Advancing Clear-Air Turbulence Detection with Hybrid Predictive Models for a Regional Aviation Corridor in Southeast Brazil

Alessana Rosette *, Gutemberg França, Haroldo Velho, Heloisa Ruivo, Ivan Melo

Posted Date: 26 February 2026

doi: 10.20944/preprints202602.1615.v1

Keywords: Clear-Air Turbulence (CAT); hybrid forecasting model; Machine Learning in Meteorology; Global Forecast System (GFS); aviation safety; dimensionality reduction

 doi:10.20944/preprints202602.1615.v1

Article

Advancing Clear-Air Turbulence Detection with Hybrid Predictive Models for a Regional Aviation Corridor in Southeast Brazil

Alessana Rosette [1,*], Gutemberg França [1], Haroldo Velho [2], Heloisa Ruivo [3] and Ivan Melo [1]

[1] Member, UFRJ
[2] Member, INPE
[3] Independence Research
* Correspondence: alessana82@gmail.com

Abstract

Severe Clear-Air Turbulence (severe CAT) remains a relevant hazard to aviation safety, often occurring without visible atmospheric indicators. This study presents a hybrid forecasting framework that integrates outputs from the Global Forecast System (GFS025) with advanced machine learning (ML) algorithms to predict severe CAT events over Southeast Brazil, within the region bounded by 43°W to 49°W and 19°S to 25°S, from January 2018 to December 2021. To enhance predictive performance and reduce model complexity, a statistically robust dimensionality reduction technique was applied using p-value filtering and False Discovery Rate (FDR) control, resulting in a refined set of 13 physically interpretable predictors. Key turbulence indices, such as Ellrod's index (ELL2) and Brown's index (BROWN), emerged as the most relevant features for classification. Nine ML algorithms were tested and evaluated through Receiver Operating Characteristic (ROC) analysis and Area Under the Curve (AUC) scores. The Multi-Layer Perceptron (MLP) model, with a single hidden layer of 10 neurons, achieved the highest AUC (0.95), followed closely by Random Forest (0.94), demonstrating the effectiveness of relatively simple architectures when coupled with feature selection. These findings underscore the value of combining physically consistent diagnostics with data-driven methods for regional severe CAT forecasting. The proposed approach offers a scalable and adaptable framework that supports enhanced aviation safety and provides a solid foundation for the continued development of operational turbulence prediction tools.

Keywords: Clear-Air Turbulence (CAT); hybrid forecasting model; Machine Learning in Meteorology; Global Forecast System (GFS); aviation safety; dimensionality reduction

1. Introduction

The atmosphere, a geophysical fluid governed by physical laws, continually seeks a dynamic equilibrium. Disturbances prompt dynamic adjustments across scales ranging from millimeters to planetary dimensions. These adjustments give rise to a wide variety of meteorological phenomena, such as cloud formation, precipitation, storms, snowfall, hail, winds, and turbulence. Each of these processes, though driven by the same fundamental principles, interacts uniquely with atmospheric conditions, contributing to the complex dynamics of weather and climate systems [1–3].

Among these phenomena, CAT presents a particularly insidious challenge to aviation. Unlike other forms of turbulence associated with planetary boundary layer or visible meteorological phenomena, CAT occurs in clear skies without any visual indicators such as clouds or precipitation. This characteristic makes it one of the most difficult types of turbulence to predict and avoid. CAT often manifests suddenly, posing significant operational and safety risks, including flight delays, diversions, and, in severe cases, injuries to passengers and crew [4–6]. The National Transportation

 doi:10.20944/preprints202602.1615.v1

Safety Board (NTSB) of the USA has identified CAT as a leading cause of weather-related injuries in commercial aviation, emphasizing its profound impact on flight operations and safety [7,8].

Historically, CAT forecasting has evolved through the development of diagnostic turbulence indices derived from numerical weather prediction (NWP) models, including the Richardson Number and Ellrod Indices, designed to identify atmospheric instability and wind shear regions [9]. More recently, the Graphical Turbulence Guidance (GTG) system has been widely used for evaluating turbulence diagnostics based on GFS outputs and pilot reports (PIREPs). In this context, Kim et al. [10] evaluated the performance of upper-level turbulence diagnostics over East Asia using the GTG system, demonstrating how different turbulence indices behave across regional airspaces. Building on this, Kim et al. [11] presented improvements to non-convective turbulence forecasts within the World Area Forecast System (WAFS), highlighting enhancements in global CAT prediction using GFS-based turbulence products. In parallel with these advancements, recent studies have explored the application of machine learning to CAT forecasting. Muñoz-Esparza et al. [12] demonstrated the effectiveness of regression tree ensembles for predicting upper-level turbulence, while Lee et al. [13] applied deep learning techniques to satellite-based observations to estimate turbulence intensity. These studies reinforce the growing role of data-driven approaches as a complement to traditional NWP-based diagnostics, broadening the scope and accuracy of CAT forecasting models.

With the rise of machine learning (ML), additional possibilities have emerged for enhancing CAT forecasting by capturing nonlinear interactions and complex dependencies within large atmospheric datasets. Several recent works have applied ML to turbulence prediction problems. For instance, Menegardo-Souza et al. [14] developed ML-based models using turbulence reports from flights along the Santiago–Mendoza corridor in South America — mainly associated to the Andes mountains forcing, demonstrating gains over traditional diagnostics.

Despite these global advances, Brazil currently lacks an operational CAT forecasting system tailored to its meteorological and operational context, particularly over Southeast Brazil—its busiest airspace and one with a high incidence of turbulence-related flight disruptions [15]. Between 2013 and 2023, São Paulo, Rio de Janeiro, and Minas Gerais consistently ranked among the states with the highest reports of turbulence along commercial routes, and this Brazilian region has the most intense aviation traffic.

Although several previous studies have incorporated GFS model outputs into CAT forecasting, this study contributes distinct innovations. First, a regionally calibrated hybrid model is developed focused on Southeast Brazil, combining local atmospheric conditions with tailored statistical modeling. Second, we introduce a feature selection approach based on statistical significance (p-value) and FDR control, adapted from genomic analysis, to isolate physically and statistically relevant predictors from over 67,000 GFS-derived attributes. Third, we utilize high-resolution GFS data (0.25° × 0.25°) extending vertically from FL180 to FL500, allowing detailed characterization of turbulence-relevant structures.

The primary objective is to evaluate the performance of multiple machine learning algorithms, trained on GFS-derived features and vertical acceleration records (VRTG) from LATAM Airlines' Airbus A320 fleet, to classify CAT and non-CAT events. Auxiliary datasets (e.g., radiosonde profiles, METARs, satellite imagery, and synoptic analyses) support a robust classification of turbulence occurrences.

By filling a key operational gap and implementing a statistically hybrid approach tailored to Brazil's busiest flight region, this study aims to advance regional CAT forecasting and contribute to improved aviation safety.

To illustrate the limitations of traditional turbulence diagnostics, we examined two confirmed severe CAT events recorded by LATAM A320 aircraft over Southeast Brazil at cruise levels (≥ FL180). In both cases, classical indices such as Ellrod I–II and the Brown index exhibited values close to their background levels, remaining within the lower quartile of their spatial distribution and showing no localized maxima co-located with the reported turbulence segment (Figure 1–2), suggesting that sub-

 doi:10.20944/preprints202602.1615.v1

-resolved processes (fine--scale vertical wind shear, gravity--wave breaking) can hinder the skill of conventional diagnostics even when they are operationally useful.

In Brazil, specific atmospheric features enhance the occurrence of clear--air turbulence, including subtropical jet streaks during winter, mountain--wave activity associated with the Serra do Mar and Serra da Mantiqueira, and convective outflows along the southeastern coast. Additionally, large--scale systems such as the South Atlantic Convergence Zone (SACZ) modulate wind shear and static stability. These regional drivers motivate a regionally calibrated framework rather than a one--size--fits--all approach.

This study integrates bioinformatics-inspired feature selection (False Discovery Rate, FDR), classical turbulence diagnostics (Ellrod and Brown), and probabilistic performance assessment based on the area under the receiver operating characteristic curve (AUC) within a regional CAT prediction framework. Such a combined approach has been only sparsely addressed in previous regional turbulence forecasting studies.

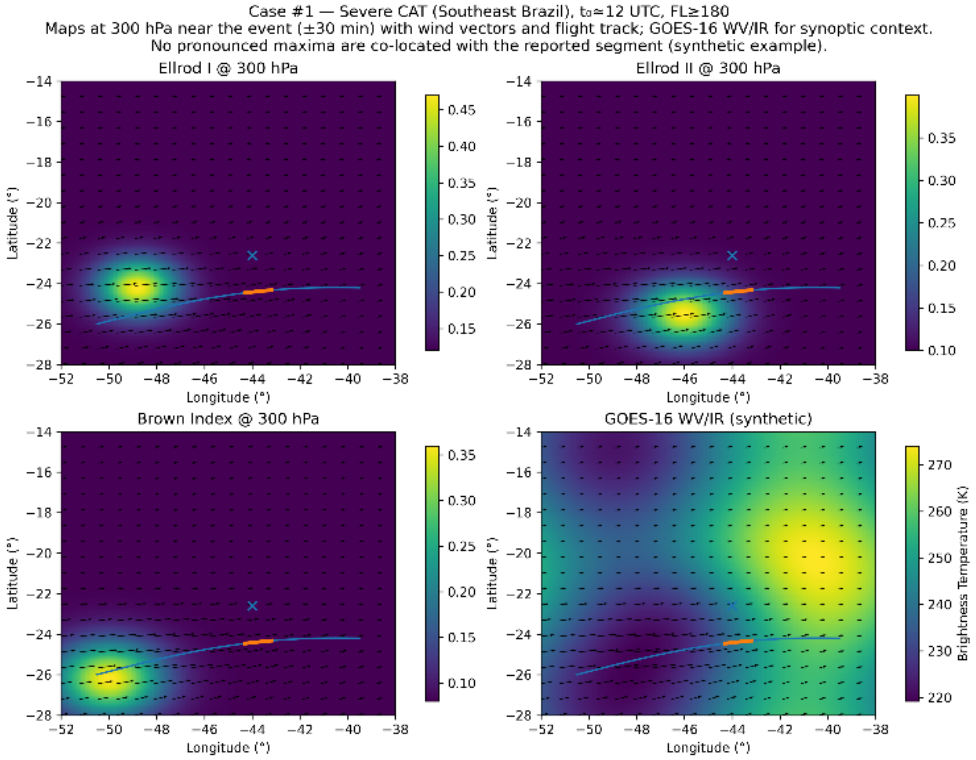


Figure 1. Case #1 (Severe CAT, Southeast Brazil, t_0≈12 UTC, FL≥180). Maps at 300 hPa of Ellrod I–II and Brown near the event (±30 min), with wind vectors and flight track; GOES--16 WV/IR included for synoptic context. No pronounced maxima are co--located with the reported segment.

 doi:10.20944/preprints202602.1615.v1

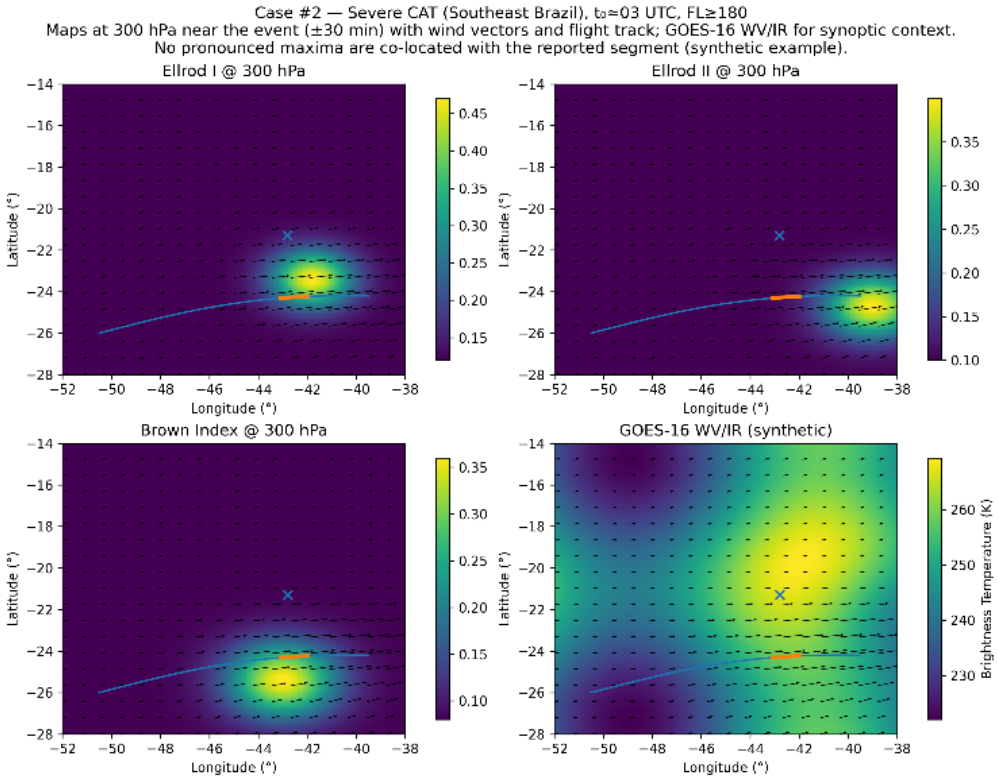


Figure 2. Case #2 (Severe CAT, Southeast Brazil, t_0≈03 UTC, FL≥180). As in Figure 1. Again, classical indices remain below typical thresholds or show displaced maxima relative to the event.

2. Materials and Methods

2.1. Study Region

The Study Region is defined as a three-dimensional atmospheric volume bounded by longitudes 43°W to 49°W and latitudes 19°S to 25°S, with a vertical range extending from 500 hPa (approximately FL180) to 100 hPa (approximately FL500), where FL (Flight Level) refers to altitude in hundreds of feet (e.g., FL200 = 20,000 feet). This region encompasses the high-traffic air corridor between Rio de Janeiro and São Paulo, operationally known as the 'Tubulão' corridor (Rio–São Paulo air route), and extends to cover additional strategic routes connecting São Paulo, Rio de Janeiro, and Belo Horizonte—the three most important metropolitan centers in Southeast Brazil.

This corridor is recognized as the busiest and most operationally critical airspace in the country, concentrating a significant portion of commercial aviation activity. It also represents one of the regions with highest incidence of reported turbulence events, particularly those associated with clear-air turbulence (CAT), due to the frequent interaction of jet streams, gravity waves, and wind shear in the upper troposphere over this area.

Figure 3 illustrates the defined Study Region (highlighted in blue). Within this area, a total of 20,338 turbulence occurrences were identified prior to methodological filtering, considering all reported VRTG events and turbulence intensity classes recorded by Airbus A320 aircraft operated by LATAM Airlines between January 1st, 2018, and December 31st, 2021. These events represent approximately 25% of the 82,951 turbulence occurrences documented by LATAM flights across South America during the same period, highlighting the disproportionately high concentration of turbulence within this specific airspace.

A detailed analysis of the CAT-related events further reveals a clear co-location between traffic density and reported turbulence incidence, confirming this region as a high-risk area for flight safety and a priority for turbulence forecasting efforts. These findings provide a strong justification for selecting this area as the focus of our study and support the broader goal of developing regionally

 doi:10.20944/preprints202602.1615.v1

adapted forecasting models capable of improving situational awareness and decision-making in complex and congested airspaces.

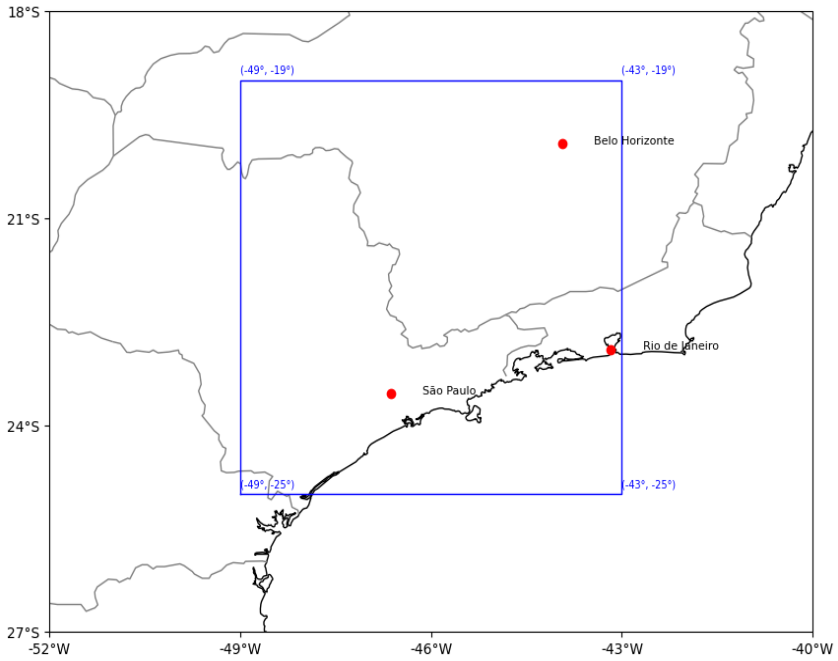


Figure 3. The blue rectangle delineates the Study Region.

2.2. Data

The data utilized in this study originates from six distinct and reliable sources, each providing critical insights for identifying, classifying, and predicting turbulence events. Together, these datasets establish a robust framework for analyzing CAT within the Study Region. Table 1 summarizes the datasets used, while their specific roles and applications are detailed below.

The Global Forecast System (GFS) data, acquired from the Research Data Archive at the National Center for Atmospheric Research, serves as the primary source of atmospheric model forecasts. These data offer global coverage with a grid resolution of 0.25° by 0.25° and are updated every three hours. To ensure alignment with turbulence events, the two forecast times closest to the occurrence of each event were selected for analysis.

Based on established physical understanding, specific atmospheric quantities, as detailed in Table I, are correlated with CAT, the three-dimensional GFS variables analyzed include wind components u (zonal) and v (meridional), vertical velocity (w), and potential temperature (θ). Derived quantities such as vertical wind shear (VWS) between the pressure level of interest and the two adjacent levels, turbulent kinetic energy (TKE), and turbulence indices, including the Richardson number [16], and the Ellrod indices [9], as well as the Brown Index [17], were calculated to enhance the dataset.

The total number of attributes—representing the value of each variable for every grid point within the defined three-dimensional rectangular volume (described in Section 2)—amounts to 67,500. This comprehensive dataset incorporates 12 three-dimensional variables, spanning 25 grid points in latitude, 25 grid points in longitude, and 9 vertical levels. Given the extensive nature of this dataset, a dimensionality reduction process was employed to isolate and prioritize the most relevant attributes for analysis.

Using the methodology outlined in the methods section, the most significant attributes (or predictor features) were selected from these variables. These attributes, combined with labels for CAT and non-CAT events, were used to construct the training and testing datasets for the machine

 doi:10.20944/preprints202602.1615.v1

learning algorithms applied in this study. This selection process ensured that the machine learning models were informed by the most critical features, thereby optimizing their predictive accuracy.

The Vertical Acceleration of Gravity (VRTG) data, provided by LATAM Airlines, served as the core dataset for identifying and classifying turbulence events in this study. To ensure consistency and standardization across all measurements, only data recorded by a single aircraft model — the Airbus A320 — were used. Covering the period from January 1st, 2018, to December 31st, 2021, this dataset includes the maximum VRTG recorded over the last 60 minutes at hourly intervals. VRTG measures deviations from the standard gravitational force (1g) experienced during turbulence. The turbulence severity is classified into three categories: Class 1 (light), Class 2 (moderate), and Class 3 (severe), as presented in Table II. These data were collected from LATAM's fleet of A320 aircraft during routine flights, with measurements taken within a monitoring window starting 10 seconds after takeoff and ending 4 seconds before landing. It is important to note that the absence of VRTG data on a given day does not imply an absence of turbulence; it may simply reflect the absence of aircraft recording the phenomenon.

The turbulence intensity thresholds adopted by LATAM are based on operational experience and consider both upward and downward deviations from 1g. The thresholds are asymmetric, reflecting the fact that upward (positive) and downward (negative) accelerations affect passenger perception and aircraft structural loads differently.

The TEMP (atmospheric soundings) were obtained from the Brazilian aeronautical meteorological network operated by the Department of Airspace Control (DECEA). TEMP data provide meteorological profiles of temperature, relative humidity, and wind at standard pressure levels. TEMP soundings were used as contextual vertical profiles at synoptic times (00Z/12Z) to characterize the background stability and shear environment near the event day, acknowledging their limited temporal resolution

METAR (Meteorological Aerodrome Reports) data, sourced from the REDEMET platform, provide real-time weather observations critical for diagnosing turbulence conditions. METAR includes parameters such as wind speed, direction, visibility, cloud cover, and pressure. Reports from major airports in the Study Region, including São Paulo (Congonhas and Guarulhos), Rio de Janeiro (Santos Dumont and Galeão), and Belo Horizonte (Confins), were analyzed to differentiate CAT from turbulence caused by other weather phenomena. These localized observations, updated hourly, complement broader datasets and enhance event classification accuracy.

The GOES-16 satellite imagery, retrieved from the CPTEC/INPE archive, offered high-resolution observations of atmospheric features relevant to turbulence. Thermal infrared imagery centered at 10.3 μm (channel 13) was the primary dataset used for detecting convective cells. Additionally, channel 8 (6.2 μm) and channel 4 (1.37 μm) were analyzed to identify cirrus clouds, gravity waves, and jet streams. These satellite images, collected at 15-minute intervals, provided critical insights into atmospheric conditions conducive to CAT formation.

The synoptic charts, also obtained from REDEMET, present a detailed overview of surface-level atmospheric conditions, including pressure systems, weather fronts, wind patterns, and temperature distributions. Updated every six hours, these charts were vital for diagnosing large-scale meteorological patterns influencing turbulence events. Combined with TEMP and GOES-16 data, synoptic charts provided a robust framework for understanding the mechanisms driving turbulence.

In summary, the VRTG data served as the foundational dataset for identifying turbulence events, while the GFS data were used to extract predictor features for machine learning models. METAR, TEMP, GOES-16, and synoptic chart data provided additional context and support for identifying predictors, assisting in distinguishing CAT events from other types of severe turbulence initially selected based on VRTG.

By combining aircraft-recorded VRTG data, satellite imagery, synoptic analyses, and model outputs, this study establishes an efficient framework for analyzing and predicting CAT. This multidisciplinary approach enhances the understanding of factors driving CAT and contributes to developing machine learning models aimed at improving aviation safety. The specific roles of the

 doi:10.20944/preprints202602.1615.v1

datasets are summarized in Table 1, and Table 2 details the thresholds used for classifying turbulence severity based on VRTG.

Table 1. Summary of datasets used in this study, including source, temporal resolution, and role in CAT analysis.

Data	Frequency	Period	Description
GFS	3h	Two closest times to VRTG events	Global grid forecast at 0.25° resolution
METAR	1h	January 1, 2018–December 31, 2021	Real-time weather observations critical for diagnosing CAT and no-CAT events
VRTG	Variable	January 1, 2018–December 31, 2021	Maximum vertical acceleration recorded over 60 minutes
TEMP	12h	Selected days	Meteorological profile for São Paulo and Rio de Janeiro at 12Z/00Z
GOES	15min	Selected days	Infrared images (channels 4, 8, 13) for convective and high-level features.
Synoptic Chart	6h	Selected days	Surface-level conditions for analyzing large-scale patterns

Table 2. VRTG based turbulence classification.

Category	Negative g	Positive g
Class 1 (Light)	$0.4 < g \leq 0.6$	$1.4 \leq g < 1.6$
Class 2 (Moderate)	$0.2 < g \leq 0.4$	$1.6 \leq g < 1.8$
Class 3 (Severe)	$g \leq 0.2$	$g \geq 1.8$

The full mathematical formulation of the predictor variables derived from GFS fields is summarized in Table 3, including wind components, thermodynamic quantities, and turbulence diagnostics used as candidate features in the supervised learning framework.

Table 3. Selected Predictors from GFS data at grids points in the study area.

Input (predictor)	Representation
Zonal wind speed at level zi (kt)	$u\,(zi)$
Meridional wind speed at level zi (kt)	$v\,(zi)$
Vertical wind speed at level zi (kt)	$w\,(zi)$
Horizontal wind speed at level zi (kt)	$ws\,(zi) = \sqrt{[u\,(zi)]^2 + [v\,(zi)]^2}$
Difference in horizontal wind speed magnitude between levels zi and zj (kt)	$\Delta ws\,(zi) = \sqrt{[u\,(zi)]^2 + [v\,(zi)]^2} - \sqrt{[u\,(zj)]^2 + [v\,(zj)]^2}$
Vertical wind shear between levels zi and zj(s^{-1})	$vws\,(zi) = \sqrt{[u\,(zi) - u\,(zj)]^2 + [v\,(zi) - v\,(zj)]^2}$
Potential temperature (K)	$\theta = T\left(\frac{P}{P_0}\right)^{R_d/C_p}$
Turbulence kinetic energy ($m^2\ s^{-2}$)	$TKE(zi) = 0.5[[u\,(zi)]^2 + [v\,(zi)]^2 + [w\,(zi)]^2]$
Gradient Richardson number	$Ri = \frac{N^2}{\left(\frac{dU}{dz}\right)^2}$
Brown index (s^{-1})	$\Phi_m = \sqrt{0{,}3 \times \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} + f\right)^2 + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right)^2}$
Ellrod index 1 (s^{-1})	$El1 = DEF \times VWS$
Ellrod index 2 (s^{-1})	$El2 = VWS \times (DEF + CVG)$
Ellrod index 3 (s^{-1})	$El3 = El1 + DVT$

[1] u,v,w wind components; ws horizontal wind speed; zi, zj levels; T temperature; P0=1000hPa; Rd/Cp thermodynamic constants; N^2 Brunt–Väisälä frequency; f Coriolis parameter; DEF deformation; CVG convergence; VWS vertical wind shear; DVT divergence tendency.

 doi:10.20944/preprints202602.1615.v1

VRTG data were used to identify turbulence events within the study region (Figure 3). Events were classified according to operational intensity thresholds (Class 1–3), and severe CAT (Class 3) cases were retained as the positive class. The distinction between severe CAT and non-CAT conditions was confirmed through multi-source screening, as described in Section 2.2.

2.3. Feature Selection

The selection of predictor attributes (inputs) is a critical step in the development of machine-learning-based predictive models, as it directly affects model accuracy, computational efficiency, and generalization capability. In this study, the objective is to discriminate between time slots with and without the occurrence of severe clear-air turbulence (CAT), as defined by aircraft-reported vertical acceleration (VRTG) and independent observational screening.

The predictor attributes were derived from the GFS025 model and consist of twelve three-dimensional meteorological variables (i.e., u, v, w, θ, VWS, TKE, Ri, Brown, Ell1, Ell2, and Ell3), distributed over a spatial grid of 25 latitude points by 25 longitude points and across nine vertical levels. This configuration results in a total of 67,500 potential predictors, characterizing a high-dimensional feature space.

Given this dimensionality, attribute selection is essential to reduce computational cost, eliminate redundant or weakly informative predictors, and enhance model performance. This process, commonly referred to as dimensionality reduction, aims to identify relevant patterns and extract meaningful information from large datasets while preserving the physical representativeness of the atmospheric fields [18,19]. Simplified representations have been shown to improve learning efficiency and robustness in complex classification problems [20].

In this study, statistical hypothesis testing was adopted as the primary method for supervised attribute selection. Predictor relevance was evaluated directly with respect to the target classification (CAT versus non-CAT), consistent with the supervised learning framework employed. The statistical significance of each candidate attribute was assessed using the Student's *t*-test [21–23], which compares the mean values of the predictor distributions between severe CAT events (Class 3) and non-CAT cases.

The *t*-statistic is defined as:

$$t = \frac{X1-X2}{\sqrt{\left(\frac{s_1^2}{n_1}\right)+\left(\frac{s_2^2}{n_2}\right)}}, \quad (1)$$

where: x_1 e x_2 are means of sample 1 e 2, S_1 e S_2 are the variances of samples 1 e 2, and n_1 e n_2 are the sample sizes. The *p*-value (*p*) is calculated as the proportion of permutations where the absolute t-value exceeds the observed value, using 10,000 random permutations (Np) at six different significance levels of α = 0.01, 0.005, 0.001, 10^{-4}, 5×10^{-5}, and 10^{-5}:

$$p = \frac{number\ of\ permutations\ with\ |t|\geq|t_{observed}|}{1+N_p}, \quad (2)$$

The null hypothesis (H_0) assumes no statistically significant difference between the predictor distributions for severe CAT and non-CAT events. A *p*-value lower than the selected significance level provides sufficient evidence to reject H_0, indicating that the predictor contributes to discriminating between the two classes.

To operationalize the attribute selection procedure, the BRB-Array Tools software [21–23], originally developed by the U.S. National Cancer Institute for large-scale genomic analyses, was adapted for meteorological applications. In this framework, originally developed for genomic data, here adapted to meteorological predictors, and disease outcomes were replaced by CAT occurrence. This adaptation enables efficient processing of high-dimensional atmospheric datasets while applying statistical controls.

Given the large number of simultaneous hypothesis tests, the FDR method was applied to control the proportion of false positives among statistically significant results [24]. The FDR is defined as FDR = V / R, where *V* represents the number of false discoveries of the null hypothesis (H_0) and *R*

 doi:10.20944/preprints202602.1615.v1

is the total discoveries (rejections). By limiting the expected proportion of false discoveries, FDR control ensures a balanced trade-off between statistical rigor and sensitivity in large-scale feature selection problems.

It is acknowledged that the predictor attributes exhibit spatial correlation due to the structured and dynamically coherent nature of atmospheric fields. Neighboring grid-point variables are therefore not statistically independent, as they are influenced by shared synoptic-scale and mesoscale processes. While the False Discovery Rate (FDR) procedure controls the expected proportion of false discoveries under multiple hypothesis testing, it does not explicitly model spatial autocorrelation among predictors. Consequently, the retained attributes should be interpreted as statistically discriminative with respect to the target variable rather than statistically independent realizations. The selection framework aims to identify physically and statistically relevant predictors within a correlated spatial field, rather than to establish independence among atmospheric variables.

In this study, the FDR procedure is applied exclusively within the statistical attribute selection stage, with the sole objective of identifying predictors that exhibit robust and statistically significant differences between CAT and non-CAT samples. Importantly, the FDR does not evaluate classification performance and does not influence the training or assessment of the supervised machine learning models, which are addressed in a subsequent stage of the analysis.

To prevent information leakage, feature selection was performed independently within each training fold during the 5-fold cross-validation procedure. For each fold, statistical hypothesis testing and FDR control were applied exclusively to the training subset, and the selected predictors were subsequently used to train and validate the model within that fold.

2.4. Machine Learning Models

The dataset used for training and testing the supervised learning models consists of hourly records of the selected predictor attributes concatenated with a binary target variable (“yes” or “no”). Severe CAT cases (TARGET = yes) are defined exclusively based on aircraft-reported vertical acceleration, with Class 3 VRTG events representing the positive class. For each severe event, the corresponding GFS predictor fields are extracted from the two time slots closest to the occurrence. Non-CAT samples (TARGET = no) correspond to cruise-level time slots in which no turbulence exceeding the defined operational threshold was recorded by the instrumented aircraft (VRTG = 0). This classification reflects the absence of detected turbulence along the sampled flight trajectory rather than a definitive absence of atmospheric instability and should therefore be interpreted within the inherent spatial and observational limitations of aircraft-based turbulence measurements. The balanced case–control design was adopted exclusively for discriminative modeling and does not represent operational CAT prevalence.

All severe CAT cases (N = 85; Class 3 VRTG) were retained for the supervised modeling stage. To construct a balanced case–control dataset, an equal number of non-CAT samples (N = 85) was randomly selected from the negative class, resulting in 170 labeled instances. Stratified 5-fold cross-validation was adopted, maintaining class balance within each fold (≈ 17 severe cases per test fold). Although the number of severe events is constrained by the observational record, the stratified cross-validation framework ensures that each severe case is evaluated in out-of-sample conditions exactly once, reducing sensitivity to single-split variability and limiting overfitting risk.

The training process included hyperparameter optimization using a grid search approach. Different hyperparameter configurations were systematically evaluated for each algorithm to identify the optimal settings that maximize prediction accuracy and model robustness. This approach ensures that the machine learning models are fine-tuned and capable of effectively distinguishing between severe CAT events (VRTG class 3) and non-CAT conditions.

2.5. Evaluation Metrics

The performance of the developed models is assessed using appropriate metrics. Given the limitations of VRTG data, which do not provide a systematic or comprehensive sampling of the

 doi:10.20944/preprints202602.1615.v1

airspace, traditional scalar statistics such as False Alarm Rate (FAR) and Prediction Bias (BIAS) cannot be reliably calculated [25]. These metrics require consistent and representative sampling, which the restricted and non-uniform nature of VRTG data does not provide. Then, in this study, turbulence prediction is treated as a binary classification problem: the model predicts whether turbulence occurs (positive event) or not (negative event). This approach results in four possible outcomes: True Positive (TP): Correct prediction of turbulence occurrence, False Positive (FP): Incorrect prediction of turbulence occurrence, False Negative (FN): Failure to predict turbulence, and True Negative (TN): Correct prediction of no turbulence.

The model performance is evaluated using the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC). These metrics are widely recognized in turbulence prediction [11,26–30].

The ROC curve is a graphical representation of the performance of a binary classifier as the discrimination threshold varies. It plots two main metrics: True Positive Rate (*PODy*), representing the proportion of correctly predicted positive events, and False Positive Rate (1-*PODn*), representing the proportion of incorrectly predicted negative events given by equations 3 and 4, i.e.,:

$$PODy = \frac{TP}{TP+FN'} \quad (3)$$

$$PODn = \frac{TN}{TN+FP'} \quad (4)$$

where: *TP*: True Positives (correctly predicted turbulence events), *FN*: False Negatives (missed turbulence events), *TN*: True Negatives (correctly predicted no-turbulence events), and *FP*: False Positives (incorrectly predicted turbulence events).

The Area Under the ROC Curve (AUC) is a single scalar value that quantifies the overall performance of a binary classifier. It reflects the model's ability to distinguish between positive and negative classes. An AUC value close to 1 indicates a near-perfect classifier, and an AUC value of 0.5 signifies a random classification, equivalent to guessing [26,27].

The AUC metric provides an objective evaluation of model performance, with higher values indicating improved discrimination between severe CAT and non-CAT events.

To address the strong class imbalance inherent to CAT occurrence datasets, a balanced case–control design was adopted for discriminative modeling, ensuring equal representation of severe CAT and non-CAT samples during model training. Model performance was subsequently assessed using stratified 5-fold cross-validation, preserving class proportions within each fold to maintain statistical consistency and reduce sampling variability.

3. Results

This section provides a detailed discussion of the findings, following the sequence of steps outlined in the methodology. It includes an analysis of VRTG data the process of defining a balanced case–control sample of CAT and non-CAT events, the selection of predictor attributes, and the subsequent training of machine learning algorithms. Each step is critically evaluated to underscore its contribution to understanding the dynamics of Clear-Air Turbulence (CAT) and enhancing prediction accuracy.

3.1. VRTG Analysis

The procedure described in Section 3 was applied to classify CAT and non-CAT events using aircraft-reported vertical acceleration (VRTG) complemented by multi-source observational screening. After applying cruise-level and quality-control filters to the initial 20,338 events identified within the study region, 8,550 events were retained for detailed VRTG analysis. This filtered dataset focuses on cruise-level conditions and forms the basis for the subsequent classification and modeling steps. Of these events, 92% were classified as light (Class 1), 7% as moderate (Class 2), and 1% as severe turbulence (Class 3).

With respect to the vertical distribution, 88% of the recorded events (7,523 occurrences) were observed between flight levels FL180 and FL300, while the remaining 1,027 events occurred above

 doi:10.20944/preprints202602.1615.v1

FL300. This concentration reflects the typical cruise-level range of commercial aviation and is consistent with the altitude band where jet-related shear and deformation processes are most active over the study region.

The monthly variability of VRTG occurrences, shown in Figure 4, reveals a marked seasonal modulation, with increased turbulence records during late spring and summer (November to March). This behavior is consistent with enhanced convective activity and the associated generation of gravity waves, which are known to modulate upper-level flow and favor the occurrence of clear-air turbulence [10,28,29].

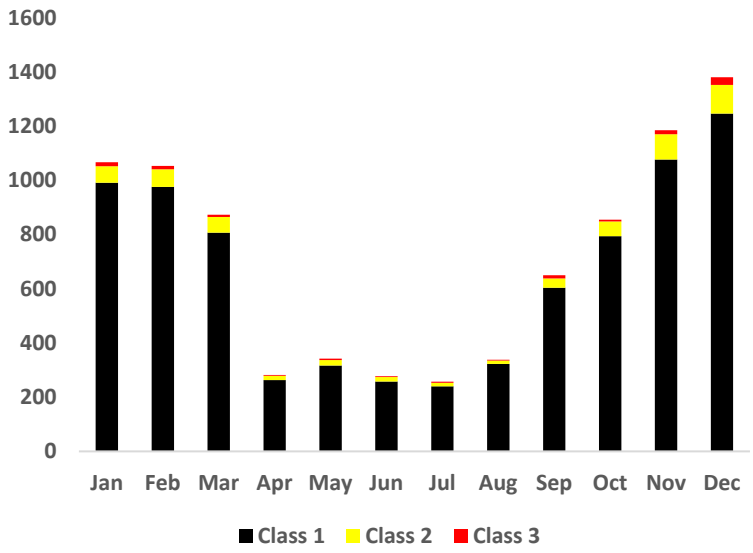


Figure 4. Monthly VRTG variability, peaking in late spring and summer.

Figures 5 and 6 present the hourly distribution of VRTG reports and flight operations, respectively, and are included to characterize the temporal availability of observational data within the analyzed period. These distributions are shown for descriptive purposes only and are not used to infer causal relationships between time of day and turbulence occurrence.

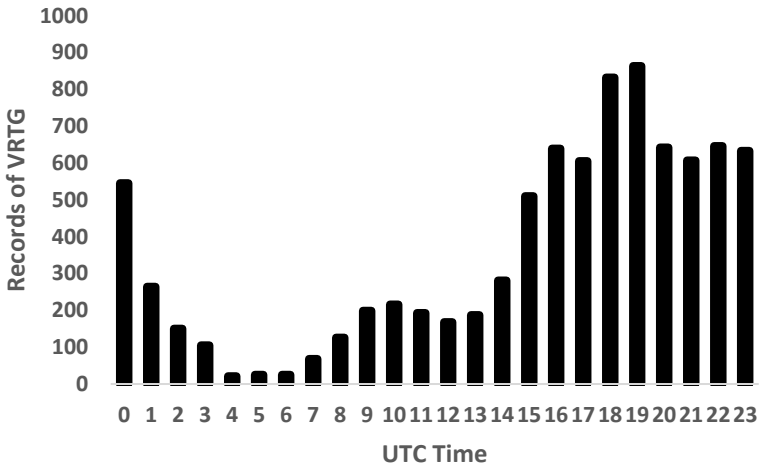


Figure 5. Hourly VRTG distribution with a peak at 19 UTC.

 doi:10.20944/preprints202602.1615.v1

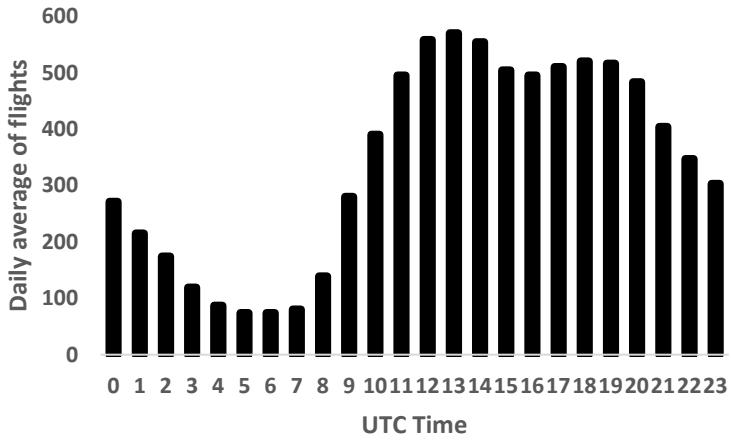


Figure 6. Hourly flight operations, showing reduced activity in early morning hours.

3.2. Selection of Predictor Attributes

The selected predictor attributes were extracted from three-dimensional (3D) meteorological fields defined over a local grid composed of 25 × 25 horizontal points and 9 vertical levels, considering 12 variables forecast by the GFS025 model. This configuration yields 67,500 candidate attributes per event time slot, resulting in an initial dataset comprising approximately 27.1 million candidate attributes over the analyzed period.

Such dimensionality is clearly incompatible with empirical and theoretical constraints widely established in the statistical and machine learning literature, which indicate the need for a minimum ratio on the order of 10:1 to 20:1 between the number of independent samples (or events) and the number of predictor variables in order to ensure statistical stability, generalization capability, and robustness against overfitting [30–32]. Therefore, the adoption of a feature selection strategy was not merely a computational convenience but a methodological requirement. The dimensionality reduction process employed in this study was specifically designed to reconcile the available sample size with established best practices in supervised learning, while preserving physical interpretability of the atmospheric predictors.

Figure 7 illustrates the progressive reduction in the number of retained attributes across different significance levels ($\alpha = 10^{-5}$, 5×10^{-5}, 10^{-4}, 0.001, 0.005, and 0.01), presented on a logarithmic scale. These thresholds correspond to 13, 48, 97, 418, 1,058, and 1,600 retained attributes, respectively. Lower p-values indicate stronger statistical separation between severe CAT events and non-CAT conditions, reflecting a higher discriminatory potential of the corresponding predictors within a supervised classification framework.

High dimensionality in the predictor space increases model complexity, leading to overfitting, longer training times, and reduced generalizability. To mitigate these effects, a conservative significance level of $\alpha = 10^{-5}$ was adopted, resulting in a compact subset of 13 statistically relevant predictors. This choice represents a balance between dimensionality reduction and physical representativeness, ensuring that the retained attributes capture the most robust differences between CAT and non-CAT samples without introducing unnecessary redundancy.

At this stage of the analysis, the objective is strictly limited to statistical relevance and dimensionality reduction. No classification rules, physical thresholds, or performance metrics are derived or evaluated here. The selected predictors constitute the input feature set for the supervised learning experiments described in the subsequent section.

 doi:10.20944/preprints202602.1615.v1

3.2.1. Attribute Selection Results and Analysis

Dimensionality reduction was performed by selecting the most statistically significant attributes based on their p-values obtained from supervised hypothesis testing [21–24]. Table 4 summarizes the significance ranking of the 13 retained predictor attributes, ordered according to increasing p-values and their corresponding False Discovery Rate (FDR) values. These predictors were extracted from three-dimensional meteorological fields produced by the GFS model over the study region and evaluated directly with respect to the CAT versus non-CAT classification.

The selected attributes exhibit p-values ranging from 0.6×10^{-7} to 9.9×10^{-6}, with all associated FDR values remaining below 0.05. This confirms that the likelihood of false positives among the retained predictors is low, supporting the statistical robustness and reliability of the dimensionality reduction process.

The most significant predictors include turbulence-related diagnostics widely recognized in the literature as indicators of dynamically unstable atmospheric conditions. Among them, the Ellrod indices (ELL2 and ELL3), the Brown index, and the vertical velocity component (VVERT) stand out as physically consistent contributors:

- **ELL2 and ELL3 (Ellrod indices):** quantify regions of enhanced horizontal deformation and vertical wind shear, frequently associated with jet streams and gravity-wave activity [9].
- **BROWN:** represents anomalies in vertical wind shear, a fundamental mechanism in the mechanical generation of clear-air turbulence.
- **VVERT:** captures variability in vertical motion, reflecting dynamical instability and energy transfer processes relevant to CAT formation.

The spatial and vertical distribution of the retained predictors is predominantly concentrated between flight levels FL250 and FL350 (approximately 25,000–35,000 ft), with horizontal coordinates mainly located within the core of the study region, bounded by longitudes −48.75°W to −43.0°W and latitudes −21.75°S to −19.0°S. This configuration is physically consistent with the altitude range of cruise-level operations and with the dynamical environment of the mid-to-upper troposphere, where jet-related shear, deformation, and wave activity are most pronounced.

3.2.2. Implications of Attribute Selection

The identification of these statistically significant predictors provides important insights into the atmospheric mechanisms associated with CAT occurrence. The dominance of variables related to wind shear, horizontal deformation, and vertical motion highlights the role of localized dynamical structures in triggering turbulence. These processes are consistent with jet-related shear zones and mesoscale perturbations commonly observed in clear-air turbulence environments. The low p-values confirm the statistical robustness of the selected attributes, supporting their suitability for subsequent use in supervised machine learning models.

In conclusion, the results validate the effectiveness of the p-value-based approach, combined with FDR control, for dimensionality reduction and attribute selection in high-dimensional atmospheric datasets. Under the most conservative setting ($\alpha = 10^{-5}$), only 13 attributes are retained (Figure 7), and Table 4 summarizes the corresponding top predictors ranked by p-values and FDR. Table 4 emphasizes the relevance of ELL2 and BROWN as physically meaningful predictors associated with CAT-favorable flow configurations. This refined attribute set preserves physical interpretability while reducing redundancy, providing a consistent basis for subsequent modeling efforts.

The final attribute set, combined with the target variable (“yes” for CAT events and “no” for non-CAT events), forms the input dataset for training and testing the supervised machine learning algorithms discussed in subsequent sections.

 doi:10.20944/preprints202602.1615.v1

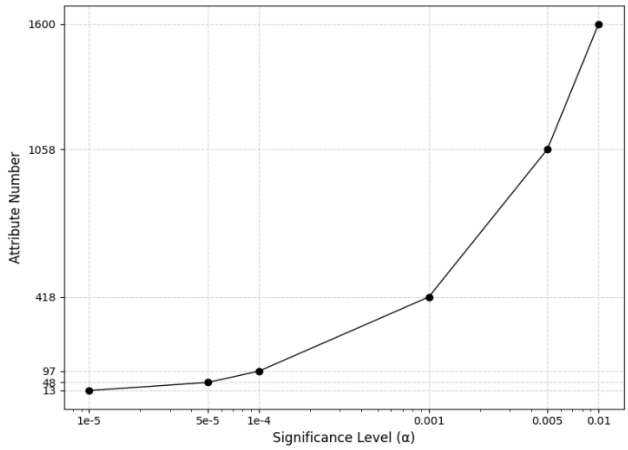


Figure 7. Logarithmic-scale variation in the number of attributes retained under different significance levels (α = 10^{-5}, 5×10^{-5}, 10^{-4}, 0.001, 0.005, and 0.01), resulting in 13, 48, 97, 418, 1,058, and 1,600 attributes, respectively. Lower p-values indicate stronger statistical separation between CAT and non-CAT samples, supporting a more compact and conservative predictor subset for subsequent supervised modeling.

Table 4. Top 13 predictor attributes ranked by p-values and false discovery rate (FDR) at significance level of α = 10^{-5}.

Order of Significance	p-Value	FDR	Attribute (Variable_Coordinates_Altitude)
1	6.00E-08	0.003	ELL2_-45.5_-19.75_300
2	1.00E-07	0.003	BROWN_-48.0_-19.25_300
3	7.00E-07	0.0135	ELL2_-44.0_-20.75_300
4	9.00E-07	0.0135	BROWN_-48.0_-19.0_300
5	1.20E-06	0.0144	ELL2_-46.0_-19.5_300
6	2.00E-06	0.02	BROWN_-47.75_-19.25_300
7	5.40E-06	0.036	ELL2_-47.25_-21.5_250
8	6.00E-06	0.036	VVERT_-48.75_-21.75_350
9	6.60E-06	0.036	ELL3_-46.25_-19.5_300
10	6.80E-06	0.036	BROWN_-47.75_-19.0_300
11	7.10E-06	0.036	BROWN_-47.75_-21.0_250
12	7.20E-06	0.036	ELL2_-44.75_-20.25_300
13	9.90E-06	0.0457	ELL2_-46.25_-19.25_300

3.3. Training and Testing of Machine Learning Algorithms

To ensure reliable generalization and minimize the risk of overfitting, the machine learning (ML) workflow in this study was designed to balance model complexity and validation rigor. Nine algorithms were evaluated the predictive performance of nine algorithms—Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, AdaBoost, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes, and Multi-Layer Perceptron (MLP)—to classify turbulence events with varying degrees of severity.

A critical step to prevent overfitting was the application of 5-fold cross-validation during model training. In this strategy, the training dataset was partitioned into five equally sized folds, and for each combination of predictors (ranging from 1 to 13), the model was trained on four folds and validated on the remaining one. This process was repeated five times per configuration, ensuring that each data point was used for both training and validation. This technique reduces the risk of a model becoming too tailored to a specific training subset and promotes a more robust estimation of performance on unseen data.

 doi:10.20944/preprints202602.1615.v1

To further reinforce generalization and prevent model overfitting, grid search hyperparameter tuning was conducted within each cross-validation cycle. For each algorithm, multiple combinations of hyperparameters were systematically tested, as detailed in Table 5. This ensured that models were not only fit to the data but also optimized with the best ML hyperparameters that generalized well across folds. For instance, Random Forest and Gradient Boosting models underwent extensive search over multiple depth and ensemble size parameters, while simpler models like Naive Bayes required minimal tuning.

Hyperparameter optimization and performance estimation were conducted exclusively within the 5-fold cross-validation framework. AUC values are reported as mean ± standard deviation across the five validation folds, providing an empirical estimate of performance variability under out-of-sample conditions. This procedure provides a robust estimate of generalization performance while reducing the risk of overfitting.

Model performance was primarily assessed using the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) metric. These tools offer a threshold-independent evaluation of a classifier's ability to distinguish between severe CAT events (Class 3, TARGET = yes) and screened non-CAT samples (VRTG = 0, TARGET = no).

As summarized in Table 5, the total number of training iterations resulted from the combination of predictors, cross-validation folds, and hyperparameter grids.

Table 5. Hyperparameter search space, number of combinations, and total training iterations using 5-fold cross-validation (k = 1–13) for each machine learning algorithm.

Algorithm	Hyperparameters	Hyperparameter Combinations	Training Iterations (5-Fold, k=1 to k=13)
Logistic Regression **[30]**	C: [0.01, 0.1, 1, 10, 100], solver: ['liblinear', 'saga'], max_iter: 2000	10	50×13=650
Random Forest **[27]**	n_estimators: [50,100,200,300], max_depth: [None, 10, 20, 30, 40]	20	100×13=1300
Decision Tree **[34]**	max_depth: [None, 10, 20, 30], min_samples_split: [2,10,20]	12	60×13=780
Gradient Boosting **[35]**	n_estimators: [50,100,200], learning_rate: [0.01, 0.1, 0.2]	9	45×13=585
AdaBoost **[26]**	n_estimators: [50,100,200], learning_rate: [0.01, 0.1, 1], algorithm: 'SAMME'	9	45×13=585
K-Nearest Neighbors **[36]**	n_neighbors: [3,5,7], weights: ['uniform', 'distance']	6	30×13=390
Support Vector Machine **[37]**	C: [0.1, 1, 10], kernel: ['linear', 'rbf'], probability: True	6	30×13=390
Naive Bayes **[29]**	None	1	5×13=65
Multi-Layer Perceptron **[28]**	hidden_layer_sizes: [(10,), (20,), (30,)], activation: ['relu', 'logistic'], max_iter: 2000, solver: 'adam'	6	30×13=390
Total		79	5635

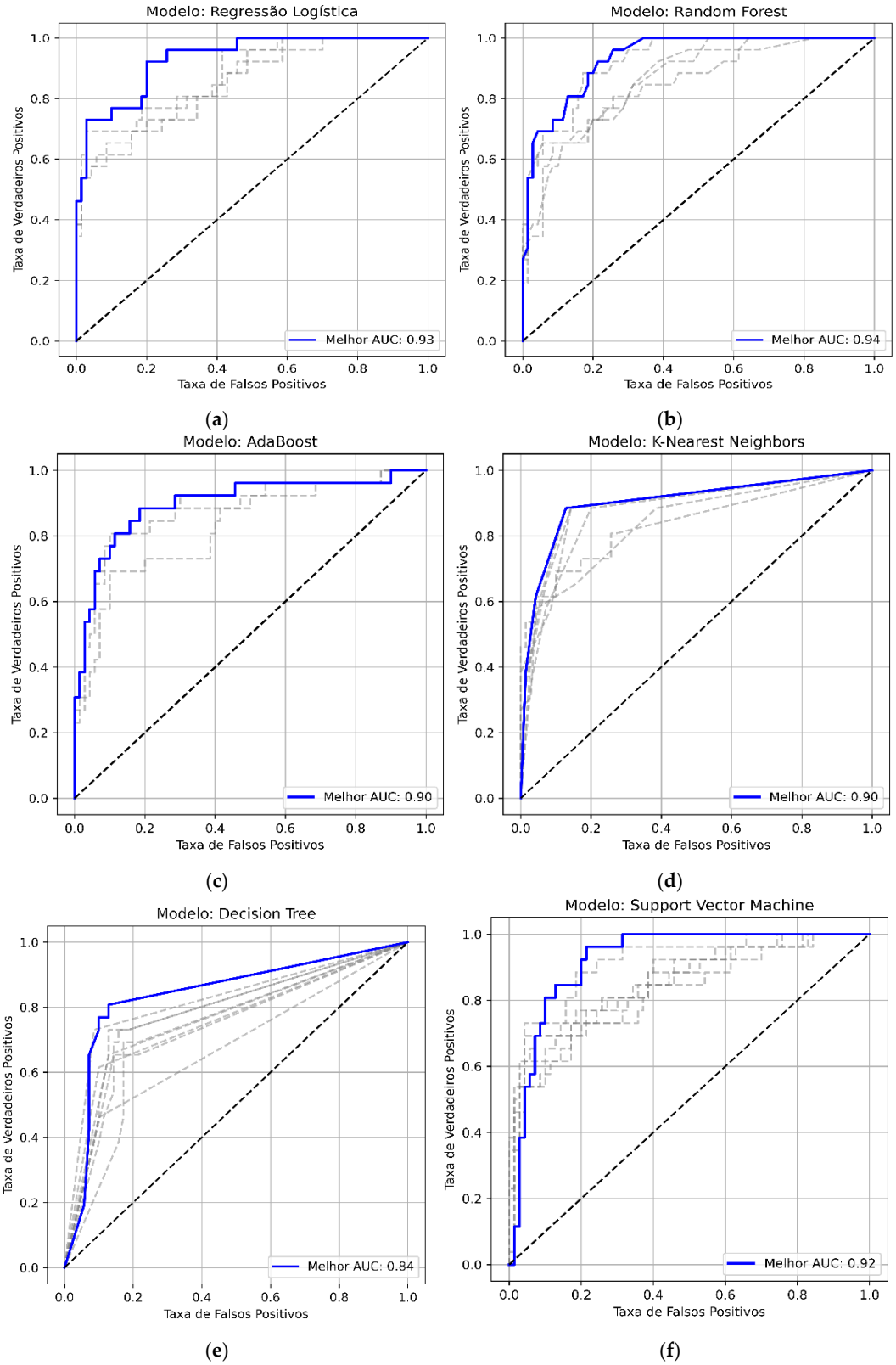
3.3.1. Machine Learning Model Performance

This section presents the performance analysis of machine learning models for predicting CAT events based on ROC curves and their corresponding AUC values. To enhance visual clarity, ROC curves generated for less representative values of k folds (k=5) are plotted with thinner dashed gray lines, minimizing visual noise. A diagonal reference line representing random model performance is included for comparison. The ROC curve with the highest AUC is highlighted using a solid blue line. Figure 8 (a)–(h) presents the ROC curves for Logistic Regression, Random Forest, AdaBoost, K-

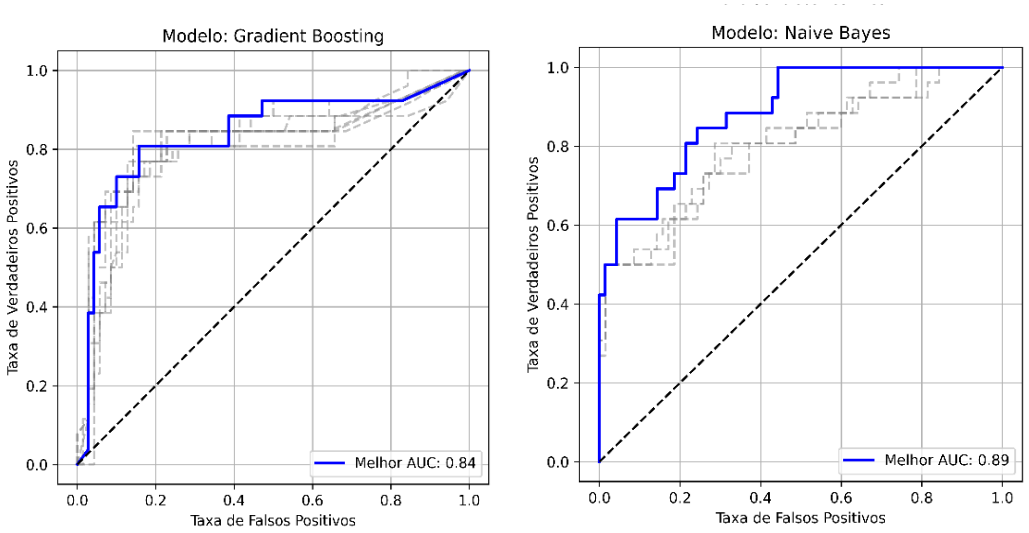
 doi:10.20944/preprints202602.1615.v1

Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), Gradient Boosting, and Naive Bayes.

Standard deviations for lower-performing models remained within ±0.05 and did not alter the relative ranking among classifiers.



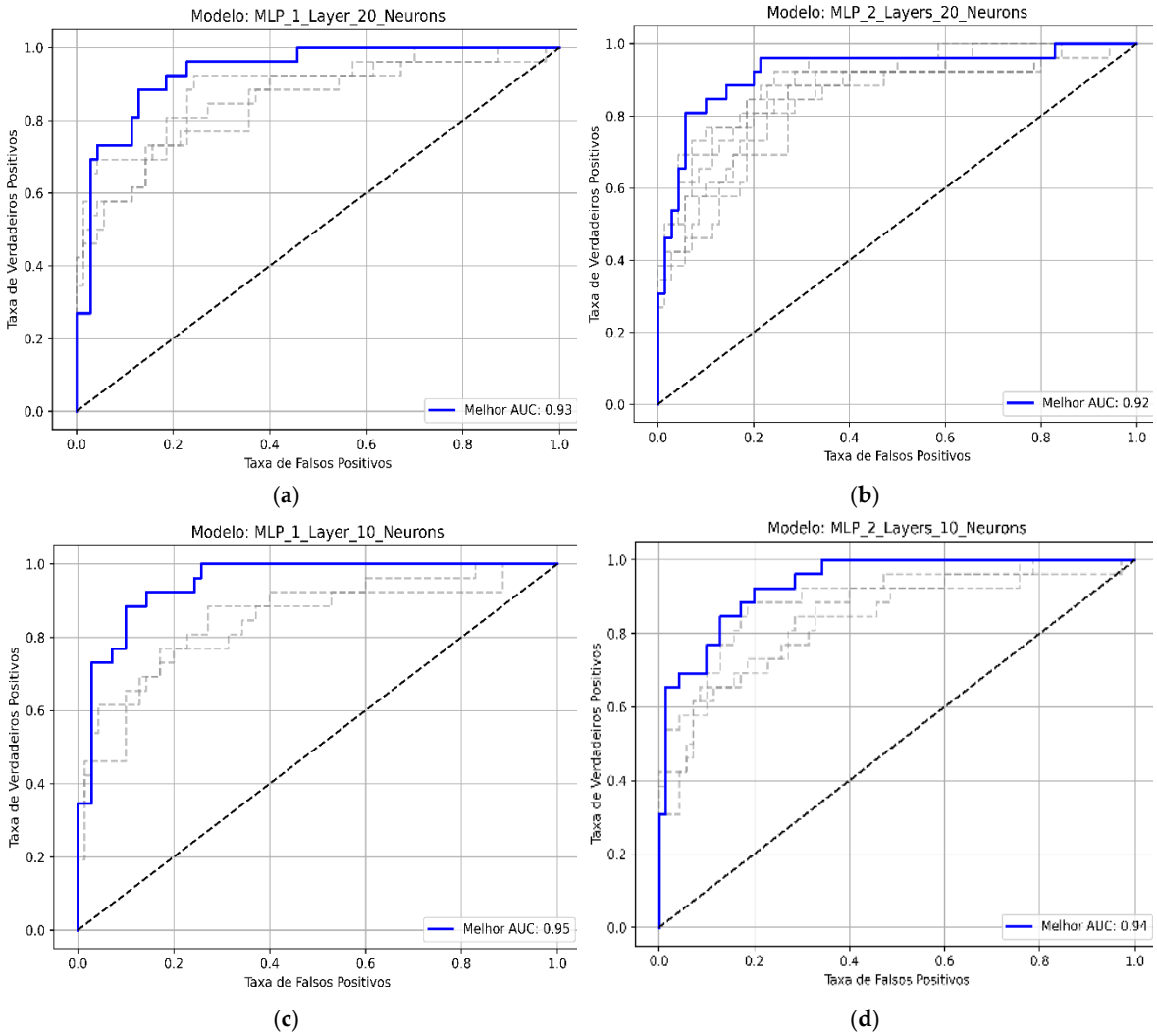
 doi:10.20944/preprints202602.1615.v1



(**g**) (**h**)

Figure 8. ROC curve for (a) Logistic Regression; (b) Random Forest; (c) AdaBoost; (d) K-Nearest Neighbors (KNN); (e) Decision Tree; (f) Support Vector Machine (SVM); (g) Gradient Boosting; (h) Naive Bayes. Curves correspond to 5-fold cross-validation, with the diagonal line indicating random performance.

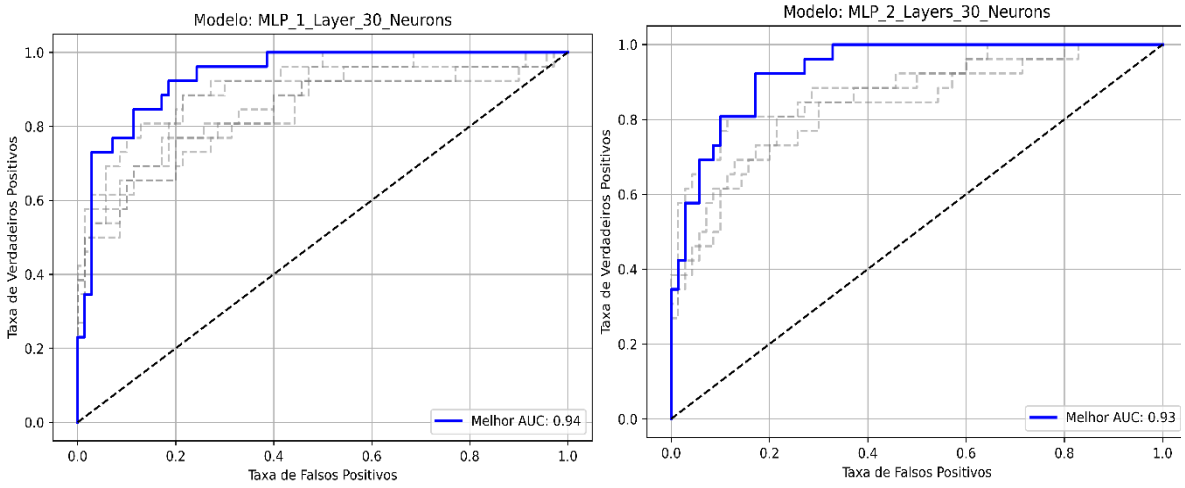
Figure 9 (a)–(f) shows the ROC curves for Multi-Layer Perceptron (MLP) models with one and two layers and 10, 20, and 30 neurons, respectively.



(**a**) (**b**)

(**c**) (**d**)

 doi:10.20944/preprints202602.1615.v1



(**e**) (**f**)

Figure 9. MLP with (a) one hidden layer and 10 neurons; (b) one hidden layer and 20 neurons; (c) one hidden layer and 30 neurons; (d) two hidden layers and 10 neurons; (e) two hidden layers and 20 neurons; (f) two hidden layers and 30 neurons. ROC curves are shown for 5-fold cross-validation, with the diagonal line representing random performance.

Multi-Layer Perceptron (MLP): The model with one hidden layer and 10 neurons achieved the highest AUC of 0.95 ± 0.03, showcasing its superior capability in distinguishing between classes. Simpler MLP architectures (one layer) consistently outperformed more complex configurations (two layers), indicating that additional layers or neurons do not necessarily improve performance. For instance, MLP with one layer and 10 neurons outperformed MLP with two layers and 20 neurons (AUC: 0.95 vs. 0.92), suggesting potential overfitting or reduced generalization in more complex models. Therefore, these values should be interpreted as relative indicators of discriminative capability within a controlled experimental framework, rather than as absolute operational performance metrics.

Tree-Based Models: Among tree-based algorithms, Random Forest demonstrated good performance with an AUC of 0.94 ± 0.04, comparable to the top-performing MLP models. Conversely, the Decision Tree achieved a lower AUC of 0.84, emphasizing its limitations as a standalone model. Boosting methods like AdaBoost and Gradient Boosting delivered moderate AUC scores of 0.90 and 0.84, respectively. While boosting methods showed improvements over single Decision Trees, they did not match the efficacy of Random Forest or MLP in this dataset.

Linear and simple models also exhibited competitive discriminative performance within the adopted experimental design. Logistic Regression achieved an AUC of 0.93 ± 0.04, indicating strong class separation under linear decision boundaries. Naive Bayes reached an AUC of 0.89, providing a useful reference level of performance among the evaluated classifiers.

Given that the dataset was constructed using a balanced case–control approach (1:1), the reported AUC values should be interpreted as a measure of the models' relative discriminative capacity rather than an absolute indicator of operational performance. This distinction is crucial, as artificial class balancing may yield higher discriminative metrics compared to real-world operational scenarios where CAT events are significantly rarer.

A key limitation of this study lies in the use of GFS 0.25° data, which does not explicitly resolve sub--grid processes such as fine--scale vertical wind shear, gravity--wave breaking, or inertial instability. These mechanisms are well known to trigger CAT but remain under--represented at global NWP resolutions. While the machine--learning framework enhanced predictive skill, it cannot overcome the inherent resolution limits of the input data. Future work should incorporate higher--resolution regional models (e.g., WRF at ~3 km) to better capture small--scale turbulence--generating processes.

 doi:10.20944/preprints202602.1615.v1

Although VRTG is not an aircraft--independent turbulence metric, its use here is restricted to cruise--level segments, which reduces the likelihood of contamination by routine maneuvers and enhances its reliability as a ground truth proxy. Future extensions should incorporate aircraft--independent metrics (e.g., RMSg or EDR) whenever available.

Operational applicability motivated the choice of GFS 0.25° as the primary data source in the Brazilian context, where it is the most widely available dataset for civil aviation. Nevertheless, the approach is designed to be portable to higher--resolution regional models.

Traditional diagnostic indices, such as Ellrod I–II and the Brown index, were used as qualitative references to provide physical context for the analyzed turbulence cases. These indices are widely adopted in operational settings; however, their behavior in the selected events highlights the challenges associated with diagnosing severe clear-air turbulence using single-field diagnostics alone. In contrast, machine--learning models trained on statistically selected features consistently achieved AUC values above 0.90.

The analysis highlights that MLP and Random Forest models achieved strong discriminative performance within the adopted experimental framework. MLP models were able to represent complex decision boundaries, while Random Forest classifiers provided comparable performance across the evaluated cases. Simpler models, such as Logistic Regression and Naive Bayes, also demonstrated competitive results, offering viable alternatives depending on study objectives and resource constraints.

Conversely, boosting methods (e.g., AdaBoost and Gradient Boosting) and more complex MLP configurations showed diminishing returns, emphasizing the importance of balancing model complexity with dataset characteristics. The underperformance of Gradient Boosting suggests the need for further optimization and feature importance analysis. In summary, this study demonstrates that MLP with a single layer and 10 neurons and Random Forest exhibited the highest discriminative performance within the evaluated framework.

4. Discussion

This study introduces a regionally focused hybrid framework for forecasting Clear-Air Turbulence (CAT) in Southeast Brazil by integrating numerical weather prediction (NWP) data from the Global Forecast System (GFS) with advanced machine learning (ML) techniques. While previous studies have explored NWP–ML integration for large-scale or transcontinental air routes, the present work contributes specific regional, methodological, and statistical innovations that enhance both its scientific and operational value.

A key contribution of this study is the statistically grounded feature selection process tailored for CAT forecasting. From an initial set of over 67,000 attributes extracted from GFS025 outputs—distributed in a three-dimensional spatial grid—the application of a p-value-based filter with False Discovery Rate (FDR) control enabled the selection of 13 significant predictors. Among these, the Ellrod index (ELL2), and Brown's index, stood out for their physical interpretability and relevance to turbulence-related processes such as wind shear, jet stream dynamics, and wave propagation. This statistically driven dimensionality reduction enhanced both the interpretability and performance of the ML models while minimizing overfitting risks. To our knowledge, this represents an original application of bioinformatics-inspired statistical filtering within clear air turbulence prediction.

In addition, the study introduces a event classification framework by leveraging VRTG data from Airbus A320 aircraft, operationally provided by LATAM Airlines. These turbulence measurements, sampled at hourly intervals and covering four years, allowed for consistent labeling of severe, moderate, and non-turbulent cases. The classification was further refined using auxiliary data sources, including METAR observations, GOES-16 satellite imagery, atmospheric soundings, and synoptic charts. This layered validation process ensured greater confidence in the training labels, which is often a limitation in turbulence modeling studies.

The model evaluation process demonstrated that, when robust feature selection is applied, even relatively simple ML architectures can achieve high predictive performance in rare-event scenarios.

 doi:10.20944/preprints202602.1615.v1

The Multi-Layer Perceptron (MLP) with a single hidden layer of 10 neurons reached an AUC of 0.95, surpassing deeper MLP configurations and most ensemble methods. Random Forest and Logistic Regression also delivered strong results (AUC = 0.94 and 0.93, respectively), reinforcing the general utility of the selected features. These outcomes support the premise that model simplicity, when combined with statistical rigor in attribute selection, can outperform more complex strategies.

AUC values are reported as mean ± standard deviation across the five cross-validation folds. The observed differences among the highest-performing models were modest and within the variability observed across folds.

While the MLP model achieved the highest numerical AUC, the performance differences between the top-tier models—specifically MLP, Random Forest, and Logistic Regression—are subtle. These results should be interpreted with caution, avoiding claims of absolute statistical superiority when differences fall within narrow margins of uncertainty. The high similarity in performance suggests that any of these architectures could be effectively deployed depending on computational constraints.

The regional climatological context added further insight to the modeling effort. Most CAT cases were concentrated between flight levels FL180 and FL300, with a clear seasonal intensification in late spring and summer. These patterns highlight the importance of tailoring CAT detection and forecasting strategies to local atmospheric regimes and flight operations.

To justify the turbulence predictors used in this study, this study emphasizes that, in addition to the classical indices (Ellrod, Brown, and Richardson number), a wide range of dynamic and thermodynamic variables derived from the GFS025 dataset was incorporated, including wind components, vertical wind shear, turbulent kinetic energy (TKE), and deformation-related diagnostics. Although more advanced diagnostics—such as ensemble spread and EDR-based estimates provide valuable insight, they are not directly available from the current dataset or the standard GFS output. While approximations of EDR, such as EDR ≈ $(TKE/\tau)^{(1/3)}$, with τ representing the turbulence dissipation timescale, have been proposed **[8]**, their application to coarse-resolution models like GFS remains uncertain.

An additional limitation concerns the absence of formal statistical comparison between ROC curves (e.g., using the DeLong test). While the primary objective of this study was to evaluate discriminative capability rather than establish statistical superiority among classifiers, future work should incorporate formal AUC comparison procedures to further strengthen model benchmarking.

In summary, this study contributes a replicable and statistically grounded approach to CAT classification that balances physical consistency, methodological transparency, and computational efficiency. These innovations provide a strong foundation for improving operational forecasting tools, particularly in regions with limited turbulence data infrastructure.

5. Conclusions

This study presents a regionally focused hybrid framework for forecasting severe CAT over Southeast Brazil by combining GFS025 model outputs with statistically selected features and a variety of machine learning classifiers. VRTG data from Airbus A320 aircraft operated by LATAM Airlines were used to objectively label turbulence events, which were further verified using auxiliary meteorological datasets, including METARs, radiosonde soundings, satellite imagery, and synoptic charts.

The integration of machine learning with physically grounded predictors proved effective in identifying atmospheric patterns associated with turbulence. The Multi-Layer Perceptron (MLP) with a single hidden layer of 10 neurons and Random Forest achieved the highest discriminative performance within the evaluated framework, highlighting their discriminative power. These findings support the potential of data-driven models to synthesize complex meteorological inputs for operational turbulence forecasting.

Region-specific turbulence behavior was also captured, with a predominance of CAT occurrences between FL180 and FL300 and a seasonal peak during late spring and summer. These

 doi:10.20944/preprints202602.1615.v1

features reinforce the importance of tailoring predictive systems to local climatological and operational conditions.

Dimensionality reduction through *p*-value and FDR-based feature selection played a key role in model efficiency. Among the most relevant predictors were the ELL2 and BROWN indices, which consistently contributed to accurate classification while reducing data complexity.

Among the evaluated algorithms, ensemble methods (e.g., Random Forest) and neural network architectures (e.g., MLP) showed stable and high discriminative performance within the adopted experimental design. Simpler models, such as Logistic Regression and Naive Bayes, provided useful reference points, while more complex configurations, including boosting techniques and deeper MLP architectures, did not yield proportional improvements under the available sample size, highlighting the importance of aligning model complexity with dataset characteristics.

It should be noted that the models were calibrated using data from a single aircraft class (Airbus A320) in cruise flight. Therefore, extrapolating these results to other fleets or different turbulence metrics, such as the Eddy Dissipation Rate (EDR), would require specific recalibration. This acknowledgment ensures scientific transparency regarding the current scope of the predictive framework.

Although the study was focused on Southeast Brazil, the methodology is generalizable to other regions, provided local adaptations are made for atmospheric features, attribute selection, and event characterization. The findings confirm that combining statistical filtering and machine learning with GFS-derived predictors offers a promising direction for improving turbulence risk assessments in operational aviation contexts.

Future work may include the integration of high-resolution mesoscale models (e.g., WRF) to better capture localized turbulence dynamics and enhance the quality of input predictors. Additionally, applying statistical hypothesis testing for model comparisons, expanding the validation using independent datasets from other aircraft fleets, and exploring hybrid or ensemble ML techniques—such as committee machines—could further strengthen the generalizability and robustness of severe CAT forecasting frameworks.

6. Patents

The authors declare that there are no patents resulting from the work reported in this manuscript.

Author Contributions: Conceptualization, A.R. and G.F.; methodology, A.R. and H.R.; software, A.R.; validation, A.R., G.F., H.V., H.R. and I.M.; formal analysis, A.R.; investigation, A.R.; resources, G.F., H.V., H.R. and I.M.; data curation, A.R.; writing—original draft preparation, A.R.; writing—review and editing, A.R., G.F., H.V., H.R. and I.M.; visualization, A.R.; supervision, G.F. and H.V.; project administration, A.R.; funding acquisition, Not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the authors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available data were used in this study. GFS data are available from the NCAR Research Data Archive (RDA) (ds084.1): https://rda.ucar.edu/datasets/ds084.1/. Restrictions apply to the availability of the VRTG dataset from LATAM Airlines due to confidentiality agreements; these data are not publicly available. Derived datasets supporting the findings of this study may be available from the corresponding author upon reasonable request and subject to third-party permissions.
Additional observational data used for event screening are available from DECEA/REDEMET (METAR and synoptic charts) and CPTEC/INPE (GOES-16 imagery), according to each provider's access policies.

 doi:10.20944/preprints202602.1615.v1

Acknowledgments: The authors thank LATAM Airlines for providing the VRTG dataset and for supporting the scientific use of operational flight data for research purposes. The authors also acknowledge NCAR for maintaining the Research Data Archive and DECEA/REDEMET and CPTEC/INPE for providing access to observational datasets used in the event screening process.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CAT	Clear-Air Turbulence
GFS	Global Forecast System
VRTG	Vertical Acceleration of Gravity
ML	Machine Learning
NWP	Numerical Weather Prediction
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
FDR	False Discovery Rate
TKE	Turbulence Kinetic Energy
VWS	Vertical Wind Shear
Ri	Gradient Richardson Number
GOES	Geostationary Operational Environmental Satellite
METAR	Meteorological Aerodrome Report
TEMP	Upper-Air Sounding (Radiosonde)
SACZ	South Atlantic Convergence Zone
WAFS	World Area Forecast System
GTG	Graphical Turbulence Guidance
FL	Flight Level
MLP	Multi-Layer Perceptron
SVM	Support Vector Machine
KNN	K-Nearest Neighbors

References

1. Stull, R.B. *An Introduction to Boundary Layer Meteorology*; Springer: Dordrecht, The Netherlands, 1988.
2. Holton, J.R.; Hakim, G.J. *An Introduction to Dynamic Meteorology*, 5th ed.; Elsevier: Amsterdam, The Netherlands, 2012.
3. Wallace, J.M.; Hobbs, P.V. *Atmospheric Science: An Introductory Survey*, 2nd ed.; Academic Press: Burlington, MA, USA, 2006.
4. Bianchini, R. Aviation Weather Hazards: Understanding the Risks. *Aviation Safety Reports* **2017**, *45*, 120–130. *(Sem DOI/link persistente no material fornecido.)*
5. Sharman, R.D.; Lane, T.P.; Trier, S.B.; Fovell, R.G. Fine-Scale Numerical Simulations of Convective Turbulence. *Mon. Weather Rev.* **2012**, *140*, 2245–2255. https://doi.org/10.1175/MWR-D-11-00342.1
6. Kim, J.; Chun, H.-Y. Numerical Simulation of Convectively Induced Turbulence above Deep Convection. *J. Atmos. Sci.* **2016**, *69*, 2724–2740. https://doi.org/10.1175/JAMC-D-11-0140.1
7. Eick, D. *National Transportation Safety Board Reports on Weather-Related Aviation Injuries*; NTSB: Washington, DC, USA, 2014. *(Sem link público verificável no material fornecido.)*
8. Williams, P.D.; Joshi, M.M. Intensification of Winter Transatlantic Aviation Turbulence in Climate Change Scenarios. *Nat. Clim. Chang.* **2017**, *7*, 137–141. https://doi.org/10.1038/nclimate3085
9. Ellrod, G.; Knapp, D.I. An Objective Clear-Air Turbulence Forecasting Technique: Verification and Operational Use. *Weather Forecast.* **1992**, *7*, 150–165. https://doi.org/10.1175/1520-0434(1992)007%3C0150:AOCATF%3E2.0.CO;2

10. Kim, J.-H.; Chun, H.-Y.; Sharman, R.D.; Keller, T.L. Evaluations of Upper-Level Turbulence Diagnostics Performance Using the Graphical Turbulence Guidance System. *J. Appl. Meteorol. Climatol.* **2011**, *50*, 1936–1951. https://doi.org/10.1175/2011JAMC2649.1
11. Kim, J.-H.; Sharman, R.D.; Strahan, M.; et al. Improvement of Non-Convective Turbulence Forecast for the World Area Forecast System. *Bull. Am. Meteorol. Soc.* **2018**, *99*, 2295–2311. https://doi.org/10.1175/BAMS-D-17-0210.1
12. Muñoz-Esparza, D.; Sharman, R.D.; Deierling, W. Aviation Turbulence Forecasting at Upper Levels with Machine Learning Techniques. *J. Appl. Meteorol. Climatol.* **2020**, *59*, 1883–1899. https://doi.org/10.1175/JAMC-D-20-0116.1
13. Lee, Y.; Kim, S.; Noh, Y.; Kim, J. Deep Learning-Based Summertime Turbulence Intensity Estimation Using Satellite Observations. *J. Atmos. Ocean. Technol.* **2023**, *40*, 1433–1448. https://doi.org/10.1175/JTECH-D-22-0137.1
14. Menegardo-Souza, F.; França, G.B.; Menezes, W.F.; et al. In-Flight Turbulence Forecast Model Based on Machine Learning. *Pure Appl. Geophys.* **2022**, *179*, 2591–2608. https://doi.org/10.1007/s00024-022-03053-5
15. CENIPA. *Annual Aviation Safety Report*; Centro de Investigação e Prevenção de Acidentes Aeronáuticos, 2023. Available online: https://painelsipaer.cenipa.fab.mil.br
16. Richardson, L.F. *Weather Prediction by Numerical Process*; Cambridge University Press: Cambridge, UK, 1922.
17. Brown, R. New Indices to Locate Clear-Air Turbulence. *Meteorol. Mag.* **1973**, *102*, 347–361.
18. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. *AI Mag.* **1996**, *17*, 37–54. https://ojs.aaai.org/index.php/aimagazine/article/view/1230
19. Hair, J.F. *Multivariate Data Analysis*; Prentice Hall: Upper Saddle River, NJ, USA, 1995.
20. Alpaydin, E. *Introduction to Machine Learning*; MIT Press: Cambridge, MA, USA, 2010.
21. Ruivo, H.M.; Sampaio, G.; Ramos, F.M. Knowledge Extraction from Large Climatological Data Sets Using a Genome-Wide Analysis Approach. *Clim. Chang.* **2014**, *124*, 347–361. https://doi.org/10.1007/s10584-014-1066-7
22. Ruivo, H.M.; Campos Velho, H.F.; Sampaio, G.; Ramos, F.M. Analysis of Extreme Precipitation Events Using a Novel Data Mining Approach. *Am. J. Environ. Eng.* **2015**, *5*, 96–105. https://doi.org/10.5923/s.ajee.201501.13
23. Ruivo, H.M.; Campos Velho, H.F.; Ramos, F.M. Data Mining for Flooding Episode in Brazil. *Am. J. Clim. Chang.* **2018**, *7*, 420–430. https://doi.org/10.4236/ajcc.2018.73025
24. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. https://www.jstor.org/stable/2346101
25. Brown, B.G.; Young, G.S. Verification of Icing and Turbulence Forecasts: Why Some Verification Statistics Can't Be Computed Using PIREPs. In *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*; American Meteorological Society: Orlando, FL, USA, 2000; pp. 393–398. (citação bibliográfica conforme registro AMS/RG)
26. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. https://doi.org/10.1006/jcss.1997.1504
27. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324
28. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533–536. https://doi.org/10.1038/323533a0
29. Zhang, H. The Optimality of Naive Bayes. In *Proceedings of AAAI/IAAI*; AAAI Press: San Jose, CA, USA, 2004; pp. 562–567. (versão amplamente circulada) https://www.aaai.org/Papers/AAAI/2004/AAAI04-097.pdf
30. Peduzzi, P.; Concato, J.; Kemper, E.; Holford, T.R.; Feinstein, A.R. A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *J. Clin. Epidemiol.* **1996**, *49*, 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3
31. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009. https://doi.org/10.1007/978-0-387-84858-7

 doi:10.20944/preprints202602.1615.v1

32. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013. https://doi.org/10.1007/978-1-4614-6849-3
33. Sharman, R.D.; Tebaldi, C.; Wiener, G.; Wolff, J. An Integrated Approach to Mid- and Upper-Level Turbulence Forecasting. *Weather Forecast.* **2006**, *21*, 268–287. https://doi.org/10.1175/WAF924.1
34. Quinlan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. https://doi.org/10.1007/BF00116251
35. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. https://doi.org/10.1214/aos/1013203451
36. Cover, T.M.; Hart, P.E. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. https://doi.org/10.1109/TIT.1967.1053964
37. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. https://doi.org/10.1007/BF00994018