Journal of Big Data

METHODOLOGY Open Access

Machine learning-based turbulence-risk prediction method for the safe operation of aircrafts

Shinya Mizuno[1*], Haruka Ohba[1] and Koji Ito[2]

*Correspondence: s.mzn.eng@gmail.com
[1] Shizuoka Institute of Science and Technology, 2200-2, Toyosawa, Fukuroi, Shizuoka 437-8555, Japan
Full list of author information is available at the end of the article

Abstract

This study has proposed a method for detecting turbulence, a primary factor that influences safe aircraft operation. The number of observed turbulence events is limited, thereby indicating the requirement of an appropriate flow for detecting turbulence events from a small number of samples. In addition, the opinions and experiences of pilots must be reflected at the initial stage to address the high risk of turbulence occurrence, which can result in airline operations being cancelled. Thus, this study proposed a method for predicting turbulence occurrence based on the turbulence occurrence date information provided by airlines as well as meteorological data sets obtained from open data available in Japan as teacher data. However, because commonly used machine learning methods are unable to detect the turbulence occurrence date, the proposed method employed principal component analysis coupled with the K-Means method to generate risk clusters with a high likelihood of turbulence occurrence and consequently perform statistical checks. Subsequently, the risk clusters were utilized as supervisory data for turbulence occurrence, while the support vector machine was used for predicting turbulence occurrence. Furthermore, the results obtained with the proposed method were statistically checked as well as practically verified by a pilot to confirm the appropriateness of the turbulence occurrence date predicted.

Keywords: PCA, k-means, SVC, Risk cluster, Open data, Meteorological data, Mountain waves

Introduction

One the most important requirements for airlines has been providing a comfortable space to customers, with avoidance and mitigation of aircraft shaking being a crucial factor. Turbulence is among the common causes of aviation accidents [1, 2]. In addition, the potential increase in aircraft turbulence owing to the effects of global warming is a prevalent concern [3].

Upon receiving a report by a pilot related encountering one or more instances of severe turbulence during a flight, the corresponding aircraft must undergo maintenance work to confirm its airworthiness. Therefore, turbulence remains a major issue for airlines. In addition, if the maximum acceleration recorded exceeds the operational acceleration limit of the aircraft, the scope of maintenance work increases considerably, thereby significantly impacting aircraft operation schedules. Therefore, airlines must strive to avoid severe turbulence to the best extent possible. However, if reports regarding turbulence rely primarily on the opinions of pilots, which tend to vary, variations in reports provided by them are inevitable.

Consequently, this study proposed a method for predicting turbulence occurrence, with an aim to contribute to the safe and comfortable operation of aircrafts. Figure 1 outlines this method, which involves the accumulation and aggregation of open data and quick access recorder (QAR) data [4, 5]. In addition, the prediction of turbulence using machine learning methods is outlined as well, the results of which are fed back to airlines and pilots. Flights to and from Matsumoto Airport in Japan, on E-170 aircrafts operated by Fuji Dream Airlines (FDA), have been observed to frequently experience turbulence during the winter season. In this study, the Matsumoto Airport was considered as the model airport representing mountainous areas subject to turbulence. The proposed technique can also be adapted to other airports.

For conducting the study, meteorological data from Japan and turbulence information provided by FDA were used. Because turbulence is a relatively rare event, first, the risk cluster was estimated. To this end, a principal component analysis (PCA) of the meteorological data was conducted to obtain a projection matrix $\boldsymbol{W}$ such that the number of dimensions of the data to be analyzed was reduced. Subsequently, using the turbulence-occurrence indicator and meteorological data transformed by $\boldsymbol{W}$, the k-means method was employed to calculate the risk cluster, which is required for predicting the days with turbulence risk for meteorological data from the year 2019 through support vector classification (SVC). The results based on this meteorological data revealed that the prediction method accurately identified the days with a risk of turbulence.

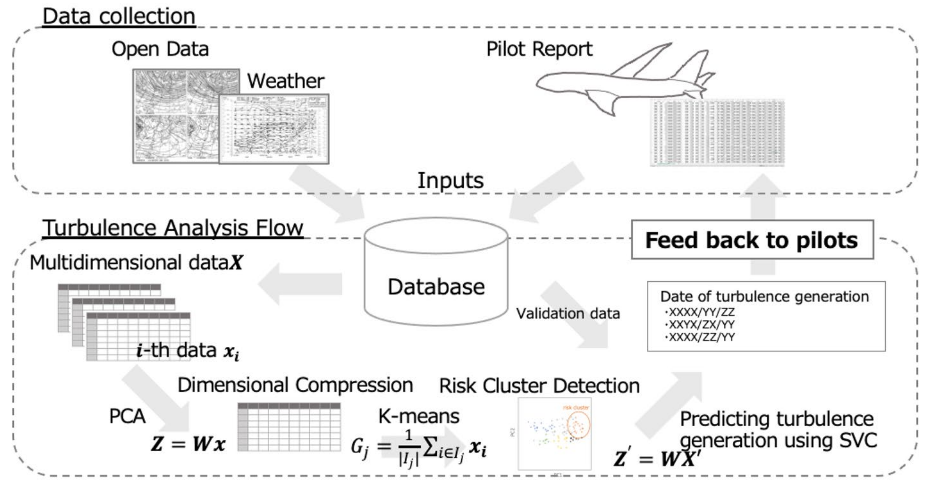
Data collection
Open Data
Weather
Pilot Report
Inputs
Turbulence Analysis Flow
Feed back to pilots
Multidimensional data$\boldsymbol{X}$
Database
Validation data
Date of turbulence generation
·XXXX/YY/ZZ
·XXYX/ZX/YY
·XXXX/ZZ/YY
i-th data $\boldsymbol{x}_i$
Dimensional Compression
Risk Cluster Detection
PCA
$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{x}$
K-means
$G_j = \frac{1}{|I_j|}\sum_{i \in I_j} \boldsymbol{x}_i$
Predicting turbulence generation using SVC
$\boldsymbol{Z}' = \boldsymbol{W}\boldsymbol{X}'$

Fig. 1 Framework of the proposed method

Related work

Most existing research concerning turbulence prediction has been performed from a meteorological perspective [6, 7], such as studies conducted to examine past turbulence incidents [8]. In an event that occurred in central Colorado on January 11, 1972, optimal conditions for strong mountain wave generation were detectable from sounding data 12–24 h in advance and approximately 1000 km upstream [9]. Further, in the case of a fatal accident involving a light aircraft near Clonvina Inn, Victoria, Australia, on July 31, 2007, the observed environment was analyzed and consequently through a three-dimensional simulation the region where turbulence intensified was identified [10]. When a Boeing 777 encountered severe clear-air turbulence (CAT) over western Greenland at an altitude of 10 km on May 25, 2010, through digital flight data recorder (DFDR) analysis and high-resolution numerical simulations the operation of a high-resolution non-hydrostatic simulation model was confirmed to predict mountain-wave turbulence (MWT) [11]. Thus, understanding past examples are crucial to identifying and predicting the conditions for turbulence.

Currently, analysis using the QAR (Quick Access Recorder) data on board the aircraft is also under consideration for predicting turbulence. Further, new methods for estimating eddy dissipation rate (EDR), considered as a measurement of turbulence, through QAR [12], comparison of calculation algorithms [13], and development of QAR data analysis software to calculate meteorological quantities such as three wind components, wind shear risk coefficient, and turbulence intensity parameters [14] have been proposed as well.

In the current aviation industry, a method for turbulence detection involves the use of Doppler lidar [15–17]. A laser beam (using a wavelength band that is safe for the pilot's eyes) can be fired into the atmosphere to observe winds in the sky. Although CAT cannot be detected by conventional aviation weather radars, airborne predictive windshear (PWS) radars enhanced with algorithms designed for turbulence detection and long-range airborne Doppler lidars have been developed and operated [18–20]. Consequently, turbulence detection using these systems has resulted in a reduction in the number of turbulence encounters by alerting pilots to the possibility of encounters.

However, in these studies, data were acquired in real time from many sensors and analyzed using a time-series approach [21]. Although turbulence forecasting with pinpoint accuracy is desirable, preparing a suitable environment for the sensors results in significant cost, and thus, it is infeasible for airlines.

In recent years, owing to the accumulation of aviation data and improvements in computation rapidity, the concept of turbulence prediction via machine learning has been introduced [22, 23]. However, studies concerning this subject are limited. Furthermore, determining an optimal machine learning approach for turbulence prediction is challenging. Moreover, there exists a need to utilize open data (such as meteorological data) to improve analysis accuracy as it can aid in the development of turbulence predictions that can be logically deduced from the data provided by the airlines. For example, in a detailed study of the causes of 700 fatal aviation accidents involving commercial airliners that occurred worldwide between 1990 and 2006, it was found that the composition of accident causes varied greatly depending region of the world, type of operation, and category of aircraft [24]. Further, a study proposed a turbulence prediction algorithm that was based on the examination of turbulent weather phenomena and aircraft operations using a stepwise multiple regression analysis model [25].

Thus, the above discussion reiterates the importance of developing of a system that can predict turbulence, among the most common aviation accidents, independent of the equipment and environment used to acquire the data.

This study attempted to approach machine learning from a non-meteorological perspective. PCA was employed to generate risk clusters for the data and determine the prediction accuracy. Several studies have followed a procedure similar to that of this study [26]. For instance, when attempting to identify the relevant genes for gene expression classification, the data was passed through PCA and independent component analysis methods, and based on the variants of the class obtained, the selected elements were individually transformed to lower dimensions. Consequently, the classification performance of the experiment was evaluated using a support vector machine kernel classifier [27]. Further, in Classifying Colon Cancer Microarray Data, PCA and Partial Least Square (PLS) have also been used to extract more features [13].

However, this has never been done in case of turbulence analysis. Therefore, this study examined the possibility of it being applied as a new method in the field for turbulence prediction.

Basic analysis of turbulence at matsumoto airport

In this section a basic analysis of the data collected at Matsumoto Airport is described.

Examples of the effects of turbulence on flights from Matsumoto Airport

Considering the topographical characteristics shown in Fig. 2, it can be inferred that flights operating from Matsumoto Airport are susceptible to mountain waves [28] from the Northern Alps, particularly on the route toward New Chitose Airport.

Table 1 summarizes the turbulence, presumably caused by mountain waves, reported by flights departing from the Matsumoto Airport. The authors were present on the flight that departed on December 27th, 2017, to gain a real-world understanding of the level of turbulence faced during a flight. Table 2 shows the meteorological conditions during

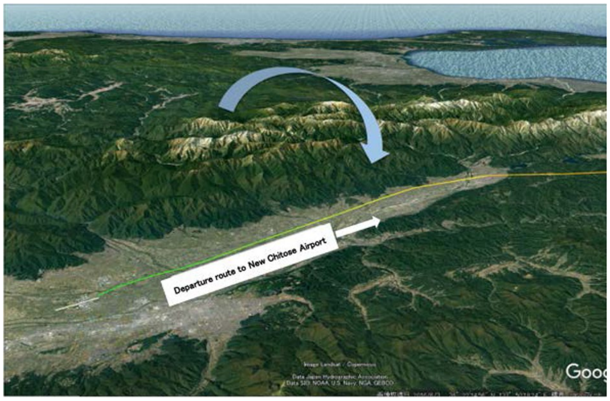
Departure route to New Chitose Airport
Image Landsat / Copernicus
Data Japan Hydrographic Association
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Goog

Fig. 2 Impact of mountain waves on flights departing from Matsumoto Airport

Table 1 Examples of the impact of turbulence on operations

Date	Flight route	Turbulence reported
12/12/2017	MMJ[a] to FUK[b]	Encountered moderate-plus turbulence while climbing
12/27/2017	MMJ[a] to FUK[b]	Encountered moderate-plus turbulence while climbing
01/23/2018	MMJ to CTS[c]	Encountered severe turbulence while climbing Destination changed to NKM[d] The maintenance inspection reported no problems

[a] Matsumoto Airport

[b] Fukuoka Airport

[c] New Chitose Airport

[d] Nagoya Airfield

Table 2 Weather conditions at the time of operation on days when turbulence occurs

Date	Duration	Altitude (min–max) feet	Airspeed (min–max) kt	Windspeed (min–max) kt	Winddirection (min–max)°	Temperature (min–max) ℃
12/12/2017	10 s	12,026–12,466	174–184	45.3–82.1	276–287	− 19.2 to − 21.5
12/27/2017	7 s	12,832–13,121	229–255	58.5–82.9	280–316	− 24.8 to − 27.0
01/23/2018	14 s	12960–13350	198–251	23.5–72.4	230–268	− 22.9 to − 27.1

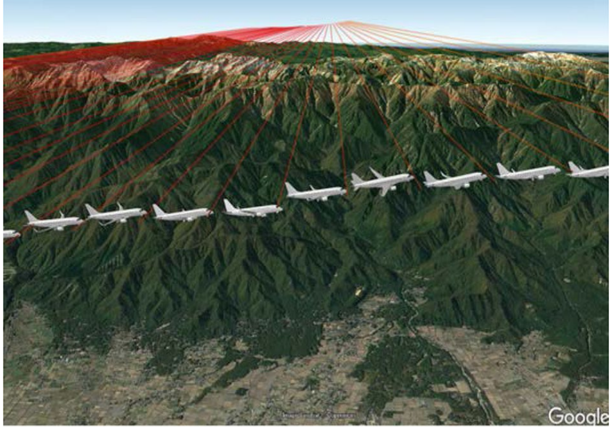

Fig. 3 Visualization of altitude changes owing to turbulence

operation on the day of the turbulence event. These values were found to be significantly different from the conditions during normal operations.

Visualization of the wind direction, speed of mountain waves, and sway of aircrafts

The first step toward solving the problem involves visualizing the turbulence and its resulting impact on operations. Thus, a visualization depicting a severe turbulence scenario was created, wherein the altitude changes during turbulence were modeled as per the flight of an E170 aircraft. Further, the aircraft altitude at every second was depicted using Google Earth Pro 7.3.4. Figure 3 shows the visual representation of a journey via FDA Flight 211 in January 2018, wherein the pilot encountered severe turbulence during ascent. Latitude, longitude, altitude, heading, pitch, and roll recorded by the aircraft were reflected in the parameters of Google Earth for accurate rendering. The average wind directions during the turbulence were represented using red lines. In addition, the wind blew over the Northern Alps directed towards the aircraft (from the back to the front of the figure). Consequently, significant altitude changes were observed during this period.

Elementary analysis of turbulence occurrences using open data

To create the dataset, weather information from October 1, 2017, through March 31, 2018, were obtained from Sunny Spot [29], which is the website homepage of the Japan Meteorological Agency [30]. In addition, an environmental database provided by Iowa State University [31] was used as well. Subsequently, a dataset with 165 rows and 45 columns was created as an explanatory variable. Table 3 summarizes the items in this dataset. Using real-world QAR data from a pilot report provided by FDA, Yes/No values were obtained for indicating whether any FDA flights that either departed from or landed at Matsumoto airport encountered a greater than moderate ("moderate-plus") or higher level of daily turbulence during the observation period. Three instances of moderate-plus turbulence exist in the data used in this study. These data were described based on "location-time-altitude-type."

Figure 4 illustrates the boxplots of fx106-03-500-spd, Wajima-12-700-temp, Matsumoto-12-500-hum, and fx106-03-500-shear; here, all data are normalized. The circle, triangle, and square in each boxplot represent the instances of turbulence in the data. In addition, on the days when turbulence was observed, the wind speed and shear were high, while the temperature was low [33].

Methods

Turbulence-occurrence analysis using PCA

Owing to a lack of sufficient data for observing patterns in annual turbulence, predicting its occurrence through supervised learning is challenging [34]. In addition, there exists a possibility of weather conditions affecting operations on days other than those on which turbulence was reported. Further, meteorological data comprises many explanatory variables, and determining the variables that contribute to turbulence is complex. Thus, in this study, to supplement the scarce information on the day of turbulence occurrence, represent the weather conditions affecting flight operations as well as contributing to the decision-making process of pilots and airlines in implementing flight operations in high-risk environments, the formation of risk clusters was determined using PCA and statistical information on their weather conditions. In addition, a method for determining forecast accuracy was applied as well. First, the limits of the explanatory variables in the PCA were reduced and the weights were used to calculate the risk clusters employing the k-means method. Consequently, the risk cluster obtained was used to predict the occurrence of turbulence through SVC. The program was executed in the Python 3.7.0 environment and scikit-learn version 0.23.2 was used as well. The algorithm is described as follows.

(1) Creation of a dataset for turbulence predictions, using open data

Table 3 List of data used

Weather data	Time (UTC)	Altitude (hPa)	Descriptive variables	
			Name	**Description**
fx106 (FXJP106) [32]	0000Z	500	spd	Wind speed (knots)
	0300Z	700	shear	Wind speed difference[a] (knots/feet)
fx502 (FXFE502) [32]	0000Z	500	low	Low pressure (yes = 1, no = 0)
			trough	Trough (yes = 1, no = 0)
			alt	Number of contour lines between Wajima and Tateno
MA (Matsumoto high-rise weather)	1200Z[b]	500	spd	Wind speed (knots)
		700	hum	Humidity (%)
			dir	Wind direction (°)
			temp	Temperature (°C)
WA (Wajima high-rise weather)	1200Z[b]	500	spd	Wind speed (knots)
		700	hum	Humidity (%)
			dir	Wind direction (°)
			temp	Temperature (°C)
TA (Tateno high-rise weather)	1200Z[b]	500	spd	Wind speed (knots)
		700	hum	Humidity (%)
			dir	Wind direction (°)
			temp	Temperature (°C)
MMJ meteorological terminal air report (METAR)[c]	2310Z		temp	Air temperature (°C)
			dwp	Dew point (°C)
			relh	Relative humidity (%)
			dir	Wind direction (°)
			spd	Wind speed (knots)
			alt	Altimeter (inches)
			vsby	Visibility (miles)
			gust	Wind gust (knots)
			vis1[d]	Cloud height level 1 (feet)
			vis2[d]	Cloud height level 2 (feet)
			vis3[d]	Cloud height level 3 (feet)

[a] 500 hPa only

[b] Previous day

[c] Refers to the weather at Matsumoto Airport. Only the names of METAR elements are used; the time and altitude are not described

[d] Not available (NA) values in this field were replaced with 10,000

(2) Calculation of turbulence risk cluster

(a) A projection matrix W is created via PCA [35]

(i) Let the i-th data be $\boldsymbol{x_i}$, and let $\boldsymbol{Y}$ be all the rows of the data matrix $\boldsymbol{X}$ minus $\overline{x}$. $\boldsymbol{Y} = \begin{pmatrix} (\boldsymbol{x_1} - \overline{x})^T \\ \vdots \\ (\boldsymbol{x_n} - \overline{x})^T \end{pmatrix} = \left(x_{ij} - \frac{1}{n} \sum_{k=1}^{n} x_{kj} \right)$. The covariance matrix $\boldsymbol{S}$ of $\boldsymbol{X}$ is as follows: $\boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{x})(\boldsymbol{x}_i - \overline{x})^T = \frac{1}{n} \boldsymbol{Y}^T \boldsymbol{Y}$

(ii) Let $\boldsymbol{S}$ be decomposed into singular values, $\boldsymbol{S} = \boldsymbol{U} \Sigma \boldsymbol{V}$, and let $\boldsymbol{V}^{(c)}$ be the number of dimensions c acquired from $\boldsymbol{V}$ following dimensionality reduction. Let $\boldsymbol{W} = \boldsymbol{V}^{(c)}$.

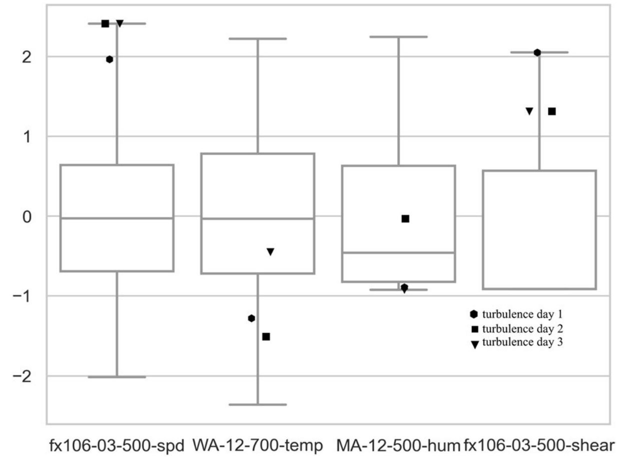

Fig. 4 Dataset comparison by wind speed, temperature, humidity, and shear difference

(b) The data are converted to principal component (PC) vector Z using W, such that $\boldsymbol{Z} = \boldsymbol{Wx}$.

(c) The risk clusters are generated based on Z using the k-means method [19].

(i) If the set of indices of $\boldsymbol{x_i}$ belonging to the j-th cluster is I_j, the center of gravity G_j of the cluster is $G_j = \frac{1}{|I_j|}\sum_{i \in I_j} \boldsymbol{x_i}$.

(ii) For each $\boldsymbol{x_i}$, calculate the distance from the center of gravity and repeat assigning to the cluster with the closest distance.

(3) Prediction of turbulence using risk clusters

(a) Prediction of turbulence-occurrence dates via SVC [35, 36]. Test data set $\boldsymbol{X'}$ is converted using $\boldsymbol{W}, \boldsymbol{Z'} = \boldsymbol{WX'}$.

(i) Consider the following optimization problem for a map $\phi : \mathbb{R}^c \to \mathbb{R}^c$,

$$\text{Maximize } f(\boldsymbol{a}) = \sum_{k=1}^{n} a_k - \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n} a_k a_l y_k y_l K\left(\boldsymbol{z'_k}, \boldsymbol{z'_l}\right).$$

$$\text{Subject to } \sum_{i=1}^{n} a_i y_i = 0, 0 \le a_i \le C.$$

$$\text{Where } \phi\left(\boldsymbol{z'_k}\right)^T \phi\left(\boldsymbol{z'_l}\right) = K\left(\boldsymbol{z'_k}, \boldsymbol{z'_l}\right),\ \boldsymbol{a} = (a_1, a_2, \ldots, a_n)^T,\ y_i \in \{-1, 1\}.$$

(b) Validation of predicted turbulence-occurrence dates.

Dimensionality reduction and coordinate transformation in PCA

PCA was employed to determine the factors that cause turbulence. Figure 5 depicts a plot for each observation date, with PCs 1 and 2 forming the x- and y-axes, respectively; the points indicated by arrows represent the three actual instances of turbulence. Flights with turbulence are plotted in the upper-right part of the figure. Figure 6 illustrates a scatter plot of the elements comprising the first and second PC

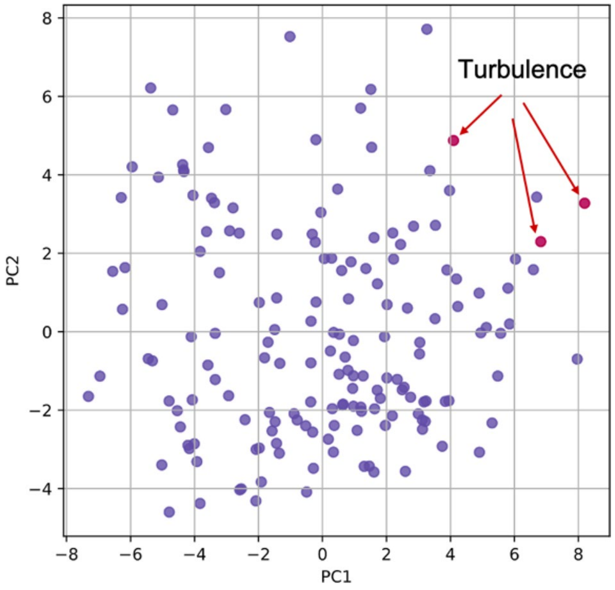

Fig. 5 Scatter plot of PC1 and PC2

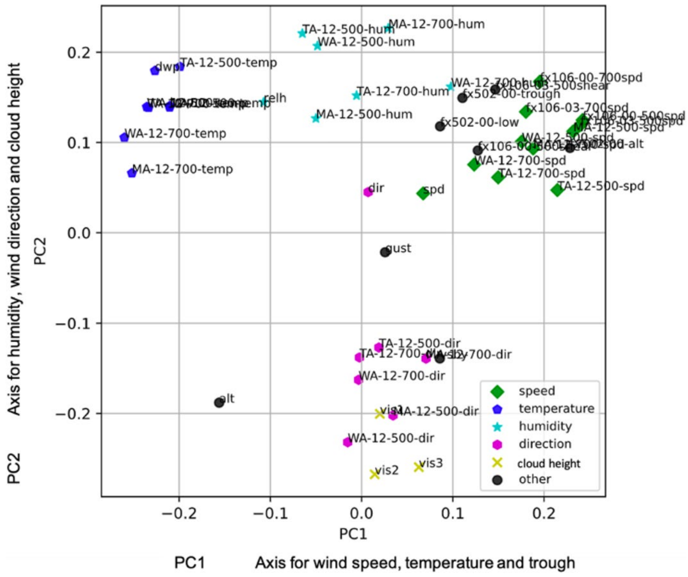

Fig. 6 Plot of each element on PC1 and PC2 planes

planes. As can be observed, the wind speed, contour lines, and trough elements are concentrated in the upper-right quadrant of the PC1-axis; the temperature elements, in the upper-left. It can be inferred that the farther the PC1 lies on the right-hand side, the higher the wind speed and the lower the temperature (i.e., there are many contour lines). Further, the humidity elements are present at the top of the plot, while wind direction and cloud height elements at the bottom indicating that in case of high humidity, the wind direction is negative. Thus, when most of the wind is from the west, on occurrence of turbulence, the wind from the southwest exerts a significant influence on the aircraft. Furthermore, the cloud height is low toward the top of the PC2 axis. The values of these PC loads are listed in Table 4. Figure 5 reveals that in PC4, troughs and cloud height were major influencing factors on the days when the turbulence occurred. Further, it can be observed that wind speed difference significantly influences the PC5.

The cumulative contribution rate from PC1 to PC2 was determined as 43.23%, with 13 components required to achieve a cumulative contribution of at least 80%. Therefore, 13 PCs were considered to obtain the matrix $\boldsymbol{W}$ that performs coordinate transformations based on $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{x}$. Here, $\boldsymbol{x}$ is the original data and $\boldsymbol{Z}$ is the coordinate after transformation.

Calculation of risk clusters by k-means method

Using the coordinate transformation matrix obtained from the PCA described in the previous Section, the risk cluster was calculated employing the k-means method, wherein the $\boldsymbol{Z}$ coordinate transformed via $\boldsymbol{W}$ was used. Figure 7 shows the resulting classification into six clusters. Clusters where turbulence was expected to occur are indicated in red, and included almost all the dates on which turbulence was observed, as presented in Table 1. However, although Cluster ID 5 might have been

Table 4 Value of each PC load

PC	Item	Value
PC1 (26.5%)	WA-12-700-temp	– 0.26
	MA-12-700-temp	– 0.2518
	fx106-00-500-spd	0.2423
	fx106-03-500-spd	0.2399
	WA-12-500-temp	– 0.2354
PC2 (16.7%)	vis2	– 0.2675
	vis3	– 0.2595
	WA-12-500-dir	– 0.2319
	MA-12-700-hum	0.2268
	TA-12-500-hum	0.2209
PC3 (7.57%)	TA-12-700-spd	0.3143
	WA-12-700-spd	0.2914
	WA-12-500-spd	0.2871
	TA-12-700-dir	0.2266
	MA-12-700-temp	0.2262
PC4 (5.42%)	spd	0.4332
	relh	– 0.3829
	fx502-00-trough	0.2476
	gust	0.2439
	vis1	0.2398
PC5 (4.05%)	dir	– 0.4083
	fx106-00-500-shear	0.406
	fx106-03-500-shear	0.3111
	WA-12-500-hum	– 0.2732
	MA-12-500-hum	– 0.2324

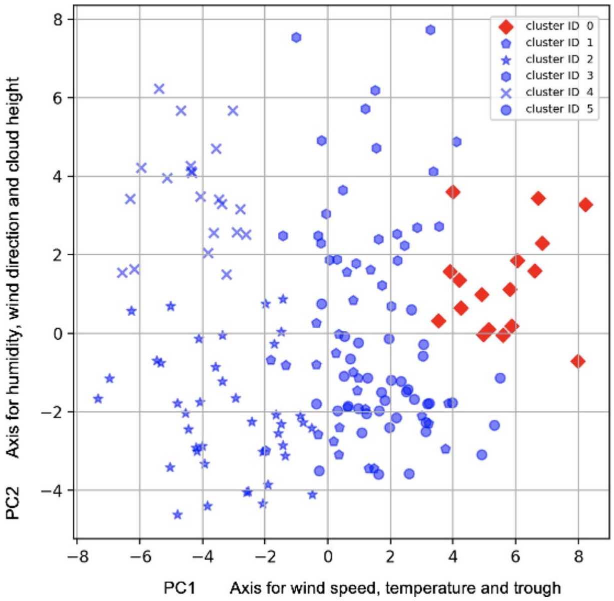

Fig. 7 Six clusters obtained via the k-means method

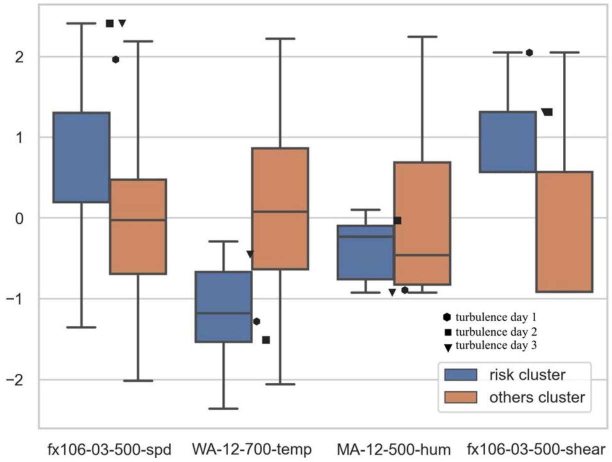

Fig. 8 Comparison of risk cluster and other clusters

affected by turbulence, it did not significantly affect flight operations. Moreover, it is also probable that the other clusters were less affected by turbulence.

Figure 8 presents a comparison of the risk clusters with other clusters. It is evident that the risk clusters exhibit faster wind speeds, lower temperatures, lower humidity, and larger wind speed differences. Further, T-test or Welch's test conducted on the risk cluster and the other clusters showed that the p-value was less than 0.05, confirming that the means of the two groups were significantly different for all the items in Fig. 8.

Table 5 Usage data and SVC parameters

Item	Value
Usage data year	2019
Corresponding dates	01/01–03/31, 10/01–12/31
Number of data points	179
Kernel function	Gaussian kernel
Gamma	1/(number of data points × variance of data)
C	1.0

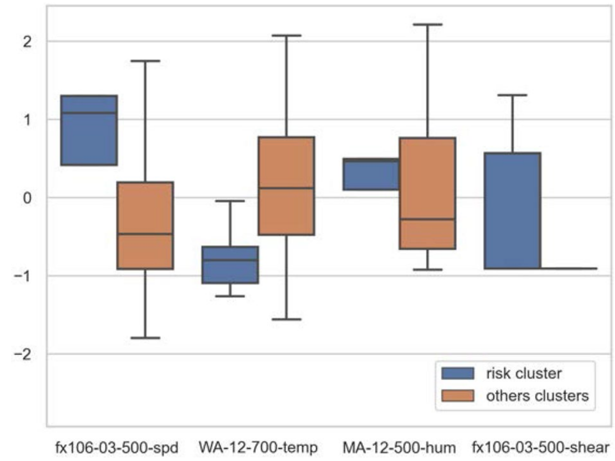

Fig. 9 Comparison of days with and without turbulence predicted for the 2019 data

Result and discussion

Turbulence prediction for validation data

The risk clusters described in the previous chapter were used to predict the occurrence of turbulence using 179 data points that were collected from Table 3 in the year 2019. Following the normalization of this data, axis transformation was performed using the transformation matrix $\boldsymbol{W}$ described in the previous section.

Calculation of risk date using the risk cluster via SVC

The risk cluster was used as the training data to predict the turbulence dates for the 2019 data using SVC. Table 5 lists the validation data and SVC parameters.

Figure 9 presents a comparison of the days that were predicted to exhibit turbulence with those that were not. It is evident that the distributions of wind speed and temperature are similar to those of the risk cluster. For fx106-03-500-shear, all values between 10/01/2019 and 12/31/2019 equaled 0. It was concluded that the days with predicted turbulence exhibited strong wind speeds, low temperatures, and large wind speed differences.

Table 6 Verification of the predicted turbulence days using the weather map

Date	Level	Overview of weather map information
01/05/2019	3. Warning	9 kt/1000 ft shear (wind speed change per altitude) is expected from FL200[a] to FL150 over Matsumoto, and there is a high probability of mountain waves
01/08/2019	2. Caution	There is a 200 kt jet over Kyushu, and it appears that the shaking will increase over the next day; however operations at this level are possible
01/09/2019	4. Critical	Under these conditions, cancellation of the flight operation is being considered. Moderate to severe turbulence is expected for a wide range of altitudes (FL180–FL7000); therefore, maximum caution is required At Hanamaki Airport, four FDA flights were canceled owing to strong winds near the airport
01/20/2019	3. Warning	Two strong jet streams are approaching. Accretion and shaking are expected on the Sea of Japan side, and the wind over the mountains appears to be strong. Aircrafts can be operated, but only with extreme caution
02/04/2019	3. Warning	Shear is expected below 10,000 ft and requires considerable caution After the cold front passes, the wind becomes stronger and mountain waves are expected

[a] FL represents the flight level, i.e., aircraft altitude at standard air pressure expressed in 100 s of feet

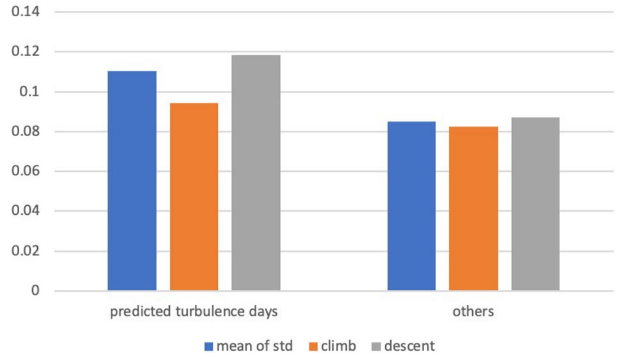

Fig. 10 Comparison of the average of the maximum SDs of vertical sway for days with and without predicted turbulence from QAR data

Verification of forecasted turbulence dates via SVC

Through the use of a weather map, the days with the risk of turbulence predicted using the risk clusters and SVC were verified; the results are summarized in Table 6. Turbulence risk was assigned based on four levels, categorized in increasing order of risk: 1 (normal), 2 (caution), 3 (warning), and 4 (critical), to render it easier to propose to airlines. Herein, the highest risk was observed on January 9, 2019, when flight cancellations were considered. Moreover, even on the dates when the turbulence risk level was at least two, passenger safety, if not flight cancellation, were seriously considered. Therefore, it was confirmed that this analysis can adequately predict turbulence-risk days.

Figure 10 presents a comparison of the per-minute average of the maximum standard deviations (SDs) of the vertical sway of the aircraft [37, 38] obtained from actual QAR data against those of the predicted turbulence date calculated via SVC and the other days. However, 02/04/2019 was excluded because QAR data could not be obtained for the said date. As can be observed, from the left, the graph shows the average of the maximum SDs of the vertical sway, and the values concerning its climb and descent. Further, the vertical sway can be observed to be generally larger on the days wherein turbulence is predicted. Moreover, there is considerable shaking observed during descent.

Comparison with other methods

The proposed method was compared with other methods. Table 7 shows the results of validation of the data in Table 3 using the cross-validation method (K = 10), where the records for the days when no turbulence occurred were set as true, and the accuracy was the highest among the methods used. The results of all methods and models show that the FN: False Negative item, which detects the days when turbulence is observed, is 0, thereby indicating that turbulence occurrence was not detected.

Conclusion

This study used open data to predict the occurrence of turbulence to render aircraft operations safer and more comfortable. Although turbulence occurs infrequently, it is a leading cause of aircraft damage and changes in flight schedules. The findings of this study are twofold. First, following the confirmation of the statistical information using the risk clusters, they were used as supervisory data to make appropriate predictions even for low frequency events such as turbulence. Moreover, the turbulence-risk cluster was derived through k-means clustering after reducing the dimensions of available data via PCA, instead of using the rare instances of turbulence as the training data. In addition, the process of creating risk clusters provided an opportunity to examine the factors that influenced turbulence occurrence. In the case of high-risk events such as aircraft operations, this can have a synergistic effect with the experience and knowledge of the pilots themselves. Further, using this turbulence-risk cluster as training data, the turbulence occurrences for 2019 were predicted through SVC, with the obtained results being confirmed to be sufficiently accurate for utilization by pilots. Second, it was found that using open data, the prediction of turbulence occurrence was possible. Further, the meteorological data used in this study is routinely used by pilots and airlines, and thus can be used at airports other than the one covered in this study.

Table 7 Comparison with other machine learning methods

Method	TP	TN	FP	FN	Model
Proposed method	148	14	1	2	Use statistical analysis as well
Tree	158	4	3	0	Fine tree, medium tree, coarse tree
LDA	160	2	3	0	Linear discriminant
Logistic regression	162	0	3	0	
Kernel Naive Bayes	162	0	3	0	Kernel Naive Bayes, kernel type: Gaussian
SVM	162	0	3	0	Linear SVM, quadratic SVM, cubic SVM, fine Gaussian SVM, medium Gaussian SVM, coarse Gaussian SVM
KNN	162	0	3	0	Fine KNN, medium KNN, coarse KNN, cosine KNN, cubic KNN, weighted KNN
Ensemble	162	0	3	0	Boosted trees, bagged trees, subspace discriminant, subspace KNN
Neural network	161	1	3	0	Narrow neural network, medium neural network, wide neural network, bilayered neural network, trilayered neural network
Kernel	159	3	3	0	SVM kernel, logistic regression kernel

TP True Positive, *TN* True Negative, *FP* False Positive, *FN* False Negative

However, there exist certain issues that need to be addressed in the future. As the present study was focused on aircraft taking off from and landing at airports in Japan, the impact of the season is significant. The occurrence of turbulence over Japan is concentrated in the winter season. Although the present data can be used to predict turbulence in Japan, further data is essential to cover all regions of the world. Moreover, to generalize the model, the availability of such data in various parts of the world must be investigated.

Although this study was conducted for predicting turbulence occurrence for Matsumoto Airport, the same method can be employed to analyze turbulence for other airports. Further, it is suggested that the prediction accuracy of the proposed technique can be improved via the combination of daily aircraft data with open data, such as weather data. The proposed method is expected to aid in turbulence prediction and also result in increased systems expertise and technological advancements in combating turbulence, to compensate for future human resource shortages in aviation.

Abbreviations
PCA: Principal component analysis; FDA: Fuji Dream Airlines; SVC: Support vector classification; QAR: Quick access recorder; SD: Standard deviation.

Acknowledgements
The authors would like to thank Fuji Dream Airlines Co., Ltd. for providing information useful for this study.

Authors' contributions
KI and SM conceptualized the study; SM designed the study; SM, HO, and KI were responsible for data acquisition and analysis; SM and HO were responsible for data interpretation; SM and HO contributed new methods or models or software; SM was responsible for writing the draft manuscript and revisions. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
All data generated or analyzed during this study are included in this published article. The publicly available dataset and the source code for the analysis can be found at the following link Github: https://github.com/smzn/Turbulence.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Shizuoka Institute of Science and Technology, 2200-2, Toyosawa, Fukuroi, Shizuoka 437-8555, Japan. [2]J.F. Oberlin University, 3758 Tokiwa-machi, Machida-shi, Tokyo 194-0294, Japan.

Received: 8 November 2021 Accepted: 18 February 2022
Published online: 07 March 2022

References
1. Digest of aircraft accident analyses for prevention of accidents due to the shaking of the aircraft. In: JTSB digest, vol. 15. Japan Transport Safety Board. 2015. https://www.mlit.go.jp/jtsb/bunseki-kankoubutu/jtsbdigests_e/jtsbdigests_No15/No15_pdf/jtsbdi-15_all.pdf. Accessed 7 Dec 2020.
2. Cabinet Office in Japan. White paper on traffic safety in Japan 2018. https://www8.cao.go.jp/koutu/taisaku/h30kou_haku/english/wp2018-pdf.html. Accessed 7 Dec 2020.
3. Regular Airlines Association. Movements towards alternative aviation fuel use in the aviation industry. https://www.meti.go.jp/committee/kenkyukai/energy_environment/biojet/pdf/001_02_00.pdf. Accessed 12 Sept 2018.

4. International Aircraft Development Fund. Trends in flight data analysis technology (FDM/FOQA). http://www.iadf.or.jp/document/pdf/24-7.pdf. Accessed 24 Dec 2020. (**In Japanese**).
5. Dubois P, AIRBUS—Airlines SMS & FDA Assistance. Flight data analysis. https://www.icao.int/SAM/Documents/2015-FDA/1.3%20AIRBUS%20PDubois%20FDA-Seminar_1%20Recording.pdf. Accessed 24 Dec 2020.
6. Japan Aerospace Exploration Agency. Smart flight technology. https://www.aero.jaxa.jp/eng/research/star/smart-flight/. Accessed 29 Aug 2018.
7. Japan Aerospace Exploration Agency. Demonstration of turbulence prevention airframe technology (SafeAvio). https://www.aero.jaxa.jp/eng/research/star/safeavio/news170313.html. Accessed 7 Dec 2020.
8. Clark TL, Hall WD, Kerr RM, Middleton D, Radke L, Ralph FM, et al. Origins of aircraft-damaging clear-air turbulence during the 9 December 1992 Colorado downslope windstorm: numerical simulations and comparison with observations. J Atmos Sci. 2000;57:1105–31.
9. Lilly DK. A severe downslope windstorm and aircraft turbulence event induced by a mountain wave. J Atmos Sci. 1978;35:59–77.
10. Parker T, Lane T. Trapped mountain waves during a light aircraft accident. AMOJ. 2013;63:377–89.
11. Sharman R, Tebaldi C, Wiener G, Wolff J. An integrated approach to mid- and upper-level turbulence forecasting. Weather Forecast. 2006;21:268–87.
12. Huang R, Sun H, Wu C, Wang C, Lu B. Estimating Eddy dissipation rate with QAR flight big data. Appl Sci. 2019;9:5192.
13. Lee JCW, Leung CYY, Kok MH, Chan PW. A comparison study of EDR estimates from the NLR and NCAR algorithms. Atmosphere. 2022;13:132.
14. Haverdings H, Chan PW. Quick access recorder data analysis software for windshear and turbulence studies. J Aircr. 2010;47:1443–7.
15. Ralph FM, Neiman PJ, Levinson D. Lidar observations of a breaking mountain wave associated with extreme turbulence. Geophys Res Lett. 1997;24:663–6.
16. Williams JK. Using random forests to diagnose aviation turbulence. Mach Learn. 2014;95:51–70.
17. Veermann H, Vrancken P, Lombard L. Flight testing DELICAT—a promise for medium-range clear air turbulence protection. Luleå, Schweden; 2014. https://elib.dlr.de/91968/. Accessed 28 Jan 2022.
18. Hamilton DW, Proctor FH. convectively induced turbulence encountered during NASA's fall-2000 flight experiments. 2002. https://ntrs.nasa.gov/citations/20030015754. Accessed 28 Jan 2022.
19. Hamada Y, Kikuchi R, Inokuchi H. LIDAR-based gust alleviation control system: obtained results and flight demonstration plan. IFAC-PapersOnLine. 2020;53:14839–44.
20. Inokuchi H, Tanaka H, Ando T. Development of an onboard Doppler Lidar for flight safety. J Aircr. 2009;46:1411–5.
21. Oikawa H, Inokuchi H, Izumi K, Kikuchi Y, Hayasaki N. Relation between resolution enhancement and accuracy in prediction of turbulence area which perturbs aircraft. Tenki. 2010;57(9):669–80 (**In Japanese**).
22. Aviation today: delta develops artificial intelligence tool to address weather disruption, improve flight operations. https://www.aviationtoday.com/2020/01/08/delta-develops-ai-tool-address-weather-disruption-improve-flight-operations/. Accessed 7 Dec 2020.
23. Muñoz-Esparza D, Sharman RD, Deierling W. Aviation turbulence forecasting at upper levels with machine learning techniques based on regression trees. J Appl Meteorol Clim. 2020;59:1883–99.
24. Oster CV, Strong JS, Zorn K, editors. Why airplanes crash: causes of accidents worldwide. In: 51st annual transportation research forum conference paper; 2010.
25. Weli V, Emenike G. Turbulent weather events and aircraft operations: implications for aviation safety at the Port Harcourt international airport, Nigeria. IJWCCCR. 2016;2:11–21.
26. Arowolo MO, Adebiyi MO, Adebiyi AA. An efficient PCA ensemble learning approach for prediction of RNA-Seq malaria vector gene expression data classification. Int J Eng Res Technol. 2020;13:163–9.
27. Arowolo MO, Adebiyi MO, Adebiyi AA, Okesola OJ. A hybrid heuristic dimensionality reduction methods for classifying malaria vector gene expression data. IEEE Access. 2020;8:182422–30.
28. Carney TQ, Bedard Jr AJ, Brown JM, McGinley J, Lindholm T, Kraus MJ. Hazardous mountain winds and their visual indicators. US Department of Commerce, National Oceanic and Atmospheric Administration; 1995. p. 55.
29. Sunny Spot Inc. Weather and climate information site. https://www.sunny-spot.net/chart/chart_archive.html. Accessed 10 Dec 2020.
30. Japan Meteorological Agency. Historical weather data retrieval (high rise), https://www.data.jma.go.jp/obd/stats/etrn/upper/index.php. Accessed 10 Dec 2020.
31. Iowa State University. Iowa environmental Mesonet: download ASOS/AWOS/METAR data. https://mesonet.agron.iastate.edu/request/download.phtml. Accessed 10 Dec 2020.
32. Japan Meteorological Agency. Aviation forecast. https://www.data.jma.go.jp/add/suishin/cgi-bin/catalogue/make_product_page.cgi?id=KokuYoho. Accessed 24 Dec 2020. (**In Japanese**).
33. Center VAA. Characteristics of high-degree disturbances near Japan. Tenki. 1967;14(5):179–87 (**In Japanese**).
34. Fukui K. Machine learning learned with Python and examples: identification/prediction/abnormality detection. Ohmsha 2018. (**In Japanese**).
35. Yang XS. Optimization techniques and applications with examples. 1st ed. New Jersey: Wiley; 2018.
36. Bishop CM. Pattern recognition and machine learning. 1st ed. New York: Springer; 2006.
37. Sato M, Endo E. Spectrum analysis for turbulence and induced accelerations. J Jpn Soc Aeronaut Space Sci. 2008;56(Supp 653):293–5 (**In Japanese**).
38. Prince JB, Buck BK, Robinson PA, Ryan T. In-service evaluation of the turbulence auto-PIREP system and enhanced turbulence radar technologies. NASA, CR-2007-214887. 2007.

Publisher's Note