Weather impact quantification on airport arrival on-time performance through a Bayesian statistics modeling approach

Go Nam Lui*[1], Kai Kwong Hon†[2], and Rhea P. Liem‡[3]

[1,3]The Hong Kong University of Science and Technology
[2]Hong Kong Observatory, Hong Kong SAR

Abstract Compared with departures, predicting the weather impact on arrival delays is more challenging because of possible non-linear, cascading effects, and higher uncertainty. Existing weather impact studies are location-dependent and often neglect the impacts of dangerous phenomena. We propose a data-driven model for severe weather impact quantification on airport arrival on-time performance based on the Bayesian approach to address these issues. Our model considers the impact of the dangerous phenomenon by evaluating the mean shift and is flexible enough to be applied to different airports. Using two years' worth of data (2017-2018) from the Hong Kong International Airport, we studied over 55,000 local meteorological reports and analyzed over 430,000 arrival flights. Across all three key performance metrics considered, a non-linear relationship with the weather score, akin to a phase transition, could be observed. This framework allows a comparison between the sensitivity of each airport's arrival performance metric towards severe weather. Delay rate is the most sensitive metric, while cancellation rate is the least. For the impacts of dangerous phenomena, cumulonimbus has the most significant impact on the delay rate. Shower rainfall/cumulonimbus has a similar and vital impact on the mean arrival delay per hour. Because of its potential applications in different airports, this framework can provide a deeper insight into weather impact on air traffic networks.

1 Introduction

Air transportation has been a key driver of worldwide commercial connections and economic development over recent decades. From 1980 to 2019, the number of carried passengers via air transportation increased from 0.6 billion to 4.2 billion in an exponential trend. With this rapid growth, concerns are growing over air traffic delays. According to the American Bureau of Transportation Statistics (2021), between 2012 to 2019, the proportion of delayed arrivals in the U.S. increased from 15.06% to 19.79%, while the proportion of delayed departures also grew from 14.67% to 19.21%. Recently, the aviation industry has been severely battered by the COVID-19 pandemic.

*Ph.D. Candidate, Department of Mechanical and Aerospace Engineering.

†Acting Senior Scientific Officer, Forecast Operation, Forecasting and Warning Services Branch.

‡Assistant Professor, Department of Mechanical and Aerospace Engineering.

However, the industry has also demonstrated its resilience in the past by bouncing back and continuing its overall upward trend after incidents such as the terrorist attacks in 2001, the war in Afghanistan that broke out in 2001, and the outbreak of severe acute respiratory syndrome (SARS) in 2003 (Mason, 2005). Recent numbers seem to agree with this propensity. In December 2021, while the revenue passenger kilometers (RPK) was still down by 45.1% compared with the same month in 2019, it was a notable improvement from March 2021, where the RPK was 74.7% lower than that of 2019 (ICAO, 2022). Therefore, despite the current downturn, addressing air traffic delays is still important and relevant in the long term, since air transportation is expected to rebound and continue its upward trend within the next few years.

This research is focused on studying the impacts of weather conditions on airport arrival on-time performance. Between 2009 and 2019, 57.83% of the American National Aviation System (NAS) delays were caused by adverse weather (Bureau of Transportation Statistics, 2021). In particular, we focus on aircraft arrival in the terminal control area, also called the terminal maneuvering area (TMA), which is regarded as the bottleneck of air traffic management (ATM) because of its high traffic density and complexity (Kistan et al., 2017; Spinardi, 2015; Erzberger and Lee, 1972). Aircraft arrival and departure delays inside the TMA attributable to convective weather conditions can cause economic losses owing to reduced performances. During arrivals, for instance, the extra maneuvering required due to weather conditions burns more fuel, which in turn increases the operating costs for airlines. Borsky and Unterberger (2019) investigated the weather shock impact on departure delays in the U.S., and the results quantified the general impact of weather conditions on aircraft delays and social costs. Rain, ice, snow, and hail increase the amount of time that aircraft spend on runways. Thunderstorms will restrict airspace capacity and produce congestion, while clouds may obstruct pilot's visibility. Hence, local weather impact quantification on airport/terminal area performance becomes increasingly important for future ATM systems. In this research, we will focus our attention on the arrival delays because of its higher level of uncertainty compared with departure delays.

Weather-impact study is location-dependent, such that the significant weather influence factors differ for different climate conditions in different geographical locations. Early researches on local aviation weather impact extracted weather features affecting a particular airport (e.g., snowstorms, thunderstorms, and surface winds) which were then correlated with flight delays (McCarthy et al., 1982; Robinson, 1989; Allan et al., 2001). Recently, the investigation of severe local weather impact on ATM has become more diverse. Some studies focused on the flight re-routing problem under specific weather conditions (Krozel et al., 2007; McCrea et al., 2008; Pfeil and Balakrishnan, 2012), which employed different weather data sources and different methodologies to re-arrange the default flight route. Advanced data-driven models have been applied to study weather impacts on flight trajectory prediction for specific origin–destination pairs (Pang et al., 2021, 2019; Pang and Liu, 2020; Zhao et al., 2019). Some other studies focused on quantifying the impact of hazardous weather on airspace capacity (Song et al., 2009; Buxi and Hansen, 2011), by investigating the impact of individual weather features on the airspace arrival/departure rate. The combined impact of the variety of weather features is complex, which motivated studies to classify the type of weather based on their impact on air traffic (Grabbe et al., 2014). Later on, another study related to weather impact quantification based on machine learning methodology was also constructed (Schultz et al., 2021). Using a combination of recurrent and convolutional neural networks, their model could perform predictive weather-dependent airport performance classification across six hours, focusing on London Gatwick Airport (LGW). de Oliveira et al. (2021) analyzed the weather impact on the delayed arrival occurrence for Brazilian domestic air traffic system using logistic regression. Rodriguez-Sanz et al. (2021) studied the local weather impact on airport arrival performance at Adolfo Suárez Madrid-Barajas airport (MAD) by applying the Bayesian network model. The recent research trend of weather impact quantification provides a growing potential for further follow-up research.

Such a local study, however, has not been performed for the Hong Kong International Airport (HKIA), which is one of the world's major transportation hubs, both in passenger numbers and cargo volumes. Besides its geographical location, its coastal and subtropical climate makes local aviation weather in Hong Kong unique. As an example, low-level wind shear and turbulence are known hazards at HKIA (Shun and Chan, 2008; Hon and Chan, 2022), which in severe instances can cause considerable traffic disruption (Chan and Hon, 2016). To the best of our knowledge, there have not been any studies that systematically investigate and quantify the weather impact on terminal area traffic pertaining to HKIA. Furthermore, existing studies do not sufficiently analyze and interpret the impacts of dangerous weather phenomena. The official weather impact research document of EUROCONTROL even stated that the real impacts of dangerous phenomena on airport operations are nearly impossible to predict (EUROCONTROL, 2011). Some works only covered a narrower range of weather scores (Schultz et al., 2018), or barely described the influence of dangerous phenomena (Reitmann et al., 2019; Schultz et al., 2021; Rodriguez-Sanz et al., 2021). A linear model has previously been used to describe weather impact on airport traffic performance (Schultz et al., 2018), which might not be sufficient to describe the non-linear relationship as revealed by data. Lastly, weather phenomena are dynamic in nature owing to several factors, such as the constant changes of global climate (Easterling et al., 2000), which will have an impact on aviation activities (Ryley et al., 2020). Therefore, it is imperative to derive a weather impact model that is adaptable to future changes.

In this paper, we aim to establish a quantitative relationship between weather conditions, including dangerous weather phenomena, and airport arrival on-time performance at HKIA. Instead of relying on black-box models, which lack interpretability (Rudin, 2019), we derive the explanatory model based on a growth function that can emulate the trend of weather impacts more realistically. The model parameters are derived based on the Bayesian approach, to allow for model updating when newly-observed data are available. As such, we can ensure that the model remains relevant albeit some weather changes due to, say, climate change, by updating the parameters as needed.

This paper commences with a description of the proposed approach, which includes airport arrival on-time performance metrics, weather data, and preliminary data analysis in Section 2. The key components for our methodologies are established in Section 3. Section 4 presents the detailed results and Section 5 briefly summarizes the work of this paper.

2 Proposed approach

The proposed framework is illustrated in Fig. 1. The pink-colored blocks refer to data processing and the red-framed blocks contain the analysis results, including (1) explanatory models describing the relationship between weather conditions and airport arrival on-time performance (with uncertainty representation), (2) traffic metric sensitivity, and (3) dangerous phenomenon impact quantification.

This analysis requires both arrival flight information data and weather data pertaining to HKIA, where we use data from 2017 to 2018. For arrival flight information, there are 214,952 records from 2017 and 218,728 records from 2018. From the raw air transportation data, we derive the corresponding airport arrival on-time performance metrics, as described in Section 2.1. For weather data, we use the weather scores derived based on Meteorological Aerodrome Reports (METAR) data, which is one of the most commonly used data sources for weather impact studies (see, for instance, Rodriguez-Sanz et al., 2021; Murça, 2021; Lemetti et al., 2020; Schultz et al., 2018, 2021). The total number of raw METAR code data used in this study is 55,313. We use the Air Traffic Management Airport Performance (ATMAP) weather algorithm, developed by EUROCONTROL

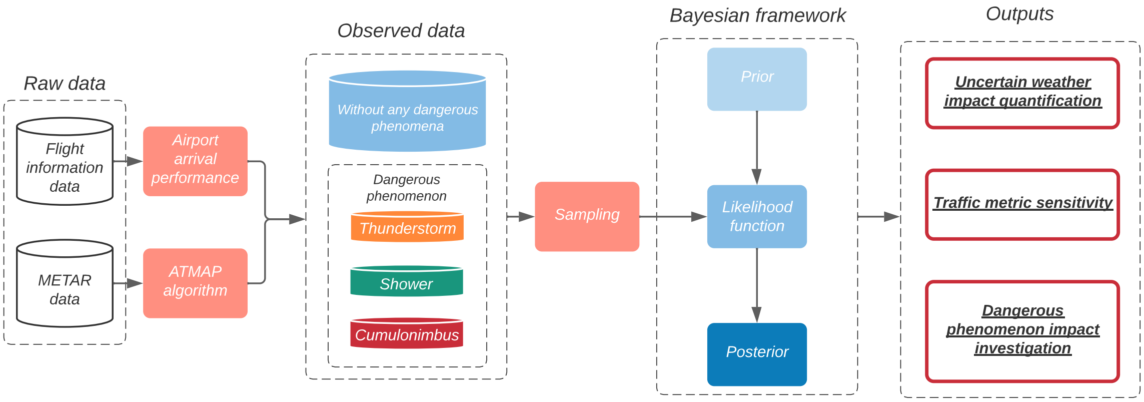


Figure 1: Schematic flowchart for Bayesian-based weather impact quantification.

(2011), to assign quantitative weather scores for different weather conditions. This algorithm has been used to assess the weather impact on aircraft performance in several works. Schultz et al. (2018) investigated the weather impact on European flights by analyzing 20.5 million flights in 2013. Based on this work, Reitmann et al. (2019) developed an ATMAP 2.0 algorithm for weather impact quantification, using unsupervised learning for clustering and supervised learning for classification. The METAR data and ATMAP algorithm will be further explained in Section 2.2, followed by a preliminary data analysis in Section 2.3.

As shown in Fig. 1, we implement the Bayesian approach to infer the parameters for the derived explanatory models based on data. Not only can the Bayesian approach quantify and characterize the output uncertainty, it can also make the models adaptable to future changes, by deriving new posteriors in the presence of newly-observed data. This will be further explained in Section 3.

Existing weather score calculation methods are insufficient to explain the impacts of dangerous phenomena on airport arrival on-time performance, as will be further elaborated in Section 2.3. As such, we propose to separate the analysis with and without dangerous phenomena. Once the respective models are derived, we use mean shift assessment to investigate the impact of dangerous phenomena. By applying this approach, we can address the limitations of some existing studies, which mostly excluded the influence of dangerous phenomena (Schultz et al., 2018; Reitmann et al., 2019; Schultz et al., 2021; Rodriguez-Sanz et al., 2021).

This quantification is aligned with ongoing efforts to integrate weather information into the new generation of ATM systems. The International Civil Aviation Organization (ICAO) has promoted closer integration of meteorological and air traffic information ("MET-ATM integration") in the Aviation System Block Upgrade (ASBU), which sets stage goals for every five years (ICAO, 2019). In Europe, the EUROCONTROL pursues a similar initiative under The Single European Sky ATM Research (SESAR) programme. Furthermore, quantifying the weather impact will help air traffic controllers (ATC) at the strategic (as early as a year before) and pre-tactical (as early as a week before) stages of flight planning and provide more accurate weather constraints for scheduling aircraft arrivals.

2.1 Airport arrival on-time performance metrics

The Federal Aviation Administration (FAA) categorizes flight delays into arrival/departure delay, taxi-in/out delay, and en-route delay. Arrival delay, which is the focus of this study, is typically hard to predict (Schultz et al., 2021). While there is already a large body of literature on aircraft delays, gaps still remain in the terminal arrival delay studies. Despite optimized schedules, flights might experience prolonged delays in the terminal area as they approach an airport to land, which is often caused by weather conditions. At this point, departure scheduling and flight rerouting are no longer possible. The vectoring and holding maneuvers that the flights need to go through under this situation may incur additional fuel burn and noise impacts, which are undesirable. Furthermore, adverse weather impacts on arrival delays are found to have snowball effects on subsequent flights, even after the weather has returned to normal (Lui et al., 2020a). Therefore, quantifying weather impact on arrival on-time performance contributes to arrival delay investigation and further enables deriving appropriate mitigation strategies.

Airport arrival *on-time performance* is assessed based on recorded flight schedule data. Such data have been frequently used in air traffic research, including operation analysis benchmarking (Gopalakrishnan et al., 2021), delay-pattern analysis (Sternberg et al., 2016), and traffic delay prediction (Rebollo and Balakrishnan, 2014). Recorded flight data reveal temporal patterns due to night flying regulations (Lui et al., 2020a,b). Fig. 2 illustrates the hourly variations of scheduled arrivals at HKIA representing air traffic activities, based on one-month's data in May 2019 (Lui et al., 2020a). As shown in this figure, the peak hourly arrival rates occur with 30 to 36 flights. This is consistent with the stated official maximum hourly movement capacity at HKIA of 68 flights under the original two-runway system with segregated mode of operations (Lo, 2015). Fig. 2 also reveals a significant reduction in air traffic between 10 pm and 9 am owing to night-time restrictions, which include any form of regulatory measures to limit aircraft noise emission exposure to residents during night time, as well as runway closure programs at HKIA. As such, we only include arrival time periods between 9 am and 10 pm (when air traffic activities are more significant) in the current study. This time period is indicated by the shaded blue region in Fig. 2. In this study, flights arriving earlier than the scheduled time are considered *on-time*, i.e., the actual and scheduled arrival times are assumed to be the same. Arıkan et al. (2013) also applied the same assumption in their air-travel infrastructure study.

For evaluation purposes, we define three airport arrival on-time performance metrics, which are briefly described below.

Mean arrival delay per hour (μ_{AD}) This non-negative integer metric measures the average of arrival delays of N flights within an hour, where *aircraft arrival delay* is defined as the discrepancy between *scheduled arrival time (SAT)* and *actual arrival time (AAT)* for each flight. This metric can be expressed as

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

where f denotes the flight index and the superscript $^+$ indicates the non-negativity of the metric. When a flight arrives ahead of the scheduled time, the negative $AAT_f - SAT_f$ value is converted to zero.

Cancellation rate per hour (RT_c) This metric is defined as the ratio of the number of cancellations per hour with respect to the corresponding total number of flights, which can be expressed as

$$RT_c = \frac{1}{N}\sum_{f=1}^{N} C_f, \text{ where } C_f = \begin{cases} 1 & \text{when flight } f \text{ is cancelled} \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

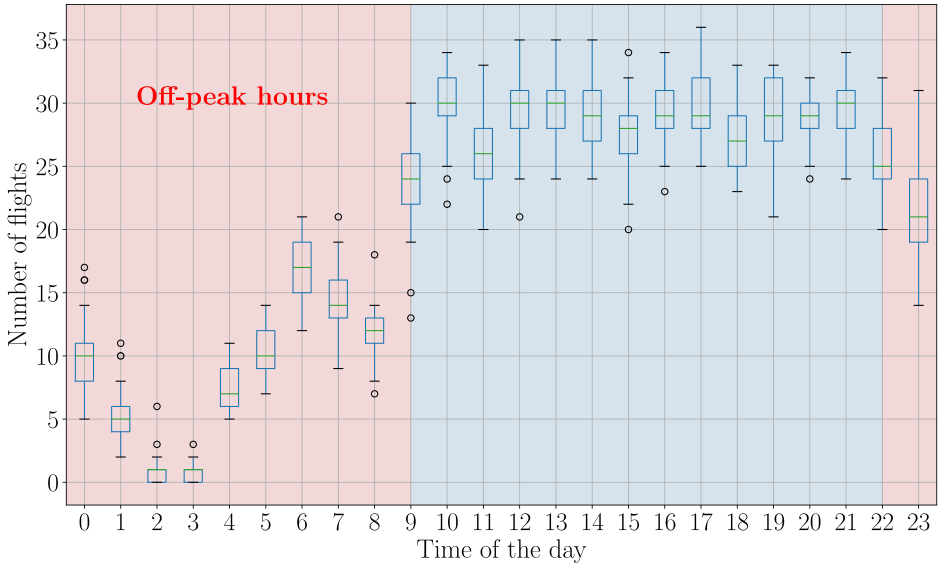


Figure 2: Temporal patterns for arrival flights at HKIA

where C_f refers to the cancellation indicator for flight f. The *flight cancellation* is identified when the actual arrival time information is missing in the flight information data set. An empty timestamp string is returned when there is no recorded arrival for a particular scheduled flight. Thus, when the actual arrival time returns an empty timestamp string, we assume that the flight is cancelled.

Delay rate per hour (RT_d) A flight is considered delayed when it fails to arrive within 15 minutes of the scheduled time (Mueller and Chatterji, 2002), following the definition from FAA and the Bureau of Transportation Statistics (BTS). To define the metric, we use a delay indicator D_f, which is a Boolean variable that depends on the state of delay. The formulation is similar to that of RT_c, which is shown below

$$RT_d = \frac{1}{N}\sum_{f=1}^{N} D_f, \text{ where } D_f = \begin{cases} 1 & \text{when flight } f \text{ is delayed} \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

The values of RT_c and RT_d range between zero and one, since they are defined as proportions. In this study, the value of μ_{AD} is normalized to be within the same range, for consistency in the computation. The normalized mean arrival delay per hour is expressed below:

$$\tilde{\mu} = \frac{\mu_{AD} - \min(\boldsymbol{\mu_{AD}})}{\max(\boldsymbol{\mu_{AD}})}, \tag{4}$$

where $\boldsymbol{\mu_{AD}}$ is the vector for μ_{AD}. This implementation allows the priors to be exchangeable among three airport arrival traffic metrics, which will be elaborated in Section 3.1.

2.2 Weather information

In this section, we first describe the weather data (in METAR format) and the algorithm to quantify the weather score. We then demonstrate the scoring algorithm with actual data and discuss the current limitations.

2.2.1 METAR data

METAR reports hourly weather conditions within a 16 km radius of an airport, though some airports might provide data every half-hour, e.g., Frankfurt International Airport. An example of raw METAR data, obtained from `Navlost.eu`, pertaining to HKIA at one timestamp is given in Table 1.

Table 1: Example features from METAR data (15:30, Mar 30th, 2021, HKG).

Raw codes	Features	Values
VHHH	ICAO airport identifier	Hong Kong International Airport
300730Z	Date Time	30th (March) 07:30 UTC
21013KT 170V230	Wind speed Wind direction	13 knots 210°, variable between 170° and 230°
9999	Visibility	9,999 meters
FEW012 SCT030	Cloud type Cloud location	Few clouds; scattered cloud FEW: 1,200 ft AGL; SCT: 3,000 ft AGL
28/22	Temperature Dew point	28°C 22°C
Q1005	Air pressure	1,005 hpa
NOSIG	Remarks	No significant change expected in the next 2 hours

Raw METAR data provide a long string that contains diverse weather information for ATM purposes including wind, moisture, visibility, etc. Airports, following the definition by the World Meteorological Organization (WMO), provide weather information in this particular format, which is considered interchangeable between airports. However, METAR contains some qualitative information that cannot be directly used in computational modeling and simulation. The ATMAP weather algorithm is typically used to convert METAR data (both qualitative and quantitative) into quantitative weather scores, which is described next.

2.2.2 ATMAP weather algorithm

The ATMAP weather algorithm can parse METAR weather information into a quantitative index, which is referred to as the *weather score* (EUROCONTROL, 2011). The algorithm has been commonly used in past aviation weather researches (Schultz et al., 2018; Reitmann et al., 2019; Lemetti et al., 2020; Schultz et al., 2021; Rodriguez-Sanz et al., 2021). The algorithm categorizes METAR weather information into five weather classes, namely visibility, dangerous phenomenon, freeze condition, precipitation, and wind condition. Within each weather class, the severity is indicated by the assigned weather score. The input features and corresponding weather score range for each weather class are tabulated in Table 2. For further details of the score calculation, readers are referred to the explanation provided in the relevant official document (EUROCONTROL, 2011). The hourly total weather score at a specific airport is then obtained by summing up the scores of these five classes.

As shown in Table 2, the scores assigned to dangerous phenomena are notably higher than those of others (Reitmann et al., 2019). Dangerous phenomena are often excluded in existing studies, or included with limited usage owing to this reason (Schultz et al., 2018; Reitmann et al., 2019; Schultz

Table 2: Weather classes and their presentations

Weather class	Input features	Score range
Visibility and ceiling	Visibility $\leq 1,500$ m, cloud type and cover	[0,5]
Wind	Wind speed $>$ 15 knots, gust	[0,5]
Precipitation	Rain, (+/-) snow, etc.	[0,3]
Freeze condition	T $\leq$ 3°C, dew point, precipitation	[0,4]
Dangerous phenomenon	TCU/CB, cloud cover, (+/-) phenomenon	[0,30]

et al., 2021). In this research, we introduce a new hierarchical approach to enable quantifying the impacts of dangerous phenomena on airport arrival on-time performance.

2.3 Preliminary data analysis

Before developing the solution methods, the arrival on-time performance and weather data are analyzed and characterized, as shown below.

2.3.1 Location-dependent weather variation

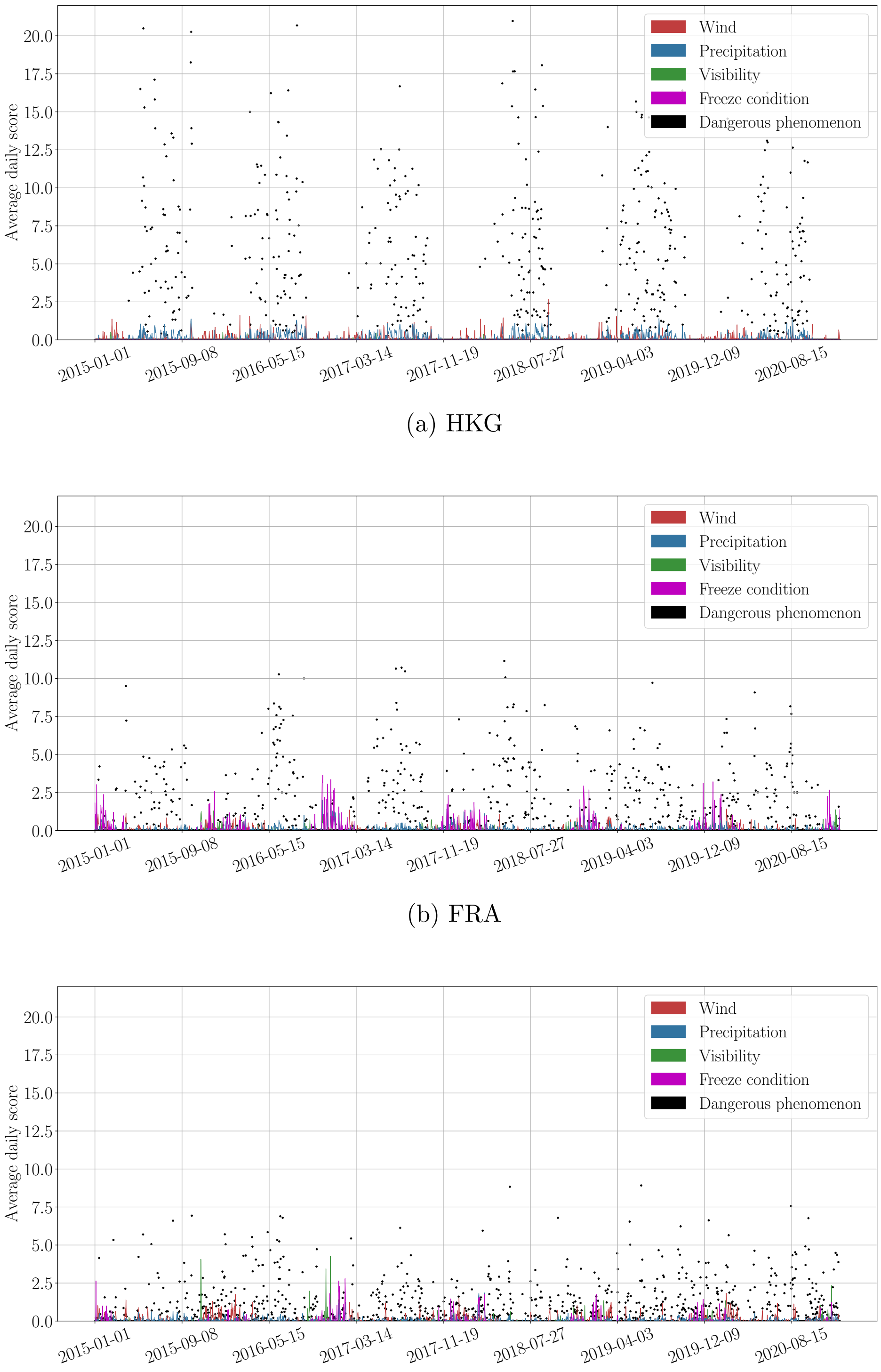
The ATMAP weather algorithm is demonstrated with the daily weather scores pertaining to three airports: Hong Kong (HKG), Frankfurt (FRA), and London (LHR), which are shown in Fig. 3. Data from 2015 to 2020 were used to generate these figures, and each weather class (refer to Table 2) is identified by a different color. The data from the three airports show some periodicity, each with a different distributional pattern.

Among the three airports, HKG notably shows more divergent weather conditions, with a larger distribution of ATMAP weather scores. This wider distribution indicates a larger contribution from dangerous phenomenon, which is identified as the most dominant weather phenomenon in HKG, followed by precipitation. At FRA and LHR, on the other hand, the freeze condition has the most frequent occurrence among all weather classes. The visibility condition also appears to be dominant at some instances at LHR. These observations highlight the uniqueness of weather conditions at different airports, which calls for more location-specific studies to investigate weather impact on airport arrival on-time performance. For HKG, in particular, the impacts of dangerous phenomena cannot be neglected for a comprehensive weather impact analysis. Note that the data pertaining to FRA and LHR are only used for weather score demonstration and discussion purposes. The aviation weather impact study is focused on HKG, using local weather and arrival flight information data.

2.3.2 Weather impact characteristics for Hong Kong

In this section, we use boxplot to visualize and observe the correlation between weather score and airport arrival on-time performance. Since the three airport arrival performance metrics exhibit a similar pattern towards weather score, we only show the results for μ_{AD} for conciseness. Fig. 4 demonstrates the trends of μ_{AD} along with the weather score for two years (2017– 2018).

The distribution of data suggests that there are four main clusters of weather-μ_{AD} trend, which are indicated by different hues. The contributions from different weather classes on weather score calculation are shown as the single-stacked bar charts above each cluster. The percentage of data that fall within each cluster is also indicated on the plot. Several scores (i.e., $8-10$ and 23) do not have any data points associated with them, as indicated by the shaded gray area. The majority of data (i.e., 92.3%) fall within the $[0,7]$ cluster, with around 80% of data having their weather scores equal to zero. Outliers are observed especially in lower weather scores. These outliers correspond to delays caused by factors other than weather, such as airspace restrictions, operation turnover,



(a) HKG

(b) FRA

(c) LHR

Figure 3: Daily average weather score for three hub airports from 2015 to 2020.

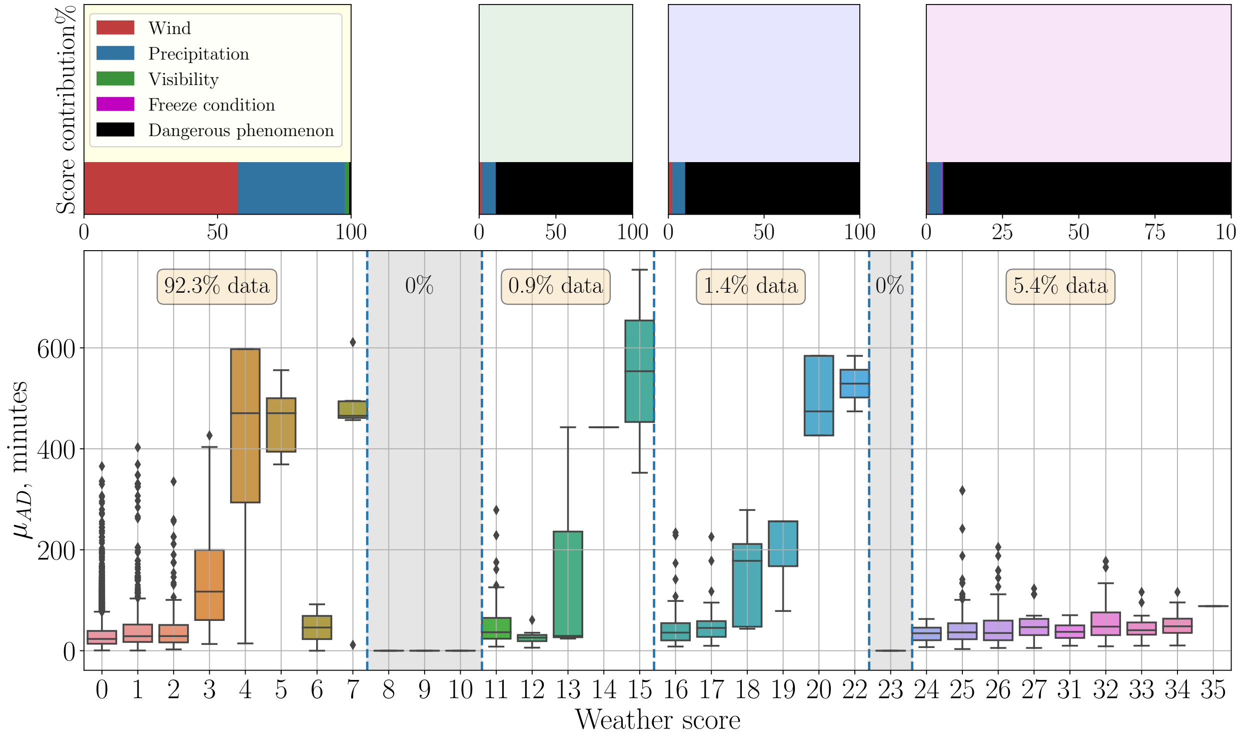


Figure 4: The trend of μ_{AD} with respect to weather score and the corresponding score contributions from five ATMAP weather classes.

runway occupation, etc.

Fig. 4 establishes two important characteristics of weather score trend for Hong Kong. First, the μ_{AD} trend does not show any linear or monotonic relationship with weather score. The non-monotonic trend can hardly reveal the real relationship between weather and airport arrival traffic performance in HKG. This observation shows that the current ATMAP weather scoring algorithm is not suitable for deriving the correlations between weather scores and airport arrival on-time performance in Hong Kong, especially for weather scores higher than seven. Second, the single-stacked bar charts show that dangerous phenomenon dominates the second, third, and fourth clusters, whereas the wind condition and precipitation are more prevalent in the first cluster. Referring to Table 2, the value range for “dangerous phenomenon” weather class is significantly wider than those of other classes, with a maximum of 30 while all others are below five. Further investigation reveals that 43.7% of the dangerous phenomenon occurrence in the fourth cluster is due to thunderstorm conditions. While thunderstorm is not an ideal condition for taking off and landing, it does not always cause severe disturbance in arrival air traffic in HKG, which explains the low μ_{AD} values in this cluster.

Based on this observation, we adopt a hierarchical approach to interpret the overall weather impacts on airport arrival on-time performance by separating data with and without dangerous phenomenon. First, we quantify weather impacts considering all weather classes except for dangerous phenomenon, which is shown in Fig. 5. A single-stacked bar chart on the right hand side, similar to those in Fig. 4, displays the contribution from each weather class. As shown in Fig. 5, excluding dangerous phenomenon weather class in weather score calculation limits the score range to $[0, 7]$. Furthermore, μ_{AD} exhibits a monotonically increasing trend with respect to weather score, thereby enabling the derivation of the correlation. Schultz et al. (2018) only considered this particular range of weather score, i.e., by excluding dangerous phenomenon, in their weather impact study on European airports in 2018. Next, we assess the shift in airport arrival on-time performance metrics caused by the presence of dangerous phenomena to quantify the overall weather impacts on arrival performance. This approach is further discussed in Section 3.

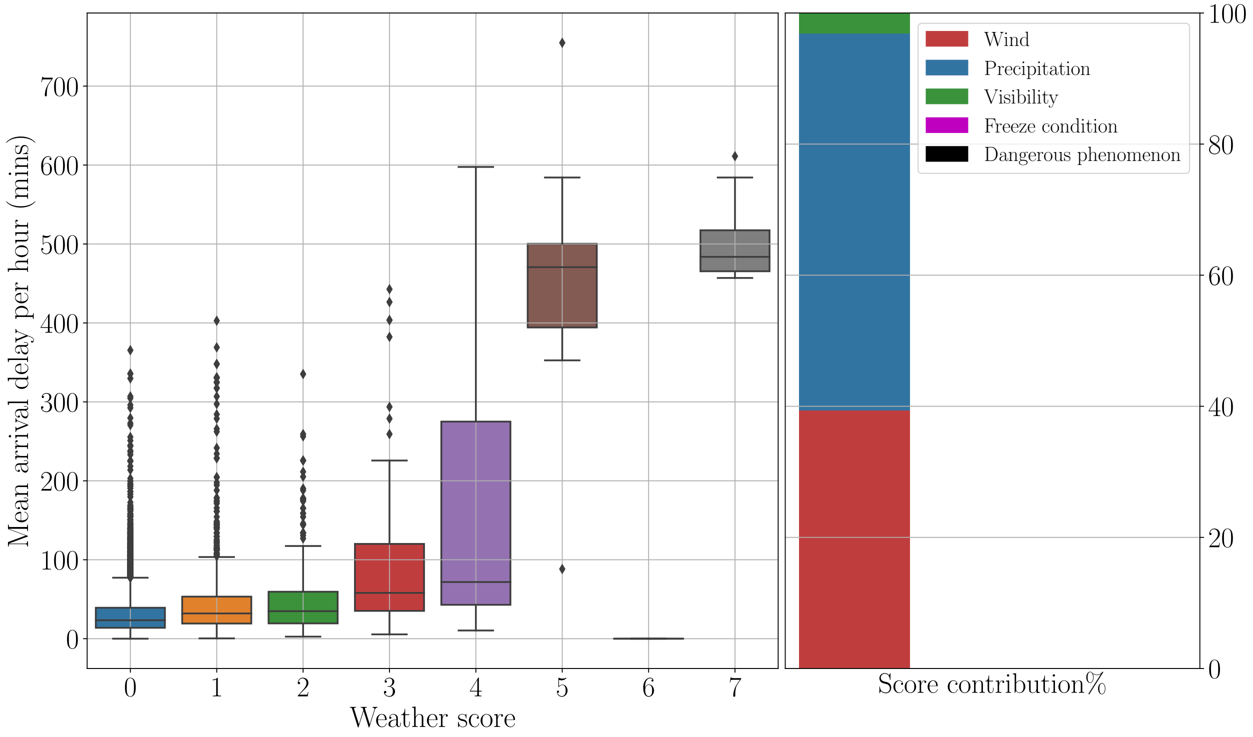


Figure 5: Weather score versus μ_{AD}, without dangerous phenomenon

2.3.3 Balanced sampling for data with long-tail features

Fig. 6 shows a histogram of weather scores (in logarithmic scale). A long-tail effect is noticeable here, with the majority of observed data having low weather scores. This imbalanced dataset reflects, to a certain extent, the local climatology and the relative occurrence frequency of those weather types classified under the ATMAP score approach at HKIA. To alleviate bias, we perform undersampling and oversampling on the imbalanced data set, which are commonly performed on data sets with this characteristic (Krawczyk, 2016). In particular, we apply the adaptive synthetic (ADASYN) method for oversampling and the cluster centroids algorithm for undersampling. ADASYN is based on the K-nearest neighbors algorithm and has the advantage of not copying the minority of data (He et al., 2008). The cluster centroids algorithm samples data by generating centroids based on K-means (Likas et al., 2003). Specifically to the problem at hand, we undersample weather scores that are less than three and oversample those above three to achieve a more balanced sample distribution, as shown in Fig. 6. The green bars show the resulting balanced data sets. The same oversampling and undersampling procedures are also applied to dangerous phenomenon data sets.

3 Methodology

In this section, we describe a new method to quantify the weather impacts on airport arrival on-time performance pertaining to all weather conditions, including dangerous phenomena. As mentioned

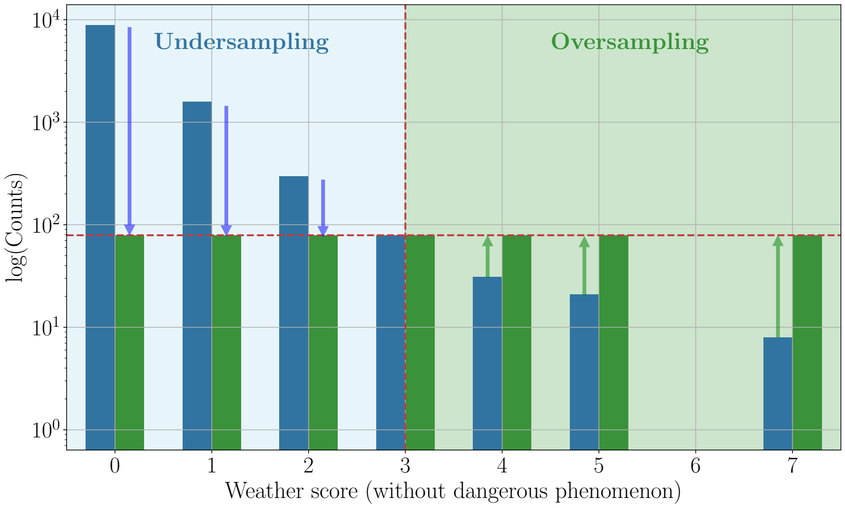


Figure 6: Oversampling and undersampling procedures to achieve a more balanced data set, due to the long-tail features of the weather data.

in Section 2.3.2, we separate the weather impact quantification pertaining to data with and without dangerous phenomena. From past data, only thunderstorm, shower, and cumulonimbus (among other dangerous phenomenon types) are commonly observed in Hong Kong. Shower refers to intense precipitation and cumulonimbus refers to dense, towering vertical cloud (EUROCONTROL, 2011). Although interactions might occur among dangerous phenomena, e.g., thunderstorms and cumulonimbus typically appear together, we assume that dangerous phenomenon indicators are mutually independent in this study. This assumption allows us to simplify the problem at the model development stage. As such, we have four data sets in our study, one for data without any dangerous phenomena, and one for each of the three dangerous phenomena considered.

An explanatory model is then derived for each combination of data set (indexed with i) and on-time performance metric (indexed with m). This model can generally be expressed as

$$y_{i,m} = \mathcal{M}\left(x\,|\boldsymbol{\theta}_{i,m}\right) + \mathcal{P}_{i,m}, \tag{5}$$

where x and y denote the weather score and airport arrival on-time performance (i.e., $\tilde{\mu}$, RT_c, or RT_d), respectively, and $\mathcal{P}$ represents the stochastic component, representing uncertainty. $\mathcal{M}$ is the deterministic mean trend of airport arrival performance metrics with respect to weather score, and the vector $\boldsymbol{\theta}$ contains the parameters of the deterministic model, which are also referred to as the *local parameters*. With four data sets and three on-time performance metrics, we have 12 models in total. The data distribution around the mean value at each weather score is represented by a probability distribution function (PDF). The model formulation shown in Eq. (5) requires assuming the same distribution type for each weather score. Hence, for generality, the Gaussian distribution, which is the most commonly used distribution type in statistics (Murphy, 2007), is selected for model derivation purposes. With this assumption $\mathcal{P}$ can then be expressed as $\mathcal{N}(0, \sigma^2)$, where $\mathcal{N}$ represents a Gaussian distribution. Therefore, we have

$$y_{i,m} \sim \mathcal{N}\left(\mathcal{M}\left(x\,|\boldsymbol{\theta}_{i,m}\right), \sigma^2_{i,m}\right). \tag{6}$$

The derived explanatory models must capture the tendency of the on-time performance to deteriorate with increasing weather score, represent the uncertainty distribution observed within each weather score, and enable quantifying the impacts of dangerous phenomena. These requirements are met by the following three key components in the model derivation:

Deterministic mean trend derivation The deterministic mean trend refers to the $\mathcal{M}$ function in Eq. (5), which quantifies how, on average, each airport arrival on-time performance varies with increasing weather score.

Model parameter derivation via Bayesian approach To ensure that the derived models are realistic and can represent actual situations, the Bayesian approach is employed to infer the appropriate model parameters based on data.

Dangerous phenomenon impact assessment The impact of each dangerous phenomenon is assessed by quantifying the *mean shift* of the derived model corresponding to a particular dangerous phenomenon, as compared to the one without dangerous phenomena. Let us index the data set without dangerous phenomena with $i = 1$, and those containing thunderstorm, shower, and cumulonimbus with $i = 2, 3, 4$, respectively. This mean-shift calculation is illustrated in Fig. 7.

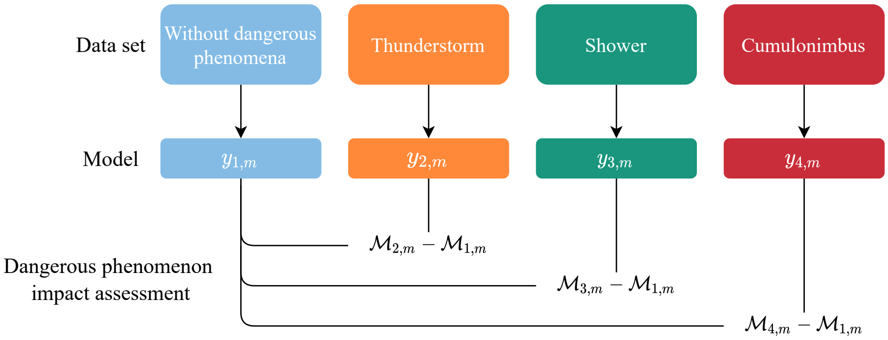

Data set
Without dangerous phenomena
Thunderstorm
Shower
Cumulonimbus
Model
$y_{1,m}$
$y_{2,m}$
$y_{3,m}$
$y_{4,m}$
Dangerous phenomenon impact assessment
$\mathcal{M}_{2,m} - \mathcal{M}_{1,m}$
$\mathcal{M}_{3,m} - \mathcal{M}_{1,m}$
$\mathcal{M}_{4,m} - \mathcal{M}_{1,m}$

Figure 7: Quantifying the impact of each dangerous phenomenon by calculating the mean shift of the derived models.

Each of these components is further discussed in the following subsections. For our research, we construct the model based on the Python numerical programming tool `NumPy` (Harris et al., 2020), and Python probabilistic programming tools `Pymc3` (Salvatier et al., 2016) and `ArviZ` (Kumar et al., 2019).

3.1 Bayesian model structure

The Bayesian model structure to derive the appropriate model parameters is explained in this section. The overall model framework is illustrated with a directed acyclic graph (DAG), as shown in Fig. 8. Each derived model (for a particular combination of i, i.e., the data set based on weather condition, and m, i.e., the on-time metric performance) has its own DAG. All of them, however, share the same structure, as described below. In this DAG, squares and circles denote constants and variables, respectively. Solid arrows indicate stochastic dependency, while dashed arrows signify deterministic dependency.

Both the deterministic and stochastic components of the model, referring to Eq. (6), are shown in this DAG. The inserted rounded rectangle with a blue frame refers to the deterministic trend function, $\mathcal{M}$, in our framework. Each $\mathcal{M}$ function is characterized by a set of *local parameters*, $\boldsymbol{\theta} = \left[\theta^1, \theta^2, \ldots, \theta^n\right]$ (highlighted in blue in Fig. 8). These local parameters and the model variance $(\sigma^2_{i,m})$ are inferred based on data by using the Bayesian approach.

i	Dataset
1	No dangerous phenomena
2	Thunderstorm
3	Shower
4	Cumulonimbus

m	Metric
1	$\tilde{\mu}$
2	RT_c
3	RT_d

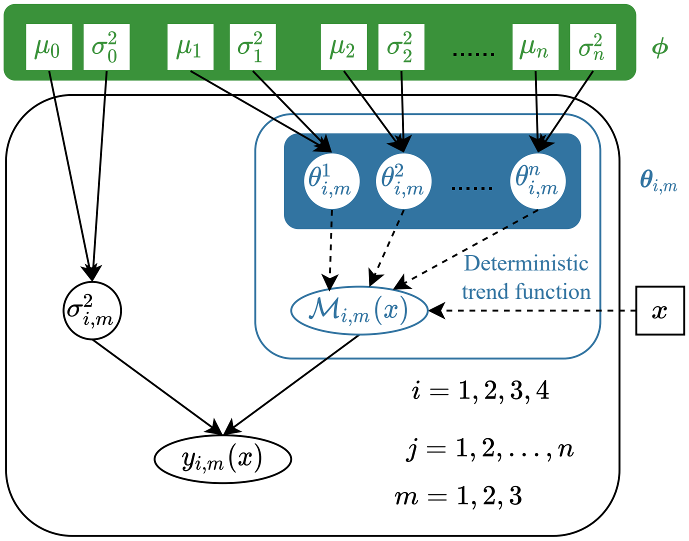


Figure 8: The directed acyclic graph (DAG) for the proposed Bayesian model structure.

We will now describe the Bayesian update procedure, where the posterior distribution of model parameters ($\boldsymbol{\theta}$ and σ^2) for each model (i, m) is expressed as

$$\begin{aligned} P\left(\boldsymbol{\theta}_{i,m}, \sigma_{i,m}^2 \left| \mathcal{D}_{i,m}\right.\right) &\propto P\left(\mathcal{D}_{i,m} \left| \boldsymbol{\theta}_{i,m}, \sigma_{i,m}^2\right.\right) P\left(\boldsymbol{\theta}_{i,m}, \sigma_{i,m}^2\right) \\ &\propto P\left(\mathcal{D}_{i,m} \left| \boldsymbol{\theta}_{i,m}, \sigma_{i,m}^2\right.\right) P\left(\boldsymbol{\theta}_{i,m} | \sigma_{i,m}^2\right) P\left(\sigma_{i,m}^2\right) \\ &\propto P\left(\mathcal{D}_{i,m} \left| \boldsymbol{\theta}_{i,m}, \sigma_{i,m}^2\right.\right) \left(\prod_{j=1}^{n} P\left(\theta_{i,m}^j | \sigma_{i,m}^2\right)\right) P\left(\sigma_{i,m}^2\right). \end{aligned} \tag{7}$$

Since we assume that $\sigma_{i,m}$ and $\theta_{i,m}^j$ are statistically independent, we have $P\left(\theta_{i,m}^j | \sigma_{i,m}^2\right) = P\left(\theta_{i,m}^j\right)$ and therefore,

$$P\left(\boldsymbol{\theta}_{i,m}, \sigma_{i,m}^2 \left| \mathcal{D}_{i,m}\right.\right) \propto P\left(\mathcal{D}_{i,m} \left| \boldsymbol{\theta}_{i,m}, \sigma_{i,m}^2\right.\right) \left(\prod_{j=1}^{n} P\left(\theta_{i,m}^j\right)\right) P\left(\sigma_{i,m}^2\right). \tag{8}$$

In this Bayesian update procedure, $P\left(\mathcal{D}_{i,m} \middle| \boldsymbol{\theta}_{i,m}, \sigma_{i,m}^2\right)$ denotes the likelihood function, whereas $P\left(\theta_{i,m}^j\right)$ and $P\left(\sigma_{i,m}^2\right)$ refer to the prior distributions of the model parameters.

Without any prior knowledge of the local parameters' distributions, the commonly used Gaussian distribution is selected due to its generality (Bishop, 2006; van de Schoot et al., 2021). As such, the prior distributions of our model parameters can be expressed as

$$\theta_j \sim \mathcal{N}\left(\mu_j, \sigma_j^2\right), \tag{9}$$

$$\sigma_{i,m} \sim |\mathcal{N}\left(\mu_0, \sigma_0^2\right)|. \tag{10}$$

$\sigma_{i,m}$ follows a half-normal distribution, predefined by (μ_0, σ_0^2). We define μ_0 as equal to zero and σ_0 as equal to 10 to avoid providing too much preliminary information for $\sigma_{i,m}$. Each local parameter θ^j has its corresponding μ and σ. In the Bayesian context, *hyperparameter* refers to the parameter of a prior distribution (Gelman et al., 1995). In our model, (μ_0, σ_0^2) and $\left(\mu_j, \sigma_j^2\right)_{\forall j}$

are the hyperparameters. For convenience, we define $\boldsymbol{\phi}$ as a hyperparameter vector containing all the prior distribution's parameters $\left(\mu_0, \sigma_0^2\right)$ for $\sigma_{i,m}$ and $\left(\mu_j, \sigma_j^2\right)$ for θ_j $\forall j$ (highlighted in green in Fig. 8).

Although a different model is derived for each (i, m) combination, some similar characteristics are observed in all of them. First, all on-time performance metrics deteriorate (i.e., display higher values) with increasing weather scores in all data sets. Second, these metrics have their values within the $[0, 1]$ range (recall that μ_{AD} is normalized to $\tilde{\mu}$ in Eq. (4)). Owing to the similarity among all data sets, their hyperparameter vectors $\boldsymbol{\phi}$ can be regarded as exchangeable. This implementation is known as *Bayesian hierarchical modeling* (Gelman et al., 1995). This scheme is convenient when considering the impact of dangerous phenomenon, and beneficial toward computational efficiency. Next, we perform an evaluation process for the deterministic mean trend selection.

3.2 Deterministic mean trend selection

The deterministic mean trend $\mathcal{M}$ is selected to represent the deterioration of airport arrival on-time performance as the weather condition worsens. For the problem at hand, we select five potential candidates, namely two sigmoid functions (logistic and Gompertz), power function, quadratic function, and linear function (Table 3). The number of parameters n (in the third column) corresponds to the number of local parameters shown in the above DAG (Fig. 8). For the Gompertz function, for instance, we have $\boldsymbol{\theta} = [c, x_0, k, y_0]$; hence, $n = 4$.

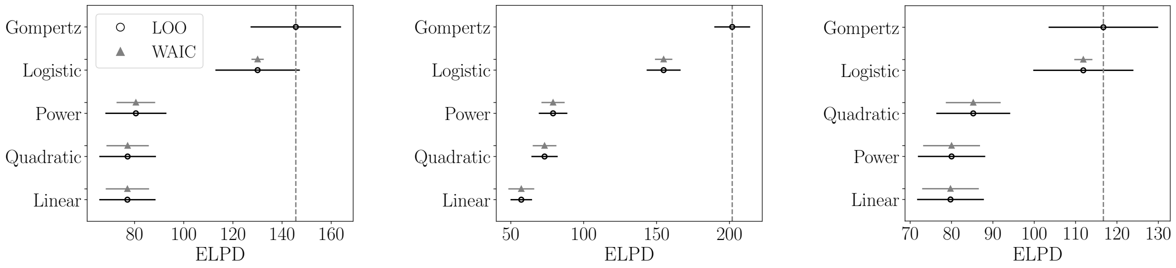
Table 3: Five potential deterministic functions for deterioration trend representation.

Model ($\mathcal{M}$)	Equation	No. of parameters (n)
Logistic	$\frac{c}{1+\mathrm{e}^{(x_0-kx)}} + y_0$	4
Gompertz	$c\exp\left[-\mathrm{e}^{(x_0-kx)}\right] + y_0$	4
Power	$kx^c + y_0$	3
Quadratic	$kx^2 + cx + y_0$	3
Linear	$kx + y_0$	2

The predictive accuracy of each derived model is evaluated using the expected log pointwise predictive density (ELPD), that can be approximated via the widely applicable information criterion (WAIC) and leave-one-out cross-validation (LOO) methods without extra running steps. WAIC and LOO are methods for measuring pointwise out-of-sample prediction accuracy from a fitted Bayesian model using log-likelihood values evaluated at posterior simulations of parameters (Vehtari et al., 2017). Eq. (11) presents the definition of ELPD, where y_u represents unobserved data, P_{true} represents true distribution of unobserved data, and P_{post} represents the posterior distribution,

$$ELPD = E\left[\log P_{post}\left(y_u\right)\right] = \int P_{true}(y_u)\left[\log P_{post}\left(y_u\right)\right] \underset{\cdot}{\mathrm{y}}_u. \tag{11}$$

Fig. 9 shows the predictive accuracy of five models on three arrival traffic performance metrics. The dark circles represent $ELPD_{LOO}$, and the gray triangles represent $ELPD_{WAIC}$. Both of them have similar mean values but different standard errors. The vertical dashed line indicates the most accurate mean ELPD value for each arrival performance metric.



(a) $\tilde{\mu}$

(b) RT_c

(c) RT_d

Figure 9: Evaluation results, predictive accuracy (ELPD) via LOO and WAIC.

For all three traffic metrics, the growth functions (Logistic, Gompertz) show their superiority in predictive accuracy. Among the growth functions, Gompertz shows a slight advantage in $\tilde{\mu}$ and RT_d compared with the logistic function. For RT_c, the Gompertz function even outperforms the logistic function in terms of predictive accuracy. It is worth mentioning, if we apply this model to data pertaining to other airports, the evaluation results might be different since the aviation weather impact study is location-dependent. To sum up, we evaluate the predictive accuracy of five deterministic mean trend functions in this subsection. The results show that the Gompertz function has the best performance for the problem at hand. Thus, only the Gompertz function will be considered in the subsequent analyses and discussions.

3.3 Posterior computation via the No-U-Turn Sampler (NUTS)

After selecting the Gompertz function as our deterministic mean trend, we now discuss the posterior distribution computation for the Bayesian inference. For our proposed framework, calculating the posterior distributions of the underlying system can be intractable (Eq. (8)). As such, we apply a Markov chain Monte Carlo (MCMC) algorithm, in particular the No-U-Turn sampler (NUTS) (Hoffman et al., 2014), as our posterior sampling method. NUTS is developed based on the Hamiltonian Monte Carlo (HMC), a sophisticated MCMC method that does not suffer from randomness and sensitivity to correlated parameters. Compared to HMC, NUTS is more advanced because of its adaptive property. When using MCMC, it is important to ensure model convergence. The informative level of priors is vital for the model's convergence performance (Gelman et al., 2017; van de Schoot et al., 2021). In this work, we use the Gelman–Rubin diagnostic ($\hat{R}$) and prior predictive check to select the appropriate hyperparameters for our weather impact quantification model. The results are shown in Fig. 10, and described below.

To the left-hand side of Fig. 10, the x-axis shows the corresponding model parameters when the Gompertz function is used (including local parameters and model standard deviation, which represents the model variance) and the y-axis displays the Gelman-Rubin diagnostic $\hat{R}$ for the MCMC convergence evaluation. $\hat{R}$ is expressed as

$$\hat{R} = \frac{\hat{V}}{W}, \tag{12}$$

where W is the within-chain variance and $\hat{V}$ is the posterior variance estimated between traces. The red dashed line in Fig. 10 (where $\hat{R} = 1$) is the convergence threshold line, below which all traces converge. $\hat{R}$ greater than one indicates that one or more chains have not yet converged,

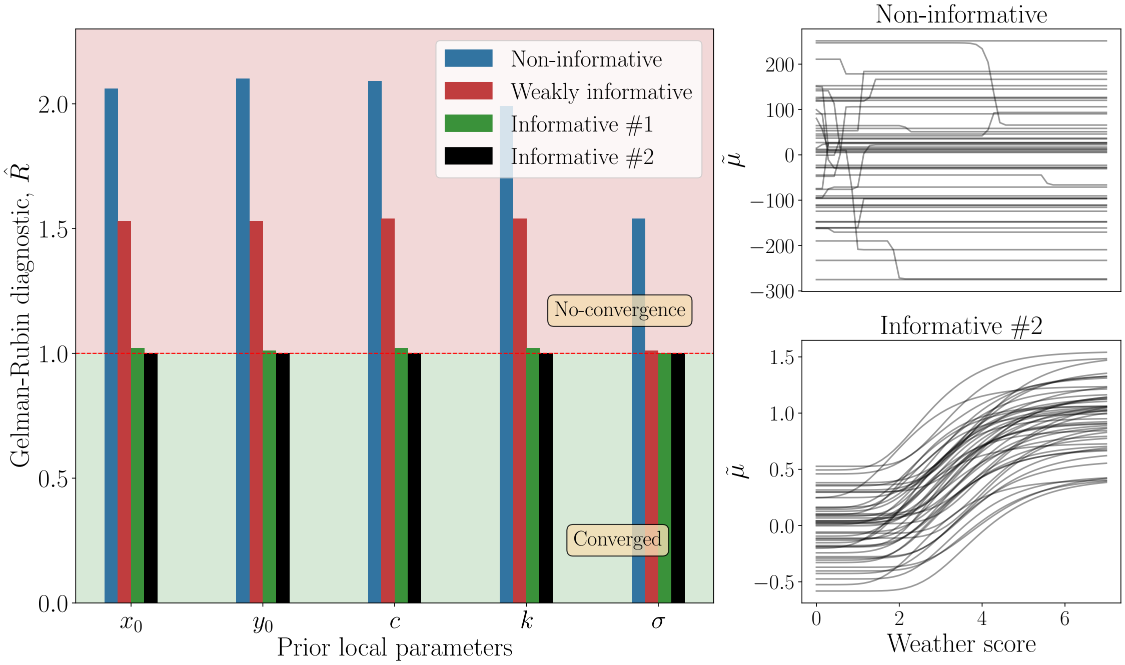


Figure 10: The Gelman–Rubin diagnostic and prior predictive check for $\tilde{\mu}$.

while $\hat{R}$ greater than 1.2 is regarded as rigorously non-convergent (Brooks and Gelman, 1998). For our case, we check four sets of prior's convergence performance on $\tilde{\mu}$, which depends on their informative level. The prior sets include non-informative ($\theta_j \sim \mathcal{N}(0, 100^2)$), weakly informative ($\theta_j \sim \mathcal{N}(0, 10^2)$), informative #1 ($\mu_j = [3, 0, 0, 1]$, $\sigma_j = 1$), and informative #2 ($\mu_j = [3, 0, 0, 1]$, $\sigma_j = [0.5, 0.25, 0.1, 0.1]$). Informative priors are decided based on the geometrical meaning of each model parameter, data observation, and the range of expected outputs. Results show that when non-informative or weakly informative priors are used, the model will not converge. A full convergence for all five parameters is only observed when the informative #2 prior is used.

To illustrate the convergence, we draw 50 samples from non-informative priors and informative #2 prior, then plot the generated function profiles in the right-hand side of Fig. 10. This process is called the *prior predictive check* (van de Schoot et al., 2021). With non-informative priors, the output range is too permissive, and the prior predictive curves are widely scattered within the $[-250, 200]$ range. In contrast, the prior predictive curves of informative #2 priors yield a more definitive trend and narrower distribution, without losing the flexibility in fitting the data. Thus, the informative #2 priors are used in our model derivation to ensure the model's convergence performance.

4 Results and discussions

In this section, we present our model applications and demonstrate its use to evaluate the impacts of adverse weather conditions on aircraft arrival performance. As mentioned in Section 2, arrival flight information data and weather data pertaining to HKIA from 2017 to 2018 are used to generate results presented in this section. Table 4 shows the detailed model setup. Input x is the weather score, and output y is each airport arrival on-time performance metric, i.e., $\tilde{\mu}$, RT_c, or RT_d. The numbers of data points shown here refer to those obtained from the undersampling and oversampling procedures, which are mentioned in Section 2.3.3, for each weather score. In other words, they refer

Table 4: Model setup for Bayesian parameter tuning.

	Feature	Content/value
Dataset	x	Weather score
	y	$\tilde{\mu}$, RT_c, RT_d
	No. of data (without dangerous phenomena)	184
	No. of data (Thunderstorm)	90
	No. of data (Shower)	208
	No. of data (Cumulonimbus)	113
Model	Prior	Informative #2
	Deterministic mean trend	Gompertz
	Posterior computation	No-U-Turn Sampler
	Tuning steps	2,000
	Drawing steps	1,000
	Number of chains	4

to the heights of green bars in Fig. 6. The number of Markov chains is four, and for each chain, there are 2,000 tuning steps and 1,000 drawing steps.

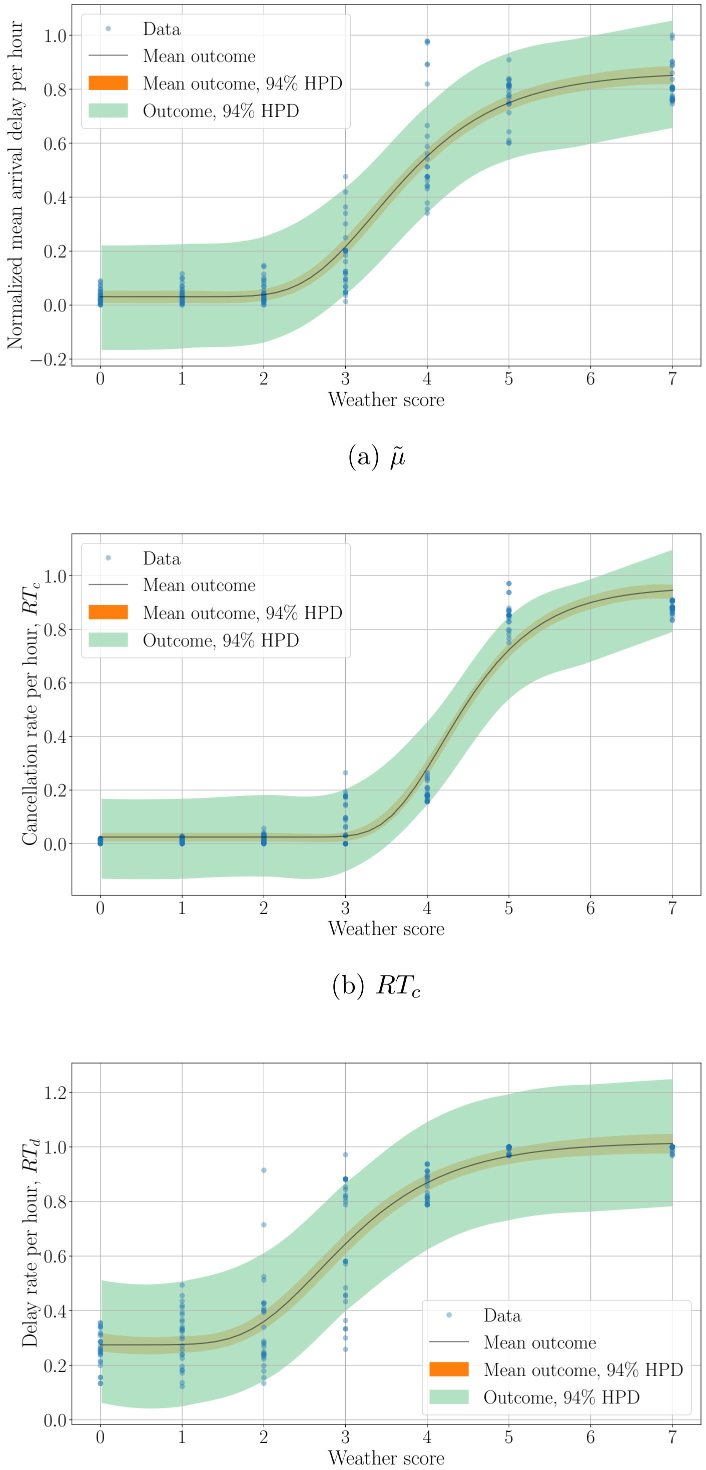
Three key results are presented in this section, namely the weather impact quantification without dangerous phenomena, the sensitivity of traffic metrics towards adverse weather conditions, and the impact quantification of dangerous phenomena.

4.1 Weather impact quantification without dangerous phenomena

To quantify the weather impact when no dangerous phenomenon is present, we generate the distribution of possible outputs, using 4,000 random draws for $\tilde{\mu}$, RT_c, and RT_d, with the obtained posterior model parameters. The mean output, its highest posterior density (HPD), and the HPD interval are illustrated in Fig. 11.

As Fig. 11 illustrates, the proposed method can adequately capture the nonlinear growth property of aircraft arrival traffic. The observed phase transition separates the situations where airport arrival performance is minimally affected and those when the impact is significant. The airport arrival performance is relatively insensitive to weather scores before and after the transition. It can also describe the uncertainty owing to the general feature of the Bayesian approach. Visually, we can observe that the fitting performances for $\tilde{\mu}$ (Fig. 11a) and RT_d (Fig. 11c) are better than that of RT_c (Fig. 11b), where the latter shows poor-fitting at around weather scores four and five. The possible reason is that the steepness level at the transition phase for the RT_c is too extreme, which might be more challenging to model. Increasing the model fidelity, such as adding a new parameter to capture the slope variation, adversely affects model convergence. Hence, the model showed in Fig. 11b is one that balances model accuracy and convergence.

It is easy to see that the fitting performance is superior in the lower range of weather scores. As we can observe from the range of HPD, the cancellation rate has the narrowest bandwidth, while RT_d has the widest bandwidth. This observation suggests that the cancellation rate is the most certain among the three metrics. In a real situation, cancellation is always the last option for flights and tends to be avoided, which causes the lowest bandwidth of HPD for RT_c. On the other hand, the higher uncertainty concerning RT_d suggests that there are a variety of causes for delays other than weather conditions, which is consistent with real situations.



(a) $\tilde{\mu}$

(b) RT_c

(c) RT_d

Figure 11: Weather impact without dangerous phenomenon

4.2 Traffic metric sensitivity towards adverse weather

In this section, we further interpret the sensitivities of the three metrics ($\tilde{\mu}$, RT_c, and RT_d) concerning adverse weather by visualizing the probability density of the local parameter samples from the posterior distribution (Fig. 12). First, we inspect the local parameter x_0, which represents the horizontal location of the model's midpoint. For the problem at hand, x_0 refers to how fast these air traffic performance metrics react towards the weather score. A larger x_0 means a slower reaction. Fig. 12 shows that RT_c has the slowest reaction, with a mean value of 5.73. On the other hand, RT_d and $\tilde{\mu}$ have mean values of 2.84 and 3.91, respectively, which indicates that more significant deterioration in these metrics is observed at lower weather scores.

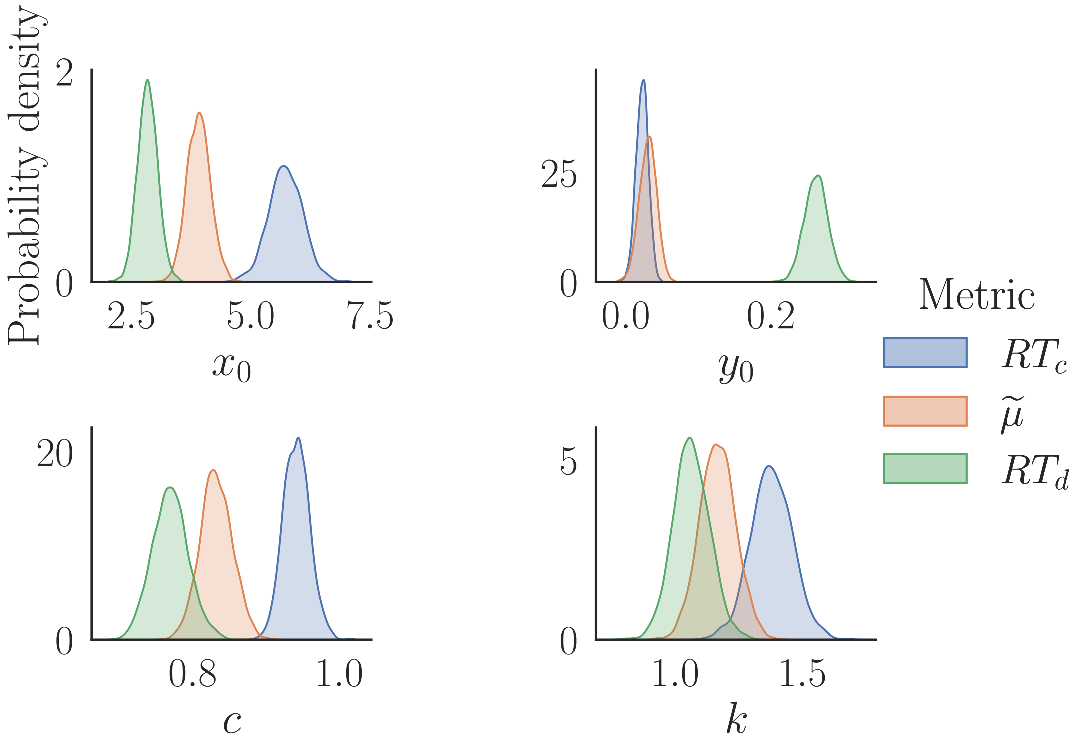


Figure 12: Comparison of the parameter posterior distribution for three arrival performance metrics

The local parameter y_0 indicates the offset distance between the curve and $y = 0$. For this parameter, the delay rate distribution has the highest mean value at 0.26. This value means that at HKIA, even at a weather score of zero, there is an average of 26% arrival flight delays per hour. These delays might be due to various reasons other than weather. The third local parameter, c, refers to the scaling factor for the exponential component; it can stretch or contract the curve to fit within the appropriate value range of the output. Since our outputs range within $[0, 1]$, the c value should never exceed one. The fourth local parameter, k, corresponds to the steepness of the growth function, i.e., how fast the performance metric deteriorates once it reaches the transition point. In this case, RT_c has the steepest curve and hence the largest k value.

Attaching physical intuitions to the model parameter's geometric interpretation, as discussed above, can be used to compare the characteristics of different airports. One potential application is to use the same framework to evaluate and compare the airport arrival on-time performance among airports within a multi-airport system (metroplex). A *metroplex* refers to a certain geographic area covering several airports within close vicinity to each other. As such, a similar weather condition can be reasonably assumed for all these airports. For example, let us consider an arbitrary airport A within the vicinity of HKIA. Upon deriving the models for the same time range, we find that for μ_{AD}, $x_{0,HKIA} > x_{0,A}$. This result indicates that while airport A's hourly mean arrival delays start to increase significantly following adverse weather, HKIA can maintain its regular operation. Furthermore, if $y_{0,HKIA} > y_{0,A}$, other factors, besides weather, have a stronger influence on HKIA's arrival delay performance than airport A's.

4.3 Dangerous phenomenon impact quantification

In this section, we train three new models (for data with the thunderstorm, shower, and cumulonimbus weather conditions) following the setting in Table 4 with the Gompertz equation. Owing to the lack of data at higher weather scores for these three weather conditions, we present the posterior distributions of y when x equals zero, which are shown in Fig. 13. Since cancellation barely happens when the weather score equals zero, we select μ_{AD} and RT_d as our outputs. The blue distribution is for no dangerous phenomenon, while orange is thunderstorm, green is shower, and red is cumulonimbus. The *mean shifts* in the airport arrival on-time performance metrics, which are described in Section 3, are indicated by the arrows.

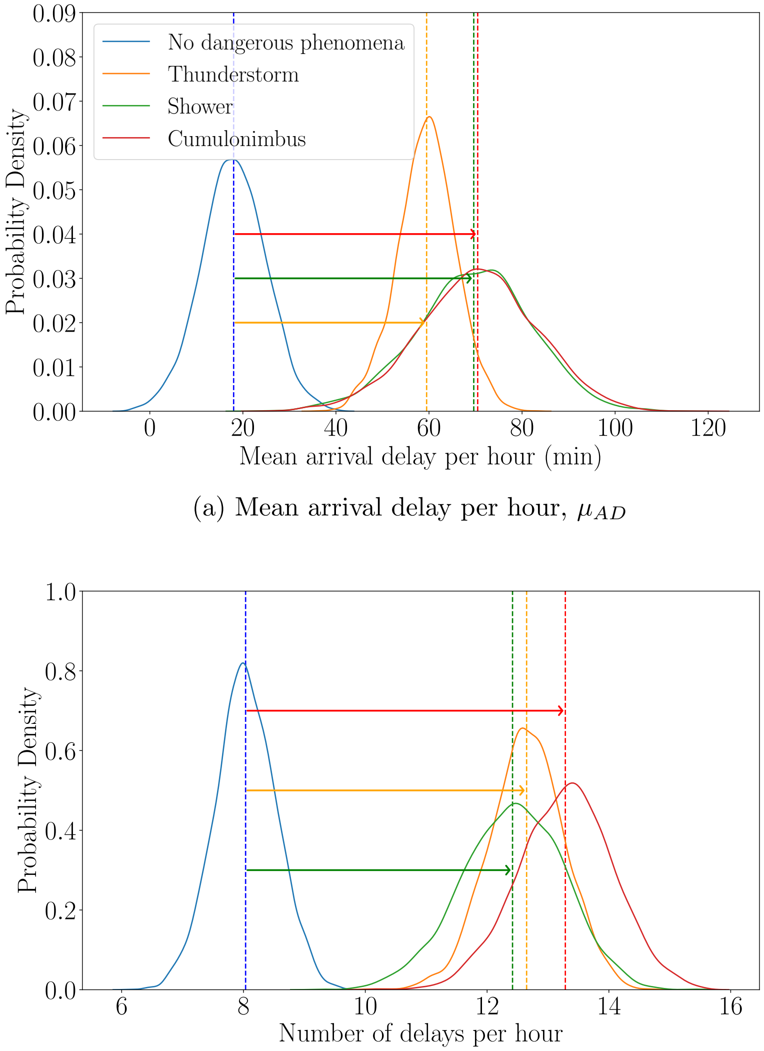
Without any dangerous phenomena, the mean arrival delay per hour (in minutes) and the number of delays per hour are shown to have the lowest mean values, i.e., the best performance, which is not surprising. From observing the mean shifts, shower and cumulonimbus have more significant impacts on the mean arrival delay per hour, while the impact of thunderstorm is relatively more moderate. Without any dangerous phenomena, the mean arrival delay is around 20 minutes. Thunderstorm adds another 40 minutes, whereas shower and cumulonimbus, which are shown to have similar impacts, cause the mean arrival delay to increase to 70 minutes. For the number of delays per hour, while no dangerous phenomenon condition has around eight delayed flights per hour, those with dangerous phenomena are expected to have 10–15 delayed flights per hour. We observe that, for cumulonimbus, the number of delayed flights increases approximately to 15, i.e., almost double. The other two dangerous phenomena increase the number of delayed flights by around four per hour. The weather impact quantification and insights presented above would not be possible if we were to use the ATMAP weather score as shown in Fig. 4. These results highlight the benefits of the proposed hierarchical approach to evaluate and quantify the impacts of dangerous phenomena on airport arrival on-time performance by comparing their performance to when no dangerous phenomena are observed, thereby addressing some existing research gaps in the literature.

5 Conclusion

In this paper, we propose a hierarchical Bayesian model to quantify the relationship between weather impact and airport arrival on-time performance, with a case study at HKIA. The key contributions from this paper are summarized below:

1. This work offers the first and largest-scale quantitative study on weather impact quantification based on data pertaining to HKIA. Since aviation weather studies are location-dependent, a study specific to HKIA can assist in policy- and decision-making analyses by the air transportation authority in Hong Kong.

2. Compared with previous studies using METAR data, the growth function employed in this work can describe the nonlinear property of airport arrival performance towards hazardous



(a) Mean arrival delay per hour, μ_{AD}

(b) Number of delays per hour

Figure 13: The impact of dangerous phenomenon at weather score = 0.

weather, where a trend akin to a phase transition is observed. Furthermore, applying the Bayesian approach ensures the model's ability to quantify the uncertainty.

3. The interpretability of growth function parameters allows us to compare the airport arrival on-time performance's tolerance under weather impact by comparing the posterior distribution of a certain parameter across different performance metrics.

4. The developed framework also includes a systematic dangerous phenomenon impact quantification procedure, offering a more comprehensive weather impact quantification that includes a wider spectrum of weather scores compared to other similar studies. Note that dangerous phenomena are commonly excluded in past aviation weather impact studies.

5. This work will enrich the literature on terminal arrival delays, which is still relatively scarce as compared to other types of flight delays.

Specific to HKIA conditions, our results reveal several essential conclusions about adverse weather impacts. First, the cancellation rate has the slowest reaction to adverse weather among these three metrics. Moreover, even when the weather score equals zero, the delay rate per hour is still noticeable because of other operational reasons. Furthermore, for weather score equals zero, cumulonimbus has the highest impact on delay rate, while both shower and cumulonimbus show significant impact on mean arrival delay per hour.

The use of METAR data in this study adds to the generalizability of the developed framework, since the data are available for more airports. However, despite its versatility, the resolution is low. METAR data are often recorded hourly and limited by their location and features. For instance, wake turbulence properties and separation requirements are essential issues affected by weather (Hon and Chan, 2017; Hon et al., 2022), yet these are not reflected in METAR.

It is a widely adopted practice to use results from computational models to support decision-making and policy analyses. The work presented in this paper is particularly relevant to assessing air traffic management and delay mitigation strategies, especially when combined with other models describing air traffic movements around HKIA. Quantifying the weather impact will help ATC at the strategic and pre-tactical stages of flight planning and provide more accurate weather constraints for scheduling aircraft arrivals. This study is currently underway in the authors' research group (e.g., Lui et al., 2020a,b; Hon, 2021).

There are several potential developments for this work. The model is developed to be generalizable, such that it can quantify the weather impacts for other airports using the relevant air traffic and weather data. Note that depending on the data, the suitable deterministic mean trend function $\mathcal{M}$ might be different for different airports. Similarly, PDFs other than Gaussian can also be assumed. The same selection criterion, i.e., the ELPD method, can still be used. The Bayesian inference approach will automatically find the appropriate model parameters based on data. Upon extending the study to other airports, we can then compare and analyze their airport arrival on-time performance, as discussed in Section 4.2. This comparison can also include the impact of dangerous phenomena. This comparison, however, is beyond the scope of the current paper due to the lack of detailed flight schedule data of other airports. Besides, the Bayesian approach employed in this framework will easily update model parameters when newly-observed data are available. This updateability property, as opposed to a static model, will make the model adaptable to future changes due to, for instance, climate change. Finally, when higher-resolution weather data are available, we can increase the level of fidelity of our model to enable modeling the weather–air traffic interaction more realistically.

List of notations

SAT	scheduled arrival time
AAT	actual arrival time
N	total number of scheduled arrival flights in specific hour
μ_{AD}	mean arrival delay per hour (minutes)
RT_c	cancellation rate per hour
RT_d	delay rate per hour
$\tilde{\mu}$	normalized mean arrival delay per hour
y	output, individual air traffic performance metric
x	input, weather score
$\mathcal{N}$	Normal distribution
$\mathcal{M}$	deterministic mean trend model
f	index for the flight
m	index for the arrival performance metric
i	index for the data set
j	index for the local parameter
θ	local parameter
$\boldsymbol{\theta}$	local parameter vector
n	number of local parameters
$\mathcal{D}$	arbitrary dataset
$\boldsymbol{\phi}$	hyperparameter vector
x_0, y_0, c, k	local model parameters for Gompertz function
$ELPD$	expected log predictive density
y_u	unobserved data
P_{post}	posterior distribution
P_{true}	true distribution of unobserved data
HPD	highest posterior density

Declaration of competing interest

The authors declare no competing interests in the development of this research.

Acknowledgements

The authors would like to thank the Hong Kong Observatory (HKO) for providing the data used in this study. The work is supported by the HKUST Start-Up Grant (Project No. R9354) and Research Support from Chau Hoi Shuen Foundation.

References

Allan, S., Gaddy, S., Evans, J.. Delay causality and reduction at the New York City airports using terminal weather information systems. Technical Report; Lincoln Laboratory, Massachusetts Institute of Technology; 2001.

Arıkan, M., Deshpande, V., Sohoni, M.. Building reliable air-travel infrastructure using empirical data and stochastic models of airline networks. Operations Research 2013;61(1):45–64.

Bishop, C.M.. Pattern recognition. Machine Learning 2006;128(9).

Borsky, S., Unterberger, C.. Bad weather and flight delays: The impact of sudden and slow onset weather events. Economics of Transportation 2019;18:10–26.

Brooks, S.P., Gelman, A.. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 1998;7(4):434–455.

Bureau of Transportation Statistics, . On-Time Performance - Reporting Operating Carrier Flight Delays at a Glance. `https://www.transtats.bts.gov/HomeDrillChart_Month.asp`; 2021. Accessed Sep 3rd, 2021.

Buxi, G., Hansen, M.. Generating probabilistic capacity profiles from weather forecast: A design-of-experiment approach. In: Proc. of USA/Europe Air Traffic Management Research & Development Seminar. 2011. p. 30–40.

Chan, P.W., Hon, K.K.. Performance of super high resolution numerical weather prediction model in forecasting terrain-disrupted airflow at the Hong Kong International Airport: case studies. Meteorological Applications 2016;23(1):101–114.

Easterling, D.R., Meehl, G.A., Parmesan, C., Changnon, S.A., Karl, T.R., Mearns, L.O.. Climate extremes: observations, modeling, and impacts. Science 2000;289(5487):2068–2074.

Erzberger, H., Lee, H.Q.. Terminal-area guidance algorithms for automated air-traffic control. volume 6773. NASA Technical Note TN D-6773, 1972.

EUROCONTROL, . Algorithm to describe weather conditions at European airports. https://www.eurocontrol.int/sites/default/files/publication/files/algorithm-met-technical-note.pdf; 2011. Accessed 15 March 2021.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.. Bayesian data analysis. Chapman and Hall/CRC, 1995.

Gelman, A., Simpson, D., Betancourt, M.. The prior can often only be understood in the context of the likelihood. Entropy 2017;19(10):555.

Gopalakrishnan, K., Li, M.Z., Balakrishnan, H.. Network-centric benchmarking of operational performance in aviation. Transportation Research Part C: Emerging Technologies 2021;126:103041.

Grabbe, S., Sridhar, B., Mukherjee, A.. Clustering days and hours with similar airport traffic and weather conditions. Journal of Aerospace Information Systems 2014;11(11):751–763.

Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. Array programming with NumPy. Nature 2020;585(7825):357–362.

He, H., Bai, Y., Garcia, E.A., Li, S.. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE; 2008. p. 1322–1328.

Hoffman, M.D., Gelman, A., et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 2014;15(1):1593–1623.

Hon, K.K.. Artificial intelligence prediction of air traffic flow rate at the Hong Kong International Airport. In: IOP Conference Series: Earth and Environmental Science. IOP Publishing; volume 865; 2021. p. 012051.

Hon, K.K., Chan, P.W.. Aircraft wake vortex observations in Hong Kong. J Radar 2017;6(6):709–718.

Hon, K.K., Chan, P.W.. Historical analysis (2001–2019) of low-level wind shear at the Hong Kong International Airport. Meteorological Applications 2022;29(2):e2063. doi:doi:https://doi.org/10.1002/met.2063.

Hon, K.K., Chan, P.W., Chim, K.C., De Visscher, I., Thobois, L., Rooseleer, F., Troiville, A.. Wake vortex measurements at the Hong Kong International Airport. In: AIAA Scitech 2022 Forum. 2022. p. 2011.

ICAO, . Global Air Navigation Plan for CNS/ATM Systems. `https://www.icao.int/publications/Documents/9750_2ed_en.pdf`; 2019. Accessed Sep 3rd, 2021.

ICAO, . Economic Development, FEB 2022: Air Transport Monthly Monitor. `https://www.icao.int/sustainability/Documents`; 2022. Accessed Mar 8th, 2022.

Kistan, T., Gardi, A., Sabatini, R., Ramasamy, S., Batuwangala, E.. An evolutionary outlook of air traffic flow management techniques. Progress in Aerospace Sciences 2017;88:15–42.

Krawczyk, B.. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 2016;5(4):221–232.

Krozel, J., Penny, S., Prete, J., Mitchell, J.S.. Automated route generation for avoiding deterministic weather in transition airspace. Journal of Guidance, Control, and Dynamics 2007;30(1):144–153.

Kumar, R., Carroll, C., Hartikainen, A., Martin, O.. ArviZ a unified library for exploratory analysis of Bayesian models in Python. Journal of Open Source Software 2019;4(33):1143. URL: `https://doi.org/10.21105/joss.01143`. doi:doi:10.21105/joss.01143.

Lemetti, A., Polishchuk, T., Polishchuk, V., Sáez García, R., Prats Menéndez, X.. Identification of significant impact factors on Arrival Flight Efficiency within TMA. In: ICRAT 2020: papers & presentations. 2020. p. 1–8.

Likas, A., Vlassis, N., Verbeek, J.J.. The global k-means clustering algorithm. Pattern recognition 2003;36(2):451–461.

Lo, N.. Constraints on HKIA Dual-runway Operation and Airspace Issue. `https://www.cad.gov.hk/english/20150409.html`; 2015. Accessed Mar 8th, 2022.

Lui, G.N., Klein, T., Liem, R.P.. Data-driven approach for aircraft arrival flow investigation at terminal maneuvering area. In: AIAA Aviation 2020 Forum. 2020a. p. 2869.

Lui, G.N., Liem, R.P., Hon, K.K.. Towards understanding the impact of convective weather on aircraft arrival traffic at the Hong Kong International Airport. In: IOP Conference Series: Earth and Environmental Science. IOP Publishing; volume 569; 2020b. p. 012067.

Mason, K.J.. Observations of fundamental changes in the demand for aviation services. Journal of Air Transport Management 2005;11:19–25. doi:doi:10.1016/j.jairtraman.2004.11.007.

McCarthy, J., Wilson, J.W., Fujita, T.T.. The joint airport weather studies project. Bulletin of the American Meteorological Society 1982;63(1):15–22.

McCrea, M.V., Sherali, H.D., Trani, A.A.. A probabilistic framework for weather-based rerouting and delay estimations within an airspace planning model. Transportation Research Part C: Emerging Technologies 2008;16(4):410–431.

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

Murça, M.C.R.. Identification and prediction of urban airspace availability for emerging air mobility operations. Transportation Research Part C: Emerging Technologies 2021;131:103274.

Murphy, K.P.. Conjugate Bayesian analysis of the Gaussian distribution. Technical Report; University of British Columbia; 2007.

de Oliveira, M., Eufrásio, A.B.R., Guterres, M.X., Murça, M.C.R., de Arantes Gomes, R.. Analysis of airport weather impact on on-time performance of arrival flights for the Brazilian domestic air transportation system. Journal of Air Transport Management 2021;91:101974.

Pang, Y., Liu, Y.. Conditional generative adversarial networks (CGAN) for aircraft trajectory prediction considering weather effects. In: AIAA Scitech 2020 Forum. 2020. p. 1853.

Pang, Y., Yao, H., Hu, J., Liu, Y.. A recurrent neural network approach for aircraft trajectory prediction with weather features from Sherlock. In: AIAA Aviation 2019 Forum. 2019. p. 3413.

Pang, Y., Zhao, X., Yan, H., Liu, Y.. Data-driven trajectory prediction with weather uncertainties: A Bayesian deep learning approach. Transportation Research Part C: Emerging Technologies 2021;130:103326.

Pfeil, D.M., Balakrishnan, H.. Identification of robust terminal-area routes in convective weather. Transportation Science 2012;46(1):56–73.

Rebollo, J.J., Balakrishnan, H.. Characterization and prediction of air traffic delays. Transportation Research Part C: Emerging Technologies 2014;44:231–241.

Reitmann, S., Alam, S., Schultz, M.. Advanced quantification of weather impact on air traffic management. In: ATM Seminar. 2019. p. 20–30.

Robinson, P.J.. The influence of weather on flight operations at the Atlanta Hartsfield International Airport. Weather and Forecasting 1989;4(4):461–468.

Rodriguez-Sanz, A., Cano, J., Fernandez, B.R.. Impact of weather conditions on airport arrival delay and throughput. Aircraft Engineering and Aerospace Technology 2021;.

Rudin, C.. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 2019;1(5):206–215.

Ryley, T., Baumeister, S., Coulter, L.. Climate change influences on aviation: A literature review. Transport Policy 2020;92:55–64.

Salvatier, J., Wiecki, T.V., Fonnesbeck, C.. Probabilistic programming in Python using PyMC3. PeerJ Computer Science 2016;2:e55.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M.G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. Bayesian statistics and modelling. Nature Reviews Methods Primers 2021;1(1):1–26.

Schultz, M., Lorenz, S., Schmitz, R., Delgado, L.. Weather impact on airport performance. Aerospace 2018;5(4):109.

Schultz, M., Reitmann, S., Alam, S.. Predictive classification and understanding of weather impact on airport performance through machine learning. Transportation Research Part C: Emerging Technologies 2021;131:103119.

Shun, C., Chan, P.. Applications of an infrared Doppler lidar in detection of wind shear. Journal of Atmospheric and Oceanic Technology 2008;25(5):637–655.

Song, L., Greenbaum, D., Wanke, C.. The impact of severe weather on sector capacity. In: 8th USA/Europe Air Traffic Management Research and Development Seminar (ATM2009), Napa, California, USA. 2009. p. 1–8.

Spinardi, G.. Up in the air: Barriers to greener air traffic control and infrastructure lock-in in a complex socio-technical system. Energy Research & Social Science 2015;6:41–49.

Sternberg, A., Carvalho, D., Murta, L., Soares, J., Ogasawara, E.. An analysis of Brazilian flight delays based on frequent patterns. Transportation Research Part E: Logistics and Transportation Review 2016;95:282–298.

Vehtari, A., Gelman, A., Gabry, J.. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 2017;27(5):1413–1432.

Zhao, X., Yan, H., Li, J., Pang, Y., Liu, Y.. Spatio-temporal anomaly detection, diagnostics, and prediction of the air-traffic trajectory deviation using the convective weather. In: Proceedings of the Annual Conference of the PHM Society. volume 11; 2019. p. 1–8.