc.1990.4790788.

Ng, K., Lee, C., Chan, F. T., & Qin, Y. (2017). Robust aircraft sequencing and scheduling problem with arrival/departure delay using the min-max regret approach. *Transportation Research Part E: Logistics and Transportation Review*, *106*, 115–136. http://dx.doi.org/10.1016/j.tre.2017.08.006.

Nguyen, C. H., & Liem, R. P. (2025). Multi-aircraft attention-based model for perceptive arrival transit time prediction. *Advanced Engineering Informatics*, *64*, Article 103067. http://dx.doi.org/10.1016/j.aei.2024.103067.

Rebollo, J. J., & Balakrishnan, H. (2014). Characterization and prediction of air traffic delays. *Transportation Research Part C (Emerging Technologies)*, *44*, 231–241. http://dx.doi.org/10.1016/j.trc.2014.04.007.

Schultz, M., Lorenz, S., Schmitz, R., & Delgado, L. (2018). Weather impact on airport performance. *Aerospace*, *5*(4), 109. http://dx.doi.org/10.3390/aerospace5040109.

Schultz, M., Reitmann, S., & Alam, S. (2021). Predictive classification and understanding of weather impact on airport performance through machine learning. *Transportation Research Part C (Emerging Technologies)*, *131*, Article 103119. http://dx.doi.org/10.1016/j.trc.2021.103119.

Sternberg, A., Carvalho, D., Murta, L., Soares, J., & Ogasawara, E. (2016). An analysis of Brazilian flight delays based on frequent patterns. *Transportation Research Part E: Logistics and Transportation Review*, *95*, 282–298. http://dx.doi.org/10.1016/j.tre.2016.09.013.

Wang, Y., Li, M. Z., Gopalakrishnan, K., & Liu, T. (2022). Timescales of delay propagation in airport networks. *Transportation Research Part E: Logistics and Transportation Review*, *161*, Article 102687. http://dx.doi.org/10.1016/j.tre.2022.102687.

Wang, Z., Liang, M., & Delahaye, D. (2018). A hybrid machine learning model for short-term estimated time of arrival prediction in terminal manoeuvring area. *Transportation Research Part C (Emerging Technologies)*, *95*, 280–294. http://dx.doi.org/10.1016/j.trc.2018.07.019.

Wang, Z., Liang, M., & Delahaye, D. (2020). Automated data-driven prediction on aircraft Estimated Time of Arrival. *Journal of Air Transport Management*, *88*, Article 101840. http://dx.doi.org/10.1016/j.jairtraman.2020.101840.

Wang, L., Mao, J., Li, L., Li, X., & Tu, Y. (2023). Prediction of estimated time of arrival for multi-airport systems via "Bubble" mechanism. *Transportation Research Part C (Emerging Technologies)*, *149*, Article 104065. http://dx.doi.org/10.2139/ssrn.4090449.

Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. (2019). Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, *125*, 203–221. http://dx.doi.org/10.1016/j.tre.2019.03.013.

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

Zhu, X., & Li, L. (2021). Flight time prediction for fuel loading decisions with a deep learning approach. *Transportation Research Part C (Emerging Technologies)*, *128*, Article 103179. http://dx.doi.org/10.1016/j.trc.2021.103179.