Assessing the risk of windshear occurrence at HKIA using rare-event logistic regression

Feng Chen1 | Haorong Peng1 | Pak-wai Chan2 | Xiaoxiang Ma1 | Xiaoqing Zeng1

1The Key Laboratory of Infrastructure Durability and Operation Safety in Airfield of CAAC, Tongji University, Shanghai, China
2Hong Kong Observatory, Hong Kong Special Administrative Region, Hong Kong (SAR), China

Abstract

Low-level windshear poses serious hazards to the flight safety of civil aircraft. The objective of the study was to assess the risk of windshear occurrence within 1 hr at runway corridors at Hong Kong International Airport (HKIA). Windshear report records from HKIA and weather data around HKIA were obtained from Hong Kong Observatory (HKO) between January 2016 and December 2018. The rare-event logistic regression was employed to evaluate windshear risk considering the imbalanced nature of the data. Several explanatory variables were found to be significantly correlated with the risk of the hourly occurrence of the windshear. Results indicated that at HKIA, March, April, July and August are associated with a higher likelihood of windshear events, while November suffered a significantly lower windshear risk. Besides, winds from the east, southeast, south and southwest are more likely to cause windshear events at HKIA. However, the study found that it was hard to predict windshear occurrence precisely only with the hourly weather variables due to the transient and sporadic nature of windshear. Therefore, more refined data are needed to improve the prediction performance of regression models for windshear occurrence. The present study is the first to analyse the windshear occurrence at airports using the rare-event logistic regression method. Findings from the study can provide some references and valuable information for related researchers to understand low-level windshear.

KEYWORDS
complex terrain, low-level windshear, rare-event logistic regression, risk analysis

1 | INTRODUCTION

Low-level windshear refers to a sudden change in wind speed and/or direction that occurs < 1,600 ft and within 3 nautical miles (nmi) from the runway thresholds of an airport (ICAO, 2016). An aircraft would experience a sudden change in its lift force if it encounters a low-level windshear, which poses potential hazards to flight safety during take-off and landing. Hong Kong International Airport (HKIA) has been significantly impacted by low-level windshear events since its opening in August 1998. Nearly one in 500 flights at HKIA reportedly encountered significant windshear,
which might bring about disastrous consequences. Two cases of tail strike of aircraft due to windshear have been documented in previous studies (Chan, 2012, 2014a).

HKIA is one of the busiest airports in the world. It is crucial to alleviate the adverse influence of windshear on the flight safety and operational efficiency of HKIA. Therefore, several studies have focused on the low-level windshear occurring at HKIA. Shun and Chan (2008) summarized that about 70% of all the windshear events reported by pilots were induced by the complex terrain (Lantau Island in Figure 1), and approximately 20% were caused by the sea breeze, accounting for the first two major causes of windshear events at HKIA (Shun and Chan, 2008). Since terrain-induced windshear is the major type of low-level windshear at HKIA with the characteristics of small spatial and temporal scale, it is extremely difficult to detect and predict precisely terrain-induced windshear. Consequently, many efforts have been put into investigating terrain-induced low-level windshear. Considering that terrain-induced windshear events mainly occurred in non-rainy weather conditions, a doppler light detection and ranging (LIDAR) system was applied to detect windshear and issue warnings operationally for HKIA (Chan et al., 2006). The scanning strategy and windshear detection algorithm have been continuously improved to increase the detection rate and reduce the rate of missing reports of low-level windshear events (Chan et al., 2011; Hon and Chan, 2014; Lee and Chan, 2014; Li et al., 2018; Wu and Hon, 2018). However, as Chan (2017) indicated, the existing technology still needs to be enhanced for the accurate detection and forecasting of low-level windshear.

Generally, the causes of low-level windshear could be classified into two categories: the geographical environment around an airport including complex terrain and buildings; and adverse weather conditions, which contains sea breezes and microbursts (Chen et al., 2019). Many studies have been conducted to analyse the influence of these factors on the occurrence of low-level windshear. Numerical simulation was employed to simulate terrain-induced windshear (Lei et al., 2013, 2017; Chan et al., 2019) and the low-level effects caused by buildings (Chan et al., 2010; Leung et al., 2012; Li and Chan, 2012). Besides, some typical windshear cases had been studied based on the observation results and flight data (Chan, 2014b, 2017). Undoubtedly, these researches have contributed to the overall understanding of low-level windshear events, which facilitated the detection and prediction of windshear occurrences at HKIA. However, to the authors’ knowledge, few, if any, reported studies examine the association between windshear occurrence and its background weather conditions, such as wind speed, wind direction and temperature.

Logistic regression has proven to be a popular and reliable method to reveal the relationship between a dichotomous response and its explanatory variables in various research fields (Wong et al., 2009; Sze et al., 2014; DeVault et al., 2016; Wang et al., 2019; Zhai et al., 2019). Regarding the risk analysis associated with airports or aircraft, multivariate logistic regression was applied to develop an accident frequency model with the purpose of assessing the risks related to aircraft accidents and in the vicinity of airports (Wong et al., 2009). Besides, DeVault et al. (2016) evaluated the

**FIGURE 1** Location of Hong Kong International Airport (HKIA)
influence of risk factors involved in wildlife strikes on the likelihood of aircraft damage using binary logistic regression analysis. However, when it comes to rare events with dozens to thousands of times fewer events than non-events, maximum likelihood estimates of ordinary logistic regression may be biased, which could sharply underestimate the probabilities of rare events (King and Zeng, 2001a). To deal with this problem, King and Zeng (2001a) proposed a rare-event logistic regression and applied it to political science (King and Zeng, 2001b), which showed great potential for predicting rare events. Over the past two decades or so, rare-event logistic regression has been applied in many fields (Van den Eeckhaut et al., 2006; Bai et al., 2011; Guns and Vanacker, 2012; Yang et al., 2015; Ren et al., 2016; Theofilatos et al., 2019). Van den Eeckhaut et al. (2006) demonstrated that the rare-event logistic regression is capable of predicting hillslope sections prone to landslide. Based on the rare-event logistic regression proposed by King and Zeng (2001a), Guns and Vanacker (2012) developed rare-event logistic regression with replications that could avoid the instability of the results due to sampling bias. Yang et al. (2015) explored the crash risk of highway work zones with relatively short durations by employing the rare-event logistic regression model. In addition, rare-event logistic regression was applied to predict red-light-running-related crashes at signalized intersections (Ren et al., 2016), which outperformed the ordinary logistic regression method with a significant improvement in the prediction rate. Moreover, this method was also used to investigate the crash occurrence on motorways with real-time traffic data (Theofilatos et al., 2019). All those aforementioned studies indicated that the rare-event logistic regression is a promising approach with reliable results.

Typically, the windshear occurrence encountered by an aircraft at HKIA is a rare event. Hours with windshear occurrence at HKIA accounted for 2.64% of the total hours during the period 2016–2018 (Table 2). Therefore, rare-event logistic regression should be employed when investigating the association between potential factors with a windshear occurrence. The objective of the study is to develop a rare-event logistic regression model in order to assess the risk of windshear occurrence at HKIA. Windshear occurrence records were obtained from windshear reports by the pilots of flights at HKIA. Potential risk factors included hourly mean wind speed, hourly prevailing wind direction, hourly mean temperature and the month of windshear occurrence, which were extracted from background weather condition profiles provided by Hong Kong Observatory (HKO). The model estimation results indicated wind directions and months with a higher risk of windshear occurrence in a quantitative form. The height of windshear occurrence, which was not involved in the rare-event logistic regression model due to lack of the control group, was also discussed descriptively. It is expected that the study would provide valuable information to assist related researchers and manage departments in developing appropriate countermeasures for low-level windshear.

The remainder of this paper is organized as follows. Section 2 introduces the method used. Section 3 provides a data description. The development of the logistic regression model is presented in Section 4, followed by the model interpretation and discussion in Section 5. Finally, some conclusions are given in Section 6.

# METHODOLOGY

The study assesses the influence of background weather conditions on the risk of windshear occurrence at HKIA using a rare-event logistic regression model. The response variable windshear occurrence (WSO) was defined as whether there was at least one windshear event occurring on runway corridors at HKIA within the time window (1 hr). More specifically, 24 hr in a day are divided into 24 time windows from 0000 to 2400 Hong Kong time (HKT), with an interval of 1 hr. Accordingly, the three years between January 1, 2016, and December 31, 2018, were divided into 26,304 time windows. For example, if a windshear event occurred at 0830 HKT on January 1, 2016, there was at least one windshear event occurring within the time window (from 0800 to 0900 HKT on January 1, 2016). WSO is a binary variable that can take on two values: 1 for at least one windshear event happening within the time window; and 0 otherwise:

\[
WSO = \begin{cases} 
1, & \text{If number of windshear event within the time window} > 0 \\
0, & \text{If number of windshear event within the time window} = 0 
\end{cases}
\]  

(1)

| Level | Name of level | Range of wind direction (WD) angle (°) |
|-------|--------------|--------------------------------------|
| 1     | North        | WD ≥ 337.5 or < 22.5                 |
| 2     | Northeast    | 22.5 ≤ WD < 67.5                     |
| 3     | East         | 67.5 ≤ WD < 112.5                    |
| 4     | Southeast    | 112.5 ≤ WD < 157.5                   |
| 5     | South        | 157.5 ≤ WD < 202.5                   |
| 6     | Southwest    | 202.5 ≤ WD < 247.5                   |
| 7     | West         | 247.5 ≤ WD < 292.5                   |
| 8     | Northwest    | 292.5 ≤ WD < 337.5                   |
The definition of windshear occurrence here refers to the windshear events reported by pilots, which were used as the “Sky Truth” for the windshear-alerting systems of HKIA (Shun and Chan, 2008).

2.1 Binary logistic regression model

A binary logistic is well suited for use when the dependent variable is of the dichotomous type. The conditional probability that at least one or no windshear occurs within the time window (1 hr) at HKIA is modelled as a logistic distribution:

\[ P(WSO = 1) = \pi_i = \frac{\exp(\alpha + \beta_i x_i)}{1 + \exp(\alpha + \beta_i x_i)} \quad (2a) \]

\[ P(WSO = 0) = 1 - \pi_i = \frac{1}{1 + \exp(\alpha + \beta_i x_i)} \quad (2b) \]

The logit transformation of the multiple logistic regression model is given by Equation 3:

\[ \logit[\pi_i] = \ln \left( \frac{\pi_i}{1 - \pi_i} \right) = \alpha + \beta_i x_i \quad (3) \]

where \( \pi_i \) is the conditional probability of the windshear occurrence; \( x_i \) is explanatory variables, which can be categorical or continuous (explanatory variables are discussed below); \( \alpha \) is the intercept; and \( \beta_i \) is the model co-efficient directly determining the odds ratio involved in windshear occurrence. These unknown co-efficients in Equation 3 are typically estimated using maximum likelihood methods.

The influence of one explanatory variable on the risk of windshear occurrence could be revealed by the odds ratio determined by Equation 4:

\[ OR = \exp(\beta_i) \quad (4) \]
The odds of windshear occurrence are ratios of probabilities of the windshear occurring against the probabilities of the windshear not occurring. The odds ratio indicates the relative odds by which the odds of the outcome increases (OR > 1) or decreases (OR < 1) when the corresponding explanatory variable is increased by one unit (Washington et al., 2010). As for the categorical explanatory variables in the study, the odds ratios represented the windshear occurrence risk comparison among different levels (Yan et al., 2005).

2.2 Rare-event logistic regression

The abovementioned ordinary logistic regression model works well for many studies with the dichotomous response variable (Wong et al., 2009; Sze et al., 2014; DeVault et al., 2016; Wang et al., 2019; Zhai et al., 2019). However, when applying the analysis of rare events, the ordinary logistic regression model can lead to an underestimation of the probability (King and Zeng, 2001a). The rare-event logistic regression proposed by King and Zeng (2001a, 2001b) has been used in many studies (Yang et al., 2015; Ren et al., 2016; Theofilatos et al., 2019) to obtain reliable results. Compared with the ordinary logistic regression, the rare-event logistic regression corrects the bias with three correction measures.

First, an endogenous stratified sampling (or choice-based sampling) of the data set should be conducted to form a new data set. This sampling data set consists of all the events and a portion of randomly selected non-event cases. The proportion of the events to non-events is recommended and set to about 1:10 according to other rare-events studies (Guns and Vanacker, 2012; Ren et al., 2016; Zhai et al., 2019). Based on the sampling data set, the maximum likelihood estimation can be used to obtain the estimated intercept \( \hat{\alpha} \) and co-efficients \( \hat{\beta}_i \).

However, the choice-based sampling design could cause significant bias to the estimated intercept parameter \( \hat{\alpha} \). Thus, a prior correction of the intercept is needed to eliminate the sampling bias. The corrected intercept parameter \( \hat{\alpha} \) is calculated based on the estimated intercept \( \hat{\alpha} \) through the maximum likelihood estimation technique, as shown in Equation 5:

\[
\alpha = \hat{\alpha} - \ln \left[ \left( \frac{1 - \tau}{\tau} \right) \left( \frac{\bar{y}}{1 - \bar{y}} \right) \right]
\]

where \( \tau \) represents the fraction of 1’s in the population (the original data set); and \( \bar{y} \) refers to the fraction of 1’s in the sampling data set.

The third correction measure aims to correct the under-estimation of probabilities \( \hat{P}_i \), which is induced by the neglect of the estimation uncertainty on the co-efficients \( \hat{\beta}_i \) when using Equation 2. The final corrected probability \( P_i \) is obtained by adding a correction factor \( C_i \) to \( \hat{P}_i \):

\[
P_i = \hat{P}_i + C_i
\]

The correction factor can be calculated from Equation 7 (King and Zeng, 2001a):

\[
C_i = (0.5 - \hat{P}) \hat{P}(1 - \hat{P}) X_0 \nu(\beta) X_0^T
\]

where \( \beta = (\alpha, \hat{\beta}_i) \); \( X_0 = [1, x_i] \); \( X_0^T \) is the transpose of \( X_0 \); and \( \nu(\beta) \) is the estimated variance–covariance matrix of the estimated co-efficients. The “relogit” function from the R package Zelig (Choirat et al., 2019) was used for the estimation of the rare-event logistic regression.

3 DATA DESCRIPTION

Windshear data for January 2016–December 2018 provided by HKO were used to develop a rare-event logistic regression model of windshear occurrence risk. The data set used in the study consists of two components: (1) windshear reports profile, which contains basic information on windshear events encountered by aircraft at HKIA; and (2) a background weather conditions profile, including hourly wind speed, hourly prevailing wind direction and hourly temperature over HKIA.

The windshear reports profile is the description of windshear events reported by pilots. For every aircraft operating at HKIA, it is encouraged to report by radio communication any encountering of low-level windshear and turbulence to air traffic control. The local airlines are particularly encouraged to do so in order to build up a comprehensive database of windshear and turbulence for the airport. In the report, the pilot is required to provide the date/time of the event, windshear magnitude and sign, height/distance of the windshear encounter (distance from the runway’s end), any encountering of a microburst, the need to conduct a missed approach (for arriving aircraft), and so on. The practice has been adopted since the opening of the airport at Chek Lap Kok in 1998, and accordingly an extensive database of windshear has been built up over the years. After the event, the cause of the windshear/turbulence is studied by trained meteorologists at HKIA. The cause is classified as a sea breeze, terrain induced, thunderstorm, low-level jet, and so on. The
performance of the windshear-alerting service is also verified by comparing it with the pilot reports.

HKIA is one of the busiest airports in the world. There were around 1,000 flights operating at the HKIA every day during 2016–2018. In the previous studies referring to low-level windshear at HKIA, the windshear reports were used as “sky truth” for the performance evaluation of windshear-alerting services and tuning of automatic windshear algorithms (Shun and Chan, 2008; Chan, 2017). Therefore, in the study it is assumed that there were no significant windshear events occurring in the runway corridors if no windshear report happened within the time window.

Based on the windshear reports profile, there were 1,192 windshear events reported by pilots at HKIA in the period 2016–2018. The objective of the study is to assess the risk of windshear occurrence using the rare-event logistic regression model. The response variable \( WOS \) was defined as windshear happening or not happening within the corresponding time window (1 hr) at HKIA. Since 1 hr might witness more than one windshear event, there were finally 693 time windows in which windshear occurred \((WOS = 1)\) on the runway corridors at HKIA during the three years.

The background weather conditions profile provides hourly mean wind speed \((SPD)\), hourly prevailing wind direction \((WD)\) and hourly mean temperature \((TEMP)\) over HKIA from January 2016 to December 2018. The hourly mean is calculated based on the time window, and each time window contains one hourly mean. There are in total 26,304 hr in the three years (January 1, 2016–December 31, 2018), so the number of time windows in the data set is 26,304. Of the 26,304 time windows, 10 time windows contain missing values. Considering the minor proportion (0.04%) of missing values, as well as no windshear having occurred in these 10 hr, these 10 entries were dropped from the data set, leaving 26,294 entries.

The windshear reports profile and the background weather conditions profile were combined into a windshear data set. The hourly prevailing wind direction \((WD)\) was given in degrees, which would cause difficulty in model interpretation since 0 and 360° indicate the same wind direction, namely, north. For the convenience of the analysis, the continuous variable hourly prevailing wind direction \((WD)\) was discretized into eight groups, representing eight wind directions (Table 1). The month was also an important potential explanatory variable because the distribution of windshear occurrence showed significant monthly differences (Figure 2). Therefore, the variable \( MONTH \) was separated into 12 levels, indicating the months January–December. Table 2 presents the descriptive statistics of the response variable and potential explanatory variables in which the mean, standard deviation, minimum and maximum of these variables are given.

4 | MODEL ESTIMATION RESULTS

To validate the logistic regression model, the windshear data set was subdivided into training and testing data sets with a ratio of 7:3 (Liu and Cocea, 2017). Based on the training data set, the backward selection process of the logistic regression process was used to develop the model. A \( p \)-value of 0.05 was chosen as the significance level. To begin with, all four potential explanatory variables, namely, \( SPD \), \( TEMP \), \( WD \) and \( MONTH \), were tested in the model without interaction terms. For categorical variables holding several levels, dummy variables were created for each, and one of these dummy variables was used as a reference in the estimation to avoid
multicollinearity. For \( WD \) and \( MONTH \), dummy variables \( NW \) and \( OCT \) were selected as the reference category, respectively. Next, according to the results obtained from fitting all potential explanatory variables, those variables with \( p > 0.05 \) were removed from the model. In each estimation process, only that with the highest \( p \)-value was dropped from the model, and this process was repeated until all the variables within the model had \( p < 0.05 \).

To deal with the binary class imbalance of the data set, three data-level approaches (under-sampling, over-sampling and hybrid method) were used for the training data set (Vluymans, 2019). These three approaches aim to reduce the imbalance ratio by resampling the training data set. In the study, under-sampling was conducted by randomly discarding some records from the majority class and leaving the minority class untouched. In fact, the first step in the rare-event logistic regression method is a type of under-sampling that forms a new data set containing all the events (the minority class) and a portion of randomly selected non-event (the majority class) cases. In the previous studies using the rare-event logistic regression model, the ratio of events to non-events in the new data set after under-sampling is recommended to be 1:10 (Guns and Vanacker, 2012; Ren et al., 2016; Theofilatos et al., 2019). In other words, the imbalance ratio was reduced to 10 from a higher value in the original data set by choice-based sampling (under-sampling). Over-sampling takes the opposite method of under-sampling. Instead of decreasing the majority class, over-sampling increases the size of the minority class by randomly selecting the minority records for duplication (Batista et al., 2004). The third resampling method is hybrid sampling, which combines under- and over-sampling to form a new subset of training data. For example, there are 693 events (\( WOS = 1 \)) and 25,601 non-events (\( WOS = 0 \)) in the original windshear data set. With a ratio of 7:3, the training data set is selected randomly by stratification. Hence, the training data set contains 485 events and 17,921 non-events. If the target ratio of events to non-events obtained from hybrid sampling is 5:10, over-sampling would be used to increase number

| Resampling method | N1:N0 | Regression model | Statistical summary of the AUC |
|-------------------|-------|------------------|-------------------------------|
|                   |       |                  | Minimum | 1st quarter | Median | Mean | 3rd quarter | Maximum |
| Under-sampling    |       | Ordinary         | 0.6045  | 0.6363      | 0.6476 | 0.6481 | 0.6605      | 0.6997  |
|                   |       | Rare event       | 0.5956  | 0.6360      | 0.6481 | 0.6480 | 0.6595      | 0.7022  |
|                   |       | Ordinary         | 0.5992  | 0.6363      | 0.6486 | 0.6486 | 0.6599      | 0.7035  |
|                   |       | Rare event       | 0.6007  | 0.6361      | 0.6486 | 0.6483 | 0.6593      | 0.7022  |
|                   |       | Ordinary         | 0.5952  | 0.6334      | 0.6464 | 0.6459 | 0.6573      | 0.7013  |
|                   |       | Rare event       | 0.5995  | 0.6358      | 0.6482 | 0.6479 | 0.6590      | 0.7014  |
| Over-sampling     |       | Ordinary         | 0.6019  | 0.6387      | 0.6510 | 0.6512 | 0.6624      | 0.7056  |
|                   |       | Rare event       | 0.6018  | 0.6388      | 0.6508 | 0.6511 | 0.6621      | 0.7049  |
| Hybrid sampling   |       | Ordinary         | 0.5979  | 0.6374      | 0.6498 | 0.6499 | 0.6614      | 0.7062  |
|                   |       | Rare event       | 0.6003  | 0.6373      | 0.6496 | 0.6498 | 0.6605      | 0.7009  |
|                   |       | Ordinary         | 0.5997  | 0.6381      | 0.6502 | 0.6505 | 0.6617      | 0.7059  |
|                   |       | Rare event       | 0.6000  | 0.6378      | 0.6501 | 0.6501 | 0.6612      | 0.7044  |
|                   |       | Ordinary         | 0.6005  | 0.6384      | 0.6506 | 0.6510 | 0.6622      | 0.7066  |
|                   |       | Rare event       | 0.6006  | 0.6382      | 0.6506 | 0.6506 | 0.6618      | 0.7042  |
|                   |       | Ordinary         | 0.6014  | 0.6381      | 0.6507 | 0.6509 | 0.6621      | 0.7045  |
|                   |       | Rare event       | 0.6010  | 0.6380      | 0.6506 | 0.6506 | 0.6617      | 0.7050  |
|                   |       | Ordinary         | 0.6015  | 0.6383      | 0.6508 | 0.6511 | 0.6623      | 0.7050  |
|                   |       | Rare event       | 0.6020  | 0.6385      | 0.6506 | 0.6509 | 0.6620      | 0.7049  |

Note: \( n1 \): number of events in the training data set obtained by dividing the original windshear data set by the ratio 7:3; \( n0 \): number of non-events in the training data set obtained by dividing the original windshear data set by the ratio 7:3; \( N1 \): number of events in the new training data set after resampling the training data set; and \( N0 \): number of non-events in the new training data set after resampling the training data set.
the minority class to 2,425 and under-sampling would be used to reduce the number of non-events to 4,850.

Besides the three resampling methods with a different target ratio of events to non-events, both the ordinary logistic and the rare-event logistic regression models were estimated to present the possible results based on the windshear data set. For each ratio of events to non-events, 1,000 random resampling and model estimations were conducted. The corresponding testing data set was used to test the performance of the estimated model. Receiver operating characteristic (ROC) curves were plotted and the areas under the curve (AUCs) were calculated. The statistical summary of the AUC obtained from 1,000 runs of the model for each ratio of events to non-events is presented in Table 3. The histogram of AUC distribution for the logistic regression model with the 15n1:20n1 ratio is shown in Figure 3. For other cases with a different ratio of events to non-events, the shape of the histogram for the AUC is similar to that shown in Figure 3.

According to the results in Table 3, with the same training data set, the AUC difference between the ordinary logistic regression model and the rare-event logistic regression model is very small. In other words, based on the wind-shear data set with the four variables in Table 1, the rare-event logistic regression model did not outperform the ordinary logistic regression model. The difference in the AUC distribution in Table 3 is mainly induced by the resampling methods. All three AUCs are < 0.71, which indicates that these rare-event logistic regression models do not provide good prediction performance when regarding windshear

---

**TABLE 4** Summary of the rare-event logistic regression model (events: non-events = 15n1:20n1)

| Variables | Estimated co-efficient | Odds ratio | Estimated standard error | z-value | p-value |
|-----------|------------------------|------------|--------------------------|---------|---------|
| Intercept | −0.7339                | 0.0349     | −21.026                  | < 0.0001|
| SPD\(^a\) | 0.0836                 | 1.0872     | 0.0058                   | 14.305  | < 0.0001|
| WD        |                        |            |                          |         |         |
| EAS\(^b\) | 0.7727                 | 2.1655     | 0.0339                   | 22.817  | < 0.0001|
| SE\(^b\)  | 0.5753                 | 1.7776     | 0.0514                   | 11.189  | < 0.0001|
| SOT\(^b\) | 0.7120                 | 2.0380     | 0.0504                   | 14.116  | < 0.0001|
| SW\(^b\)  | 0.4527                 | 1.5726     | 0.0518                   | 8.739   | < 0.0001|
| MONTH     |                        |            |                          |         |         |
| MAR\(^b\) | 0.5309                 | 1.7005     | 0.0446                   | 11.900  | < 0.0001|
| APR\(^b\) | 0.4549                 | 1.5760     | 0.0458                   | 9.934   | < 0.0001|
| JUL\(^b\) | 0.2416                 | 1.2733     | 0.0489                   | 4.945   | < 0.0001|
| AUG\(^b\) | 0.3984                 | 1.4895     | 0.0471                   | 8.468   | < 0.0001|
| NOV\(^b\) | −0.7731                | 0.4616     | 0.0679                   | −11.379 | < 0.0001|

\(^a\)Statistically significant at the 0.05 level.
\(^b\)Statistically significant at the 0.01 level.
occurrence at HKIA. This is reasonable given the transient and sporadic nature of windshear.

Among these estimated models, the results of the ratio (15n1:20n1) is presented in Tables 4 and 5. According to the \( p \)-value, the results show that three variables, namely, \( SPD \), \( WD \) and \( MONTH \), significantly affected the risk of windshear occurrence at HKIA at the \( p = 0.05 \) significance level. The variable \( TEMP \) was found to be statistically insignificant. Besides, the variance inflation factor (VIF) was used to test the multicollinearity of the model. Each explanatory variable showed a VIF < 2, so there is no multicollinearity issue in the models.

### 5 | DISCUSSION AND LIMITATIONS

Odds ratios were used to interpret the model results. An odds ratio > 1 increases the likelihood of a windshear occurrence within the time window at HKIA. Since the difference in the estimated co-efficients between the ordinary logistic (Table 5) and the rare-event logistic (Table 4) regression models is small, it could be neglected. The following subsections document the interpretation of the rare-event logistic regression model developed in the study. In addition, the height of windshear occurrence is also discussed.

#### 5.1 | Wind speed

As a continuous variable, hourly mean wind speed (\( SPD \)) was found to affect the risk of windshear occurrence significantly. The odds ratio of \( SPD \) is > 1, which shows that the risk of windshear occurrence would increase as hourly mean wind speed increases. However, as for the occurrence of windshear, the variation in wind speed is more important than mean wind speed. Since the duration of a windshear event encountered by an aircraft is generally between a few seconds and several minutes, hourly mean wind speed can hardly provide an accurate indication of windshear occurrence, which is revealed by the AUCs in Table 3. Therefore, more refined data about wind conditions such as a 1 min mean in turbulence intensity are required to improve the performance of the regression model.

#### 5.2 | Wind direction

Based on the results presented in Table 4, hourly prevailing wind direction (\( WD \)) showed a significant association with the risk of windshear occurrence at HKIA. Four levels of \( WD \) (\( EAS \), \( SE \), \( SOU \) and \( SW \)) were found to bring about a higher risk of windshear occurrence on runway corridors at HKIA within the time window compared with the reference wind directions (\( NOR \), \( NE \), \( WES \) and \( NW \)). This indicates that the windshear event could be more likely to happen under the easterly, southeasterly, southerly and southwesterly directions, a finding consistent with previous studies (Shun and Chan, 2008; Chan, 2014a). HKIA is located on a reclaimed island surrounded by sea on three sides, and the mountainous Lantau Islands lies on the other side (Figure 1). Therefore, the background winds from the east, south and southwest

| Variables | Estimated co-efficient | Odds ratio | Estimated standard error | \( z \)-value | \( p \)-value |
|-----------|-----------------------|------------|--------------------------|--------------|-------------|
| Intercept | −1.2755               | 0.0350     | −36.443                  | <0.0001      |
| \( SPD^b \) | 0.0838               | 1.0874     | 0.0059                   | 14.314       | <0.0001     |
| \( WD \) |                       |            |                          |              |             |
| \( EAS^b \) | 0.7716               | 2.1632     | 0.0338                   | 6.407        | <0.0001     |
| \( SE^b \) | 0.6044               | 1.8302     | 0.0514                   | 2.45         | <0.0001     |
| \( SOU^b \) | 0.7291               | 2.0732     | 0.0514                   | 4.216        | <0.0001     |
| \( SW^b \) | 0.5147               | 1.6731     | 0.0511                   | 2.345        | <0.0001     |
| \( MONTH \) |                    |            |                          |              |             |
| \( MAR^b \) | 0.5185               | 1.6795     | 0.0448                   | 11.584       | <0.0001     |
| \( APR^b \) | 0.5326               | 1.7034     | 0.0455                   | 11.711       | <0.0001     |
| \( JUL^a \) | 0.1070               | 1.1129     | 0.0491                   | 2.181        | 0.0292      |
| \( AUG^b \) | 0.2953               | 1.3435     | 0.0480                   | 6.152        | <0.0001     |
| \( NOV^a \) | −0.7589              | 0.4682     | 0.0664                   | −11.421      | <0.0001     |

\( a \)Statistically significant at the 0.05 level.

\( b \)Statistically significant at the 0.01 level.
could be disturbed by the complex terrain, which induced airflow disturbances over HKIA.

Besides, in either Tables 3 or 4, the odds ratio of SW was the smallest compared with the other three levels of WD, which indicated that the increasing effect of the SW on the risk of windshear occurrence is relatively lower than that of the EAS, SE and SOU. Since Lantau Island runs from northeast to southwest, the southwest flow was less disrupted compared with the easterly, southeasterly and southerly flows. This could be an explanation for the difference mentioned above.

It can be inferred that the higher risk of windshear occurrence resulted from easterly, southeasterly, southerly and southwesterly is probably related to the complex terrain of Lantau Island. Some previous studies using numerical simulation could give support to the inference (Lei et al., 2013; Li and Chan, 2016). Using a computational fluid dynamics (CFD) model, two cases of vortex/wave shedding at HKIA have been successfully simulated, with the background wind direction of east on the ground and south or southwest aloft (Li and Chan, 2016). Besides, buildings near and at HKIA could also result in the increase in windshear occurrence risk under the southerly or southeasterly, which was reflected by a wind tunnel study recently (Chen et al., 2019, 2020).

5.3 Month

Results in Table 4 show that the influence of month on the risk of windshear occurrence at HKIA was statistically significant. Four months (March, April, July and August) were associated with a higher risk of windshear occurrence compared with the other months not shown in Table 4. As shown in Figure 2, there were more hours with windshear occurrence in these four months (March, April, July and August) compared with the other months. Among these four months, July witnessed the least number of hours with windshear occurrence. This was reflected in odds ratios of MONTH.

Months with a higher risk of windshear occurrence and their odds ratios presented in Table 4 were fairly reasonable since the terrain-induced windshear events, which accounted for nearly 70% of reported windshear events, mainly occurred in spring and summer (Chan, 2014a). In springtime, the prevailing wind in Hong Kong is easterly, which flows over Lantau Island and then reaches HKIA. In summertime, the cross-mountain airflow occurs over HKIA in the southwest monsoon or during passages of tropical cyclones. These terrain-disrupted airflows would result in several windshear events, which adversely impact the flight safety and operational efficiency of HKIA.

Besides, results in Table 4 show that November had lower risks of windshear occurrence at HKIA as compared with the reference months. The odds ratio indicated that the risk of windshear occurrence in November could be much less than that in the reference months. At HKIA, terrain-induced windshear and sea breeze are two major types of windshear events, which account for about 70% and 20% of the total events reported by pilots, respectively (Shun and Chan, 2008). Since the complex terrain is located to the south of HKIA, terrain-induced windshear often occurs near the flight corridors when winds blow across the hills from the east, southeast, south and southwest (HKO, 2010). Besides, a sea breeze usually develops when prevailing westerly winds or prevailing west to southwesterly winds blow in the background at HKIA (HKO, 2010). However, in November, the prevailing wind at Hong Kong mainly comes from the north or northeast, so that both terrain-induced windshear and sea breeze seldom occur at HKIA during this period. Therefore, November suffered a lower windshear risk significantly at HKIA. As shown in Figure 2, the number of hours with windshear occurrence in November was the smallest for the windshear data.
set used in the study. In this context, November could be regarded as a reference month when analysing the risk of windshear occurrence of different months in future.

5.4 Height of windshear occurrence

When there is no windshear occurrence in a time window (1 hr), the height of the windshear report would be marked as “N/A”. Since there was a lack of a reference group, the height of the windshear report was not involved in the logistic regression model in the study. However, the height of the windshear encounter was also a critical variable for the risk of windshear occurrence, which might be associated with the characteristics of terrain-disrupted airflows. Based on the windshear reports from 2016 to 2018 at HKIA, the distribution of the height of windshear occurrence is presented in Figure 4. Most of the windshear reports occurred < 600 m, which corresponded to the definition of low-level windshear. Especially, nearly 55% of the windshear events had a height < 200 m, indicating the most dangerous areas with a high risk of encountering windshear. The wind tunnel study by Chen et al. (2019) also showed that the complex terrain and buildings in the vicinity of the airport would bring about higher turbulence intensities along the glide path below around 200 m. Therefore, the low-level wind effects caused by buildings near and at the airfield should be taken into consideration when analysing the windshear occurrence with a height < 200 m. The absence of height information about windshear occurrence may be one significant reason for the poor performance of the estimated model in the study.

6 CONCLUSIONS

In order to assess the risk of windshear occurrence at HKIA, the rare-event logistic regression method was applied to a windshear data set provided from Hong Kong Observatory (HKO). Considering the high imbalance of the outcome variable, under-sampling, over-sampling and a hybrid method were applied to reduce the imbalance ratio in order to improve the performance of the estimated model. However, both the ordinary logistic and the rare-event logistic regression models resulted in small areas under the curve (AUCs). Based on the estimated rare-event logistic regression model, hourly mean wind speed, hourly prevailing wind direction and month were found to be statistically significant. The effects of these explanatory variables on the risk of windshear occurrence were revealed by the odds ratio compared with the reference level. Conclusions are summarized as follows:

- It seems that windshear occurrence risk increases as hourly mean wind speed increases.
- The risk of windshear occurrence at Hong Kong International Airport (HKIA) was greater under the hourly prevailing wind direction of east, southeast, south and southwest. The complex terrain and buildings in the vicinity of the airport were contributing factors to the increase of windshear occurrence risk.
- March, April, July and August were found to be involved in a higher risk of windshear occurrence compared with other months, which was consistent with the fact that most of the terrain-induced windshear events occurred in spring and summer. Besides, November was found to be involved in the lesser risk of windshear occurrence at HKIA; November could be used as a reference month to assess the windshear risk of other months.
- Nearly 55% of the windshear events reported by pilots occurred at a height < 200 m, which was the high-risk area of windshear occurrence with low-level wind effects of airport buildings.

It should be acknowledged that the results from rare-event logistic regression models developed in the study did reflect some features associated with the risk of windshear occurrence at HKIA. Although some of these findings could be observed from descriptive statistical analysis, the application of the rare-event logistic regression model provided some quantitative results about these explanatory variables. However, considering the transient and sporadic nature of windshear experienced by an aircraft, it is almost impossible to predict precisely windshear occurrence based on hourly weather data and windshear reports. This was confirmed by the low AUCs of these rare-event logistic regression models developed in the study. Therefore, data with smaller spatial and temporal scale are need in order to improve the prediction ability of regression models.

ACKNOWLEDGEMENT

The study was supported by Project U1733113, sponsored by the National Natural Science Foundation of China.

ORCID

Pak-wai Chan https://orcid.org/0000-0003-2289-0609
Xiaoxiang Ma https://orcid.org/0000-0001-9080-0316

REFERENCES

Bai, S., Lü, G., Wang, J., Zhou, P. and Ding, L. (2011) GIS-based rare events logistic regression for landslide-susceptibility mapping of Lianyungang, China. Environmental Earth Sciences, 62 (1), 139–149.
Batista, G.E.A.P.A., Prati, R.C. and Monard, M.C. (2004) A study of the behavior of several methods for balancing machine learning training data. *Acm Sigkdd Explorations Newsletter*, 6(1), 20–29.

Chan, P.W. (2012) An event of tail strike of an aircraft due to terrain-induced wind shear at the Hong Kong international airport. *Meteorological Applications*, 19(3), 325–333.

Chan, P.W. (2014a) A tail strike event of an aircraft due to terrain-induced wind shear at the Hong Kong international airport. *Meteorological Applications*, 21(3), 504–511.

Chan, P.W. (2014b) Observation and numerical simulation of vortex/wave shedding for terrain-disrupted airflow at Hong Kong international airport during typhoon Nesat in 2011. *Meteorological Applications*, 21(3), 512–520.

Chan, P.W. (2017) Severe wind shear at Hong Kong International Airport: climatology and case studies. *Meteorological Applications*, 24(3), 397–403.

Chan, P.W., Hon, K.K. and Li, Q.S. (2019) A comprehensive study of terrain-disrupted airflow at Hong Kong international airport—observations and numerical simulations. *Weather*, 75(7), 199–206.

Chen, F., Peng, H., Chan, P. and Zeng, X. (2019) Low-level wind effects on the glide paths of the north runway of HKIA: a wind tunnel study. *Building and Environment*, 164, 106337.

Chen, F., Peng, H., Chan, P. and Zeng, X. (2020) Wind tunnel testing of the effect of terrain on the wind characteristics of airport glide paths. *Journal of Wind Engineering and Industrial Aerodynamics*, 203, 104253.

Choirat C, Honaker J, Imai K, et al. (2017) Zelig: everyone's statistical software. Available at https://cran.r-project.org/web/packages/Zelig/index.html [Accessed 6th December 2019], Version, 2019.

DeVault, T.L., Blackwell, B.F., Seamans, T.W. and Belant, J.L. (2016) Identification of off airport interspecific avian hazards to aircraft. *The Journal of Wildlife Management*, 80(4), 746–752.

Guns, M. and Vanacker, V. (2012) Logistic regression applied to natural hazards: rare event logistic regression with replications. *Natural Hazards and Earth System Sciences*, 12(6), 1937–1947.

HKO. (2010) *Windshear and Turbulence in Hong Kong – information for pilots*, 3rd edition. Hong Kong: Hong Kong Observatory.

Hon, K.K. and Chan, P.W. (2014) Application of LIDAR-derived eddy dissipation rate profiles in low-level wind shear and turbulence alerts at Hong Kong international airport. *Meteorological Applications*, 21(1), 74–85.

ICAO. (2016) *Meteorological Service for International Air Navigation*, Annex 3 to the Convention on International Civil Aviation, 19th edition. Montréal, Quebec, Canada: International Civil Aviation Organization.

King, G. and Zeng, L. (2001a) Logistic regression in rare events data. *Political Analysis*, 9(2), 137–163.

King, G. and Zeng, L. (2001b) Explaining rare events in international relations. *International Organization*, 55(3), 693–715.

Lee, Y.F. and Chan, P.W. (2014) LIDAR-based F-factor for wind shear alerting: different smoothing algorithms and application to departing flights. *Meteorological Applications*, 21(1), 86–93.

Lei, L., Chan, P.W., Li-Jie, Z. and Hui, M. (2013) Numerical simulation of terrain-induced vortex/wave shedding at the Hong Kong international airport. *Meteorologische Zeitschrift*, 22(3), 317–327.

Lei, L.I., Chan, P.W. and Tse, S.M. (2017) Observation and simulation of mountain wave trains in a tropical cyclone situation. *Mausam*, 68(2), 253–260.

Leung, D.Y.C., Lo, W.Y., Chow, W.Y. and Chan, P.W. (2012) Effect of terrain and building structures on the airflow in an airport. *Journal of Zhejiang University Science A*, 13(6), 461–468.

Li, L. and Chan, P.W. (2012) Numerical simulation study of the effect of buildings and complex terrain on the low-level winds at an airport in typhoon situation. *Meteorologische Zeitschrift*, 21(2), 183–192.

Li, L. and Chan, P.W. (2016) LIDAR observation and numerical simulation of vortex/wave shedding at the eastern runway corridor of the Hong Kong international airport. *Meteorological Applications*, 23(3), 379–388.

Li, M., Xu, J., Xiong, X., et al. (2018) A novel ramp method based on improved smoothing algorithm and second recognition for Wind-shear detection using LIDAR. *Current Optics and Photonics*, 2(1), 7–14.

Liu, H. and Cooce, M. (2017) Semi-random partitioning of data into training and test sets in granular computing context. *Granular Computing*, 2, 357–386. https://doi.org/10.1007/s41066-017-0049-2.

Ren, Y., Wang, Y., Wu, X., Yu, G. and Ding, C. (2016) Influential factors of red-light running at signalized intersection and prediction using a rare events logistic regression model. *Accident Analysis & Prevention*, 95, 266–273.

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

Sze, N.N., Wong, S.C. and Lee, C.Y. (2014) The likelihood of achieving quantified road safety targets: a binary logistic regression model for possible factors. *Accident Analysis & Prevention*, 73, 242–251.

Theofilatos, A., Yannis, G., Kopolias, P. and Papadimitriou, F. (2019) Impact of real-time traffic characteristics on crash occurrence: preliminary results of the case of rare events. *Accident Analysis & Prevention*, 130, 151–159.

Van den Eeckhaut, M., Vanwalraven, M., Poesen, J., et al. (2006) Prediction of landslide susceptibility using rare events logistic regression: a case-study in the Flemish Ardennes (Belgium). *Geomorphology*, 76(3–4), 392–410.

Vluymans, S. (2019) Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods. Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-04663-7.

Wang, L., Abdel-Aty, M., Lee, J. and Shi, Q. (2019) Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors. *Accident Analysis & Prevention*, 122, 378–384.

Washington, S.P., Karlaftis, M.G. and Mannering, F. (2010) *Statistical and Econometric Methods for Transportation Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Wong, D.K.Y., Pitfield, D.E., Caves, R.E. and Appleyard, A.J. (2009) The development of a more risk-sensitive and flexible airport...
safety area strategy: part I. the development of an improved accident frequency model. *Safety Science*, 47(7), 903–912.

Wu, T.C. and Hon, K.K. (2018) Application of spectral decomposition of LIDAR-based headwind profiles in windshear detection at the Hong Kong international airport. *Meteorologische Zeitschrift*, 27, 33–42.

Yan, X., Radwan, E. and Abdel-Aty, M. (2005) Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model. *Accident Analysis & Prevention*, 37(6), 983–995.

Yang H, Ozbay K, Xie K, et al. *Modeling crash risk of highway work zones with relatively short durations*. Washington D.C.: Transportation Research Board 94th Annual Meeting; 2015.

Zhai, X., Huang, H., Sze, N.N., Song, Z. and Hon, K.K. (2019) Diagnostic analysis of the effects of weather condition on pedestrian crash severity. *Accident Analysis & Prevention*, 122, 318–324.

**How to cite this article:** Chen F, Peng H, Chan P, Ma X, Zeng X. Assessing the risk of windshear occurrence at HKIA using rare-event logistic regression. *Meteorol Appl*. 2020;27:e1962. [https://doi.org/10.1002/met.1962](https://doi.org/10.1002/met.1962)