ISSN 1648-7788 / eISSN 1822-4180

VILNIUS TECH
Vilnius Gediminas Technical University

AVIATION

2025
Volume 29
Issue 4
Pages 210–221
https://doi.org/10.3846/aviation.2025.25334

AI FRAMEWORK FOR AUTOMATED TERMINAL AERODROME FORECASTING

David SLÁDEK

Department of Military Geography and Meteorology, University of Defence, Kounicova 65, Brno, Czechia

Article History:
- received 10 July 2025
- accepted 5 November 2025

Abstract. Accurate Terminal Aerodrome Forecasts (TAFs) are essential for aviation safety and operational efficiency worldwide. This study develops an AI framework for automated TAF generation, including data preprocessing, model development, and evaluation. Using GFS and ECMWF datasets from 2020–2023 and real TAF forecasts from Brno International Airport the study explores the effectiveness of ML approaches for wind speed and visibility prediction. Principal Component Analysis (PCA) efficiently reduced dimensionality for wind speed predictors but proved less effective for visibility, highlighting its complex nature. Feature importance analysis identified initial observations and seasonal patterns as dominant predictors, underscoring the influence of data quality. Regression models for wind speed met ICAO standards. While Gradient Boosting (GB) classification outperformed human forecasts in raw accuracy, it suffered from poor probability calibration due to dataset imbalance. A critical evaluation of accuracy metrics – such as log-loss and F1-score – revealed their advantages and limitations, particularly in handling imbalanced datasets and probabilistic forecasting. Beyond its empirical findings, the study provides a theoretical foundation for integrating machine learning (ML) into TAF generation, discussing methodological considerations and the interaction between model performance and forecast interpretability. Future research is recommended to focus on the local models, explore advanced models, and expand the framework to diverse climatic conditions.

Keywords: accuracy metrics, aviation forecasts, machine learning forecasting, terminal forecast, TAF, weather forecasting.

Corresponding author. E-mail: *david.sladek@unob.cz*

1. Introduction

Terminal Aerodrome Forecasts (TAFs) are standardized aviation forecasts issued globally under ICAO regulations, detailed in Annex 3 (International Civil Aviation Organization [ICAO], 2016). These forecasts, typically valid for 24 to 30 hours (Long TAF) or shorter periods like 9 hours, provide critical predictions on wind speed and direction, visibility, weather phenomena, cloud base and coverage, significant convective clouds, wind gusts, and temperature extremes. Structured into groups – including a main group for prevailing conditions and change groups for significant weather changes (Table 1) – TAFs adhere to ICAO Annex 3's specifications, which define the TEMPO group and criteria for group inclusion.

The complex nature of TAFs, encompassing probabilistic, deterministic, categorical, and continuous data, poses ongoing challenges in quality assessment. This assessment involves evaluating individual TAF accuracy, forecasting skill, and regulatory compliance, as well as comparing TAF performance across locations within regions like a Flight Information Region (FIR).

Previous research has explored how numerical weather prediction (NWP) models contribute to weather forecast issuance. Wong et al. (2013) examined the implementation of a fine-scale NWP system at Hong Kong International Airport, which generates hourly updated forecasts. While their study focused primarily on wind shear within a 9-hour validity period, such high-resolution models demonstrate the potential to enhance TAF accuracy. Similarly, Jacobs and Maat (2005) highlighted how TAFs can be improved through a combination of NWP models and statistical or physical post-processing techniques, allowing for greater accuracy in aviation-related meteorological parameters. This approach facilitates the automatic generation of TAFs and reduces the need for manual adjustments.

Table 1. Sample TAF groups with explanation

Group	Code	Meaning
Header	TAF LKTB 0206/0306	TAF from Brno-Tuřany Airport valid from Day 2, 6:00 UTC to Day 3, 6:00 UTC
Main group	23008KT 3000 RA BKN010	Wind: 230 deg., 8 knots, visibility: 3000 m, rain, broken clouds at 1000 ft
Change group	TEMPO 0212/0218 23018G38KT 1200 SHSN SCT015TCU	Temporarily between 12 and 18 UTC, wind: 18 with gusts of 38 knots, visibility: 1200 m, snow showers, scattered towering cumulus clouds at 1500 feet

Several commercial products have been developed to support TAF issuance, including the guidance systems provided by the German Weather Service (Deutscher Wetterdienst, DWD). The DWD AUTO TAF system applies Model Output Statistics (MOS) to refine numerical weather predictions from the European Centre for Medium-Range Weather Forecasts (ECMWF), incorporating additional inputs such as METAR, SYNOP, lightning observations, and radar measurements. By implementing statistical corrections based on observational data, this system adapts global model outputs to specific aerodromes, with a particular emphasis on airports that maintain a continuous and reliable observational dataset. Users of this product can access either tabular data with multiple guiding parameters or a familiar TAF-coded output. Similarly, airports without a dedicated TAF issuance can utilize the Localized Aviation MOS Program (LAMP), which provides accurate forecasts for both visual and instrument flight rules (Boyd & Guinn, 2021). The use of first-guess TAFs, which leverage NWP model data to generate simple and readable forecasts, further enhances efficiency by minimizing the need for meteorologist intervention and allowing experts to focus on refining model outputs (Lanyon et al., 2020).

Expert systems offer an automated approach to assess the accuracy and reliability of Terminal Aerodrome Forecasts (TAFs), ensuring adherence to ICAO standards and identifying discrepancies to improve forecast quality. For instance, expert systems are integral to decision-support frameworks like the Automated Decision Tool for Operations Support (ADTOS). ADTOS leverages a data warehouse of aviation information, including TAFs and METARs, to aid traffic flow specialists in optimizing arrival and departure strategies (Ayhan et al., 2013). Furthermore, advancements in machine learning enhance TAF-based predictions. Altinok et al. (2018) demonstrated this by modelling weather and traffic demand data to predict runway configurations, enabling real-time forecast adjustments.

Given the mandatory and continuous nature of TAF issuance, it generates substantial amounts of data that can be leveraged for performance-based evaluations aligned with ICAO standards. Simone et al. (2022) noted that traditional forecast verification approaches rely on accuracy indicators to assess forecast reliability, assisting decision-makers in evaluating past performance. More recent methodologies incorporate machine learning techniques to enhance anomaly detection in historical weather data, utilizing past bulletins and previous forecasts to identify potential inaccuracies (Patriarca et al., 2023). Techniques such as anomaly detection and hierarchical clustering facilitate the calculation of an error propensity metric, offering insights into the likelihood of forecasting errors. These advancements contribute to improved decision-making processes in aerodrome weather management by identifying critical areas of forecast inaccuracy.

Enhancing TAF reliability hinges on robust forecast verification. Novotný et al. (2021) investigated verification methods designed for consistent application across diverse airports. Similarly, Anggoro et al. (2019) highlighted the necessity of systematic verification procedures to minimize forecasting errors and improve operational preparedness. While direct TAF prediction is crucial, related prediction systems can also enhance airport operations. For example, Buxi and Hansen (2013) and Kicinger et al. (2016) demonstrated the use of historical and real-time weather data to create probabilistic airport capacity scenarios, enabling better strategic planning. Machine learning further strengthens these predictive capabilities. Dhal et al. (2013) utilized multinomial logistic regression to forecast airport arrival rates. Ultimately, automating processes from TAF generation to runway configuration or airport capacity can improve objectivity of airport management.

Unlike previous meteorological studies that often focus on general forecasting, this research specifically examines the technical and regulatory thresholds of wind speed and visibility within Terminal Aerodrome Forecasts (TAFs). It further investigates the impact of probabilistic change groups on forecast accuracy and skill, and compares the performance of machine learning (ML) models against human forecasters. In contrast to the DWD product, this study conducts a detailed analysis of accuracy metrics, evaluating their suitability for assessing TAF performance. This research aims to provide a comprehensive comparison of methodologies, preprocessing strategies, and a diagnostic evaluation of the entire TAF generation process. The central research question is:

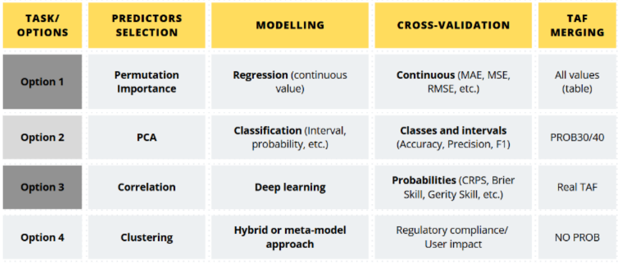
Developing an objective, global automated prediction framework that establishes a baseline for accuracy and skill in TAF generation, serving as a reference for evaluating human forecasters and comparing forecast complexity across airports.

The inherent complexity of TAFs motivates the need to establish a baseline for forecasting accuracy and skill. Previous studies (Novotný et al., 2021; Sládek et al., 2024) have demonstrated this inherent complexity.

Theoretical anticipation of this study is that human forecasters demonstrate the highest prediction skill due to their ability to integrate expert knowledge and contextual understanding, surpassing ML-based outputs. Global models are expected to perform less accurately than local models, with accuracy varying based on model configuration and data quality. ML model accuracy is expected to decrease with longer lead times. However, initially, all models should exhibit similar performance due to their reliance on nowcasting and persistence. While accuracy may be comparable especially for rare phenomena, skill differences are likely to be more pronounced.

The framework proposed in this study aims to serve as a reference standard for forecast evaluation, supporting and, in the future, outperforming human forecasters consistently. The steps involved in creating an automated TAF forecast are outlined in Figure 1.

Figure 1 presents a schematic overview of the automatic TAF forecasting framework, highlighting that at each



TASK/ OPTIONS	PREDICTORS SELECTION	MODELLING	CROSS-VALIDATION	TAF MERGING
Option 1	**Permutation Importance**	**Regression** (continuous value)	**Continuous** (MAE, MSE, RMSE, etc.)	All values (table)
Option 2	**PCA**	**Classification** (Interval, probability, etc.)	**Classes and intervals** (Accuracy, Precision, F1)	PROB30/40
Option 3	**Correlation**	**Deep learning**	**Probabilities** (CRPS, Brier Skill, Gerity Skill, etc.)	Real TAF
Option 4	**Clustering**	**Hybrid or meta-model approach**	Regulatory compliance/ User impact	NO PROB

Figure 1. Schematic overview of the automated TAF forecasting framework, depicting decision options at each step. Any combination of choices, one from each column, is permitted

of its four stages – predictors selection, modelling, cross-validation and TAF merging – different methods can be chosen. These results vary in both accuracy and formal correctness. Formal correctness, which refers to adherence to ICAO TAF coding standards, is primarily influenced by the "TAF Merging" method, as it dictates how predicted values are transformed into the TAF code. However, the first three stages of the framework directly influence the model's performance metrics. These metrics, such as accuracy and skill scores, are dependent on both the quality of the input data and the chosen model architecture:

- The goal of this study is to establish a framework that can help achieve the following objectives:
- Develop a tool for generating first-guess TAFs or supporting products.
- Define the lower bound of human forecasters' performance.
- Identify the most suitable accuracy metrics based on the specific nature of TAFs.

To achieve these goals, an experiment has been designed that follows the selected steps in Figure 1. This experiment presents the effects of each step on real data through a case study.

2. Methods

According to the Figure 1, this chapter expands on each selected method in detail. Specifically, the process was tested using GFS and ECMWF data from 2020–2023 (National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce NCEP, 2015) and real TAF forecasts from Brno-Tuřany international airport (ICAO indicator: LKTB) 2021, which were then compared against METAR reports. A demonstrative application is presented in the case study section.

2.1. Predictors selection

The first step is predictor selection. Forecasters may rely on expert judgment, but to objectify the process, this study comments on the advantages and disadvantages of the four options from the first column of Table 2.

Table 2. Overview of the predictors' selection methods

Method	Pros	Cons
Permutation Importance (Hubbard et al., 2005)	Directly related to model performance Easy to understand and interpret.	Computationally expensive for complex models. May not accurately capture feature importance in highly correlated datasets.
PCA (Jolliffe, 2005)	Reduces dimensionality Can identify underlying patterns	Can be difficult to interpret May not always improve model performance Can lead to information loss.
Correlation (Kenny, 1979)	Simple and easy to understand. Can identify highly correlated features that may be redundant	May miss non-linear relationships between features. Can lead to oversimplification and neglect important interactions.
Clustering (Everitt et al., 1974)	Can identify groups of similar features. Can help selecting representative features from each cluster.	Choice of clustering algorithm and parameters can significantly impact results. Interpretation of clusters can be subjective

The selection of appropriate predictors is critical for the development of an effective automated TAF generation framework. Given the high dimensionality of Numerical Weather Prediction (NWP) data and the complex interactions between meteorological variables, this study employed a two-step approach utilizing Principal Component Analysis (PCA) and Permutation Importance. The selected methods are compared to assess their effectiveness.

2.2. ML models

ML classification models were employed to estimate both the probability and frequency of occurrence within a given time window, while regression models were used to predict the exact values reported in the forecast. Additionally, simple Multi-layer perceptron neural networks were tested for comparison. Apart from other models used solely for benchmarking and not included in the Case Study, the primary models considered are listed in the Table 3.

For training and testing, the data were split into an 80/20 ratio, while ensuring that no inter-day shuffling can occur. This was applied to prevent data from the same day from appearing in both the training and testing sets, thereby avoiding potential data leakage.

Table 3. Classification and regression models used (Chase et al., 2022)

Method	Name	Purpose/ advantages
ML Classification	Logistic Regression	Simple, interpretable, works well with linearly separable data
	Decision Tree	Handles non-linearity, interpretable, prone to overfitting
	Random Forest	Reduces overfitting, handles imbalanced data with class weighting
	Gradient Boosting	Powerful, reduces bias, effective with imbalanced data
	AdaBoost	Focuses on misclassified cases, robust to outliers
	KNN	Non-parametric, works well with small datasets, sensitive to imbalances
ML Regression	Linear Regression	Simple, interpretable, works well as a reference
	Bayesian Ridge	Regularized, prevents overfitting, robust to multicollinearity
	Random Forest	Captures non-linearity, robust, handles missing values
	AdaBoost	Boosts weak learners, handles complex relationships
	Gradient Boosting	Highly accurate, handles non-linear patterns well
	Decision Tree	Non-linear regression, interpretable, risk of overfitting
	K-Nearest Neighbors	Non-parametric, simple, can adapt to data distributions
	Polynomial Regression	Captures non-linearity, prone to overfitting
Deep Learning Classification	Multilayer perceptron	Captures complex relationships, works well with large datasets

2.3. Evaluation

A consistently selected set of metrics was used to evaluate the performance of the prediction. These metrics were divided depending on the nature of the predicted variable, i.e., whether it was a categorical or continuous variable (Table 4). Thus, both easily interpretable and more complex methods were used to determine prediction accuracy, skill, etc.

The selection of appropriate evaluation metrics is crucial for a comprehensive assessment of TAF forecast performance, categorized by the nature of the predicted variable: deterministic continuous, deterministic categorical, and probabilistic categorical. The first two categories, using metrics like MAE and Accuracy, focus on the accuracy of point predictions, while probabilistic metrics such as RPS, Log-loss, and BSS evaluate the reliability and calibration of probability estimates, essential for TAFs. Finally, "Regulatory Compliance" assess adherence to aviation standards, ensuring that generated TAFs are usable and compliant. A high-performance model must balance accurate predictions with regulatory requirements for operational viability, although this trade-off is not defined in the current directive.

Table 4. Selected model performance metrics

Variable	Name	Advantage
Deterministic continuous	MAE	Straightforward
	Max Error	Highlights worst-case errors
Deterministic Categorical	Accuracy	Straightforward
	F1	Objective (Hubbard et al., 2005)
	ROC	Measures model performance across all classification thresholds
	AUC	Quantifies the ability of the model to distinguish between classes (Chase et al., 2022)
Probabilistic categorical	Ranked Probability Score	Suitable for probabilistic multicategory (Ferri et al., 2008)
	Log-loss	Proper for probability, penalizes confident wrong predictions
	Brier Skill Score	Proper for probabilistic multicategory (Mason, 2004)
	Expected calibration error	Measures how well forecast probabilities match observed frequencies (Famiglini et al., 2023)
Regulatory Compliance	TEMPO usage	Evaluates adherence to formal prediction standards and guidelines
	Groups usage	Evaluates compliance with regulations for the use of change groups

2.4. Formal processing options

Since NWP model outputs are available at three- and six-hour intervals, this study opted for a time window approach. This discretization allows for the aggregation of probabilistic and deterministic information within operationally relevant timeframes, facilitating the generation temporally coherent TAFs. Each three-hour time window was characterized by three key attributes:

1. Change from prevailing conditions, denoting a significant deviation from the preceding forecast, essential for determining the need for change group;
2. Probability of occurrence of a specific category within the window. This facilitates assigning of the probability of the group (PROB30, PROB40 or main group);
3. Frequency of occurrence, indicating the proportion of time the category is expected to prevail, facilitating support for TEMPO group involvement;

This method generates a structured table of attributes for each time window. A schematic representation of a possible scenario is provided in the following Table 5.

Thus, the final TAF code for this time window would be: *0106/0109 9999 PROB40 TEMPO 0106/0109 3000*=.

Table 5. Schematic representation of the ML model output for the categorical variable Visibility

Time	Visibility conditions for time window DD06/DD09								
	Conditions 1 (9999)			Conditions 2 (3000)			Conditions 3 (0600)		
	Chan.	Prob.	Freq.	Chan.	Prob.	Freq.	Chan	Prob.	Freq.
06-09	0	80	4/6	1	42	2/6	1	10	5/6

Note: 'D' denotes day.

By its very nature, TAF involves categorical values and identifies the most likely value (ICAO, 2016). Additionally, the process must balance probabilistic, deterministic, and formal requirements, which is made in accordance to national regulations (Sládek et al., 2024). The Table 6 presents the modules that process raw model values into final forecasts, or, in the case of raw outputs, how the system directly delivers predictions to the user. Colors indicate attribute quality: green for desirable, orange for neutral, and red for low quality.

Table 6. Formal modules for translating raw values into coded forecasts.

	Description	Acc.	Redundancy	Formal
Raw outputs	All values in all categories and changes	Max	High	Low
	All predicted categories and probabilities	Max	High	Low
	Most likely or most frequent value in the time window	Med	Low	Med
	All values with probability higher than 35%	Med	Med	Med
	All values with probability higher than 45%	Low	Low	Med
Regulatory Compliant	Only Annex 3 changes	Low	Low	Max
	Color codes changes only	Low	Low	Max
Special	Dangerous values or phenomena only	Low	Very Low	Low
	VMC changes only	Low	Very Low	Low
	Warning thresholds only	Low	Very Low	Low

Table 6 outlines the various processing modules that determine how the ML model's predictions are presented to the user. For instance, the version with complete probabilities and range of cases is visualized in Table 7.

Table 7. Example table output of the ML visibility classification refined by the regression within the category

Category	Value	Probability	Regression value
0	0–800 m	1%	750
1	800–1500 m	30%	1200
2	1500–3000 m	17%	2100
3	3000–5000 m	1%	4000
4	More than 5000 m	51%	9999

Due to ICAO Annex 3 (ICAO, 2016) regulations, which mandate a minimum probability threshold of 30% for inclusion of probabilistic information (PROB groups) in TAFs, the refined visibility forecast from Table 7 is constrained. Consequently, the ICAO_strict module outputs a simplified TAF code:

DDHH 9999 PROB30 1200=.

This highlights the regulatory impact on the presentation of probabilistic forecasts, where detailed model outputs are condensed to meet operational requirements.

Formal processing is the final step, where significant compromises in the determined values may occur due to coding requirements and constraints on the intervals at which changes can be included. Given that historical data indicate low frequency of short term changes crossing the regulatory threshold values, the inclusion of TEMPO groups was deemed less critical for the presented case studies. However, it is acknowledged that TEMPO groups play a vital role in capturing temporary weather phenomena, and their incorporation is considered in future extensions of this research.

3. Case study

For the case study, we used data from Brno-Tuřany International Airport in the Czech Republic. Observations were taken from METAR reports, while predictors came from GFS in 3 hours step and ECMWF in 6 hour step spanning four years (2020–2023).

Following Figure 1, a procedure for wind speed and visibility prediction was designed, consisting of the following steps:

1. Predictor Selection: PCA was applied to reduce the data from surrounding points, with PCA Loadings used to track with important features. Additionally, Permutation Importance and Feature Importance were compared as alternative approaches.
2. Model Comparison: Different regression models were evaluated for wind speed prediction (accuracy), and classification models were assessed for visibility categories.
3. Cross-Validation: Various cross-validation metrics were analyzed to evaluate model performance, highlighting their strengths and weaknesses.
4. TAF Comparison: Finally, the model's accuracy was compared with real TAF forecasts from professional distributions.

3.1. Predictors selection

Following the procedural framework outlined in Figure 1 the initial step in this case study involved the task of predictor selection. Given the high dimensionality of NWP data and the complex nature of visibility prediction, a combination of dimensionality reduction and feature importance analysis was employed. In this section, the application of PCA for dimensionality reduction is presented, specifically for the surrounding four points in case of the GFS model (Figure 2) and the model ECMWF (Figure 3). PCA was then followed by the calculation of the feature importance.

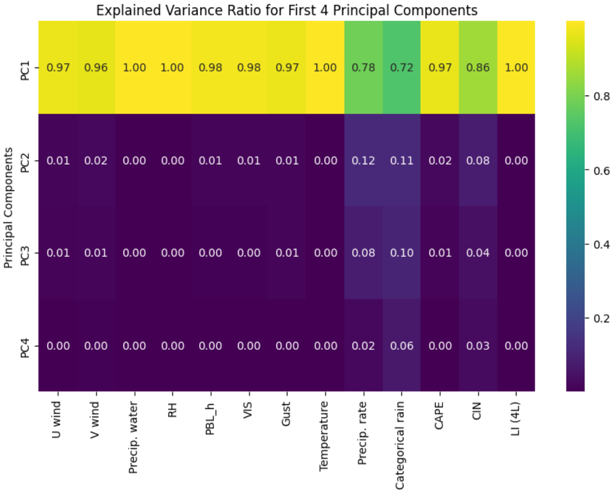


Figure 2. Principal Component Analysis (PCA) applied to GFS visibility predictors, explaining 99% of variance from four surrounding points. Predictors: U and V wind components, precipitable water, relative humidity, planetary boundary layer height, visibility, and best 4-layer lifted index

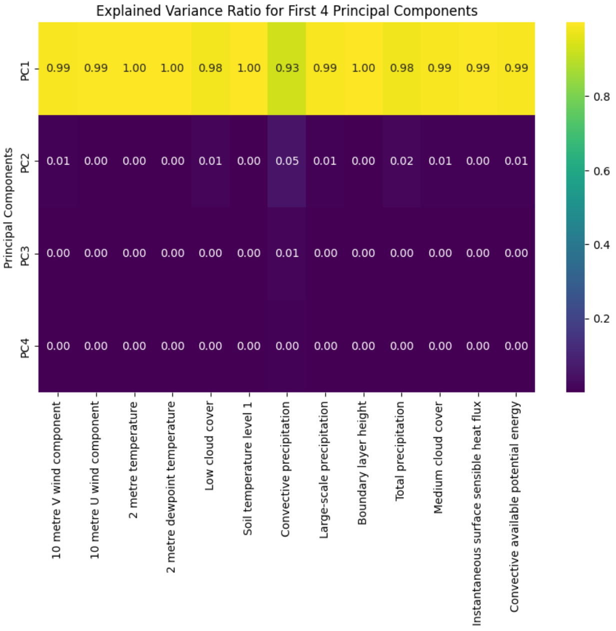


Figure 3. PCA for potential ECMWF visibility predictors explaining 99% of variance from four surrounding points

From the figures (Figure 2, Figure 3), it is evident that PCA effectively reduced data dimensionality from surrounding points near the airport, retaining 99% of the variance. However, instability and precipitation predictors exhibited higher local variability. Furthermore, it can be inferred that higher-resolution models, capturing finer spatial details, may require fewer principal components to explain the same variance, as they provide a more detailed representation of the atmosphere. Nevertheless, with increased resolution, the number of surrounding points used probably needs to be increased. Caution should be exercised when applying PCA to binary or categorical data, such as visibility categories, as its suitability depends on the nature of the features. For example, one hot encoding of categorical predictors before PCA may improve results. Therefore, categorical predictors should ideally be treated separately to ensure a more robust dimensionality reduction. To further understand the contribution of each original predictor to the derived principal components, the loadings – representing the coefficients of the original variables in the principal component space –were visualized (Figure 3).

Based on the results from Figure 2, Figure 3, and Figure 4, PCA can be applied to reduce the dimensionality of surrounding points. The resulting principal components that correspond to the initial predictors are then used as model features. This approach is particularly relevant for visibility prediction, where at least eight principal components are required to explain 99% of the variance. A key drawback of applying PCA twice is the increased difficulty in interpreting the results. Specifically, tracing the influence of individual parameters on the final prediction would become challenging.

An alternative and more intuitive approach involves utilizing Feature Importance, which is inherently available in tree-based methods. This technique directly quantifies the contribution of each predictor to the model's accuracy. However, it is important to distinguish between the built-in Scikit-learn method, which quantifies the Mean Decrease in Impurity, and the more general Permutation Importance, which provides a broader, model-agnostic evaluation (Figure 5).

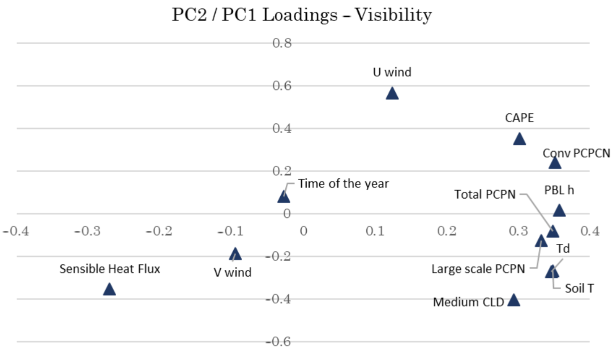


Figure 4. PCA Loadings of first two principal components of ECMWF predictors applied on visibility, explaining 99% variance

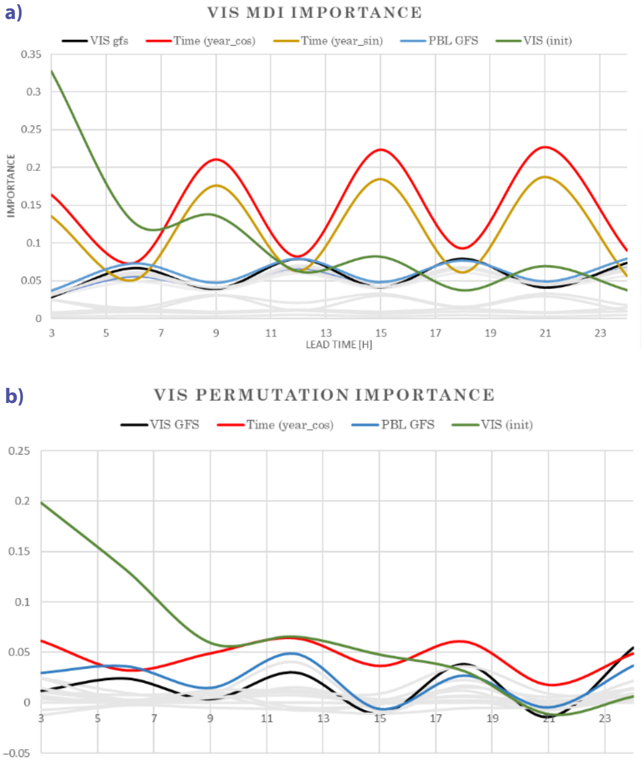


Figure 5. Comparison of: (a) – Mean Decrease in Impurity and (b) – Permutation Importance for Random Forest features predicting visibility categories. Low importance predictors are shown in gray

When evaluating the importance of predictors across lead times, three key considerations can be identified:

1. Decrease in the importance of the initial observation. Due to the persistence of atmospheric conditions, the initially observed visibility class remains the most influential predictor for approximately 6 to 9 hours. This may indicate a significant persistence effect and, consequently, the inherent "complexity" of the weather system.
2. Periodicity of temporal parameters. Features derived from the time of year exhibit a high contribution in cases where data quality is lower. This suggests that when direct forecasts from the model lack sufficient information, the model relies more heavily on temporal patterns as the most relevant source of predictive power rather than solely GFS values.
3. Increasing importance of GFS visibility. Visibility predictions from the GFS model have minimal impact on forecast accuracy until a lead time of approximately 24 hours is reached.

From the comparison of feature importance metrics, two key observations can be made:

1. Consistent feature importance scores between Mean Decrease in Impurity (MDI) and permutation importance suggest that the feature is genuinely influential in the prediction.
2. Higher MDI importance with lower permutation importance may indicate the presence of high-cardinality features or class imbalance, which can skew impurity-based importance measures.
3. Higher permutation importance with lower MDI importance may highlight complex, non-linear relationships between the feature and the target variable, which are not fully captured by impurity-based methods.

A scenario with higher Mean Decrease Impurity (MDI) and lower permutation importance was frequently observed. This suggests that a significant imbalance in the visibility classes likely hinders modelling performance. Consequently, the predictor selection process yielded three key conclusions:

1. Principal Component Analysis (PCA) effectively reduces the dimensionality of Numerical Weather Prediction (NWP) data source points, including wind speed predictors. However, applying PCA to visibility classes results in information loss.
2. Given the anticipated complexity of predicting visibility, Random Forest feature importance was visualized. For the initial 6–9 hours, the model relies heavily on the initial observations. The importance of the time of year, which emerged as the most significant feature, underscores the influence of the data quality.
3. Comparing feature importances reveals significant class imbalance. In this context, the model relies on combined features, such as the Planetary Boundary Layer (PBL), which is predominantly determined by wind speed and temperature (Nielsen-Gammon et al., 2010), while underutilizing parameters like visibility or relative humidity.

A sensitivity analysis, performed using the Spearman Rank Correlation (ρ) between Mean Decrease in Impurity (MDI) and Permutation Importance (PI), confirms that the framework's limitations stem primarily from predictor quality during specific meteorological regimes and lead times. For stable conditions, the model's feature importance demonstrated an unusual alternating pattern of consistency, achieving high agreement (0.93–0.96) at lead times of 6, 12, 18, and 24 hours, but severely diverging ($\rho \approx 0.40$) at hours with a lower data quality (3, 9, 15, and 21 hours). Conversely, analysis restricted to windy conditions (Ws > 15 KT) revealed a breakdown in model interpretability, with ranging from a perfect 1 down to a critical negative correlation of $\rho \approx -0.5$ at the 18-hour lead time. This negative correlation quantitatively demonstrates that the features the model was internally prioritizing (high MDI) actively harmed predictive performance (low PI).

3.2. Models' performance

For wind speed prediction, all models exhibited similar performance trends, albeit with varying accuracy levels (Figure 6). The maximum error fluctuated based on data availability, with lower errors coinciding with ECMWF model output times (0, 6, 12, 18 UTC). Across all models, the Mean Absolute Error (MAE) remained consistently below 2.5 knots.

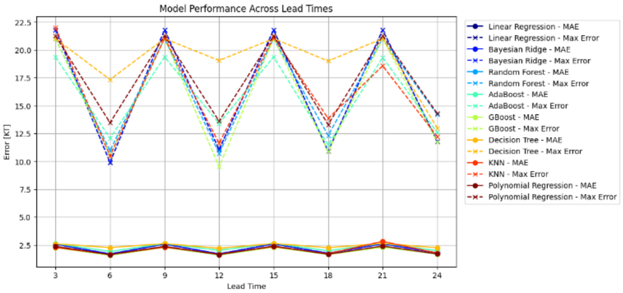


Figure 6. Accuracy metrics (Mean Absolute Error and Maximum Error) of tested regression models (uncalibrated hyperparameters) across 3- to 24-hour lead times for wind speed forecasting

Furthermore, it is significant that all models maintained wind speed accuracy within 5 knots in over 80% of predictions across all lead times, with some exceeding 90%. This performance aligns with ICAO's long-term desirable accuracy criteria for wind forecasts. For subsequent machine learning modelling, Gradient Boosting (GB) was selected due to its superior baseline performance, ability to manage data imbalances through iterative refinement, provision of probabilistic outputs, and suitability for multiclass predictions (Figure 8).

To explore the potential for further performance enhancements, a Multilayer Perceptron (MLP) neural network was optimized. The MLP architecture, a fully connected feedforward network designed for regression, incorporated multiple dense layers with ReLU activation, Batch Normalization, Dropout regularization, and L2 kernel regularization to mitigate overfitting. Training was conducted using the Adam optimizer and Mean Squared Error (MSE) loss, with Mean Absolute Error (MAE) monitored. Early stopping, based on validation loss, was implemented to prevent overfitting and ensure optimal weight retention. For comparative purposes, Gradient Boosting with grid search calibrated hyperparameters was also evaluated, demonstrating very similar performance in wind speed modelling (Figure 7).

We can observe the lowest levels of errors in the 3 hours lead time, where the contribution of the observation is the highest. Then in the 6h lead time, and 12 h, high quality of the forecast results into low errors, whereas in 9 and 21 hour lead times, both MAE and Max Error are the highest due to the usage of solely GFS data. In the Max error curve, the increasing trend is also observable, confirming assumption of the lead time dependence.

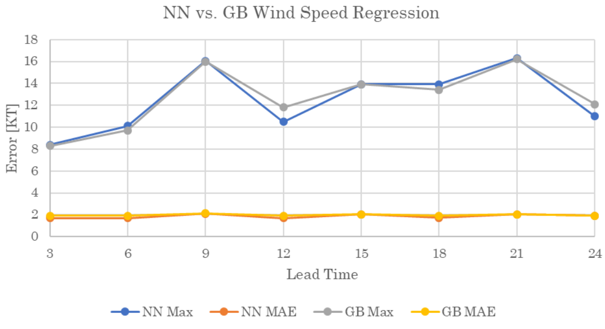


Figure 7. Performance comparison of the Multilayer Perceptron (MLP) neural network and Gradient Boosting (grid search calibrated hyperparameters) for wind speed prediction, including initial observed wind speed

In summary, the choice of model appears to have a limited impact on the results, with data quality being the primary driver of fluctuations. This is supported by the comparable performance of Neural Network and Gradient Boosting models with calibrated hyperparameters. While improvements are most evident when incorporating the initial observation from the forecast release time, the largest deteriorations occur during periods with limited dataset or model availability. Conversely, technical preprocessing aspects, such as the use of scalers or variations in model architecture, have a comparatively minor influence.

3.3. Performance comparison

To evaluate visibility prediction, we categorized visibility according to ICAO standards (0, 800, 1500, 3000, 5000 m or more) and compared Gradient Boosting classification results with professional TAF forecasts. Gradient Boosting was chosen for its suitability in probabilistic forecasting of multicategory variables, as supported by previous research (Natekin & Knoll, 2013).

Superficially, the classifier's accuracy suggests strong performance. However, a closer examination reveals discrepancies in the F1-score and log-loss, indicating potential issues with the classifier's reliability. The exceptionally high Expected Calibration Error (ECE), exceeding 0.9, signifies a severe lack of calibration. These patterns can be attributed to:

1. substantial dataset imbalance, where the 5–10 km category constitutes approximately 92% of observations, challenging algorithmic performance and smoothing metric application in multicategory probabilistic predictions and
2. ECE curve's indication of low data quality, which is worsened during runs with limited data availability.

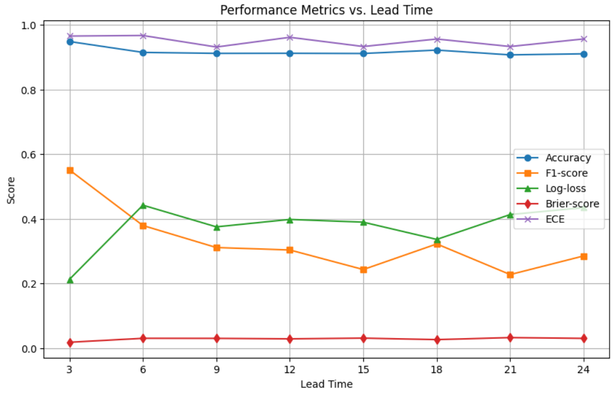


Figure 8. Performance metrics (Accuracy, F1-score, log-loss, Brier Score, Brier Skill Score, Expected Calibration Error) of the Gradient Boosting classification algorithm predicting visibility categories

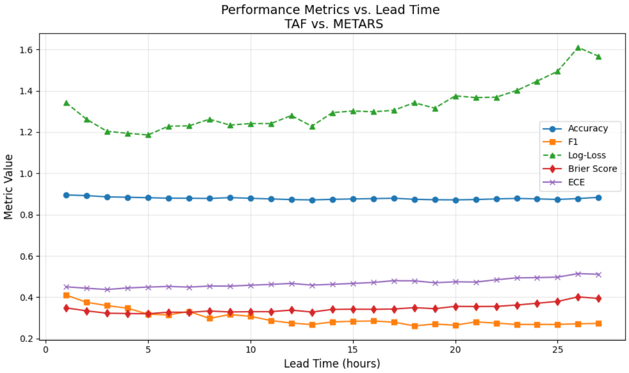


Figure 9. Performance metrics (accuracy, F1-score, log-loss, Brier Score, Expected Calibration Error) of professional TAF visibility forecasts from Brno-Tuřany International Airport

For the professional TAF forecasts, the same evaluation was performed and visualized in the Figure 9.

Values in the Figure 9 demonstrate that the TAF accuracy is comparable to Gradient Boosting. Higher expertise is reflected in a slightly lower, though still elevated, Expected Calibration Error (ECE). The Brier Score compared to the visibility class distribution is slightly higher by TAF (where 0 represents perfect skill). Evaluating human-issued predictions with log-loss presents challenges due to its sensitivity to probabilities of 0 and 1, common in Terminal Aerodrome Forecasts (TAFs) when a class is absent. To mitigate this, an epsilon parameter (0.01) was introduced, though a lower epsilon value would increase log-loss. Normalization of probabilities to sum to 1 (or 100%) is also necessary for objective evaluation, despite potentially altering the original intent of the prediction.

3.4. Case Study: January 2nd 2020

A specific case study is the production of a TAF forecast for a situation with a ridge of higher pressure over the Czech Republic, with reduced visibility in Brno on January 2, 2020.

TAF in Brno (LKTB):
202001012300 TAF AMD LKTB 020303Z
0203/0306 VRB02KT 7000 SCT003
TEMPO 0203/0209 2000 BR BKN003
PROB30 TEMPO 0203/0209 0600 FZFG OVC002
BECMG 0209/0212 CAVOK=

A Gradient Boosting model, trained on independent data to predict visibility multiclass categories, was evaluated on unseen data from a specific day. The model's predicted probabilities for each visibility class effectively captured the observed trend (Figure 10).

Following the evaluation of the Gradient Boosting model on unseen data, it was proceeded to a detailed comparison with the professional TAF forecast for January 2nd, 2020. Table 8 presents the accuracy metrics calculated for both the Gradient Boosting model and the human-issued TAF, evaluated at 3-hour and 30-minute intervals.

From the Figure 11 combined with Table 8, several observations can be drawn regarding the forecasting performance at both 3-hour intervals and 30-minute intervals.

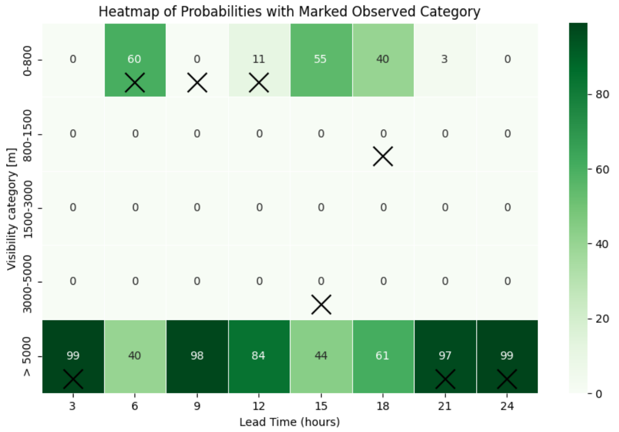


Figure 10. Predicted probabilities of visibility categories from the Gradient Boosting model for January 2nd, 2020, demonstrating potential TAF support product

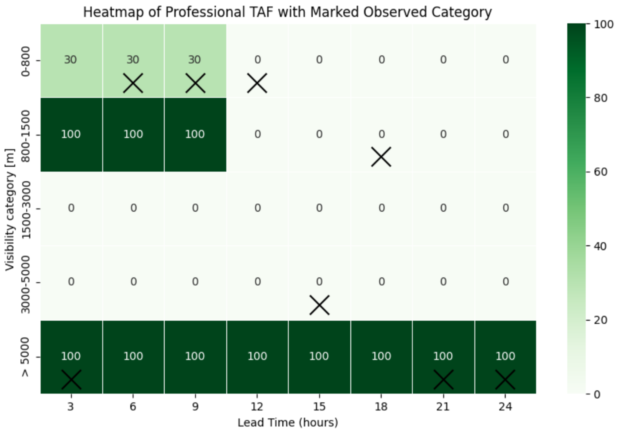


Figure 11. Visibility categories as forecast by the professional TAF for January 2nd, 2020, with clearly separated classes and probabilities leading to better illustrativeness in comparison to GB output

Table 8. Comparison of accuracy metrics of GB issued TAF supporting table for 2nd January 2020 and professional TAF forecast

	GB 3h	Human 3h	GB 30′	Human 30′
Accuracy	0.50	0.25	0.4	0.35
F1-score	0.40	0.18	0.43	0.38
Log-loss	3.50	4.06	3.04	2.40
Brier Score	0.49	0.62	0.45	0.41
Brier Skill Score	0.23	0.06	0.45	0.51
ECE	0.05	0.09	0.15	0.15

In the three hour step, Gradient Boosting (GB) outperforms the human forecast in accuracy, F1-score, and probabilistic calibration, as reflected in a better Brier Score and lower log-loss. The GB model provides more reliable probability estimates and captures fluctuating afternoon visibility trends more effectively. However, it struggles with middle-range categories, likely due to sharp boundary conditions during low visibility periods.

The Brier Skill Score (BSS) suggests that the human forecast is less skilful relative to a baseline, whereas GB shows an improvement. However, human forecasts still offer room for improvement in explanation, particularly in handling the TEMPO group, which affects the users' understanding of the conditions.

For finer time resolutions, key observations include:

1. Accuracy: GB (0.40) performs slightly better than the human model (0.35), though both remain well below 0.5, highlighting the difficulty of predicting visibility categories at this scale.
2. F1 Score: GB (0.43) outperforms the human model (0.38), demonstrating better handling of imbalanced classes.
3. Log Loss: GB (3.05) is higher than the human model (2.40), indicating that GB produces less confident probability estimates, even if it classifies categories more accurately.
4. Brier Score: GB (0.45) is higher than the human model (0.41), meaning the human model has better-calibrated probability estimates.
5. Brier Skill Score (BSS): The human model (0.51) achieves a higher skill score than GB (0.45), suggesting that while GB is more accurate overall, the human forecast provides more skillful probability estimates relative to a baseline model.
6. Expected Calibration Error (ECE): Both models show relatively low calibration errors (GB: 0.16, Human: 0.15), with human forecasts having a slight edge.

Gradient Boosting provides better raw classification performance in line with what it was trained for, which is reflected in higher accuracy and F1 scores. However, human forecasts demonstrate better probability calibration, leading to a lower Brier Score and higher BSS. This suggests that while GB is more consistent in categorical prediction, the human product offers more reliable probabilistic assessments.

4. Discussion and conclusions

This study investigated the feasibility of developing an automated framework for generating and evaluating Terminal Aerodrome Forecasts (TAFs), focusing on wind speed and visibility predictions at Brno-Tuřany International Airport. Key findings include:

1. Principal Component Analysis (PCA) effectively reduced dimensionality for wind speed predictors but not for visibility categories. In most predictors, one component explained more than 99% of the variance within four neighbouring points. However, PCA was less effective for precipitation (convective and rate) and Convective Inhibition (CIN), partially for Convective Available Potential Energy (CAPE).
2. Random Forest feature importance highlighted the dominance of initial observations and seasonal patterns in visibility prediction, revealing significant class imbalance.
3. Regression models for wind speed showed comparable performance, with data quality being the primary driver of accuracy. Notably, 100% of wind speed forecasts had a margin of error of 5 KT, indicating that even using a global model, the ICAO-required accuracy can be met.
4. Gradient Boosting (GB) classification for visibility demonstrated high accuracy but suffered from poor calibration and dataset imbalance.
5. A case study on January 2, 2020, revealed that GB outperformed human forecasts in raw classification metrics, while human forecasts exhibited better probability calibration.

4.1. Interpretations

The successful application of PCA for wind speed prediction underscores the predictability of wind fields through dimensionality reduction. However, the ineffectiveness of PCA for visibility value suggests that visibility is a more complex phenomenon, requiring alternative feature selection methods. The dominance of initial observations and seasonal patterns in visibility prediction highlights the persistence of atmospheric conditions and the influence of data quality on model reliance. The comparable performance of various regression models for wind speed, with data quality as the primary driver, indicates that model selection is less critical than data preprocessing and availability.

The high accuracy of the GB classification for visibility, contrasted with low calibration and dataset imbalance, underscores the challenges of predicting multicategory variables with skewed distributions. The case study on January 2, 2020, reveals a trade-off between raw classification performance (GB) and probability calibration (human forecasts), suggesting that human expertise still plays a crucial role in refining probabilistic predictions. The discrepancies between 3-hour and 30-minute interval forecasts highlight the scale-dependent challenges in visibility prediction. Furthermore, the formal characteristics of the TEMPO group were not explicitly incorporated in the evaluation, despite their operational importance. Since TEMPO conditions are expected to be reached in less than half of the cases and only temporarily, their omission may have led to biases in assessing forecast accuracy. Consequently, this study must emphasize that the goal of the proposed framework is not to replace human forecasters by simply slightly outperforming them. A direct comparison that dismisses human expertise is highly undesirable, as regulatory compliance, operational necessity, and the structured format of the coded forecast introduce constraints that a purely Machine Learning system does not have to deal with.

The findings have main implications for the development of comprehensive TAF generation systems. The study demonstrates the potential of machine learning models to provide accurate first-guess TAFs, particularly for wind speed and visibility prediction. However, it also highlights the need for robust feature selection techniques and calibration methods to address the complexities of visibility forecasting. The observed trade-off between classification accuracy and probability calibration suggests that hybrid approaches, combining machine learning with human expertise, may be most effective.

The identification of data quality as a primary driver of model performance underscores the importance of reliable and comprehensive observational data. The observed influence of seasonal patterns suggests that incorporating long-term climatological data could further enhance prediction accuracy. The study also provides a framework for evaluating TAF performance, revealing the strengths and weaknesses of different accuracy metrics. Given that extending the training dataset back to 2016–2024 could potentially double the training set size, applying undersampling or balancing techniques may improve model performance on minor categories. Specialized models could also be developed to handle underrepresented conditions effectively.

4.2. Limitations

This study is limited by its focus on a single airport (Brno-Tuřany) and a specific set of meteorological variables (wind speed and visibility). The findings may not be generalizable to other airports or weather phenomena. The use of historical data from 2020–2023 may not fully capture the evolving atmospheric dynamics and including more extensive database could improve the models' performance. The study's reliance on specific models (PCA, Random Forest, Gradient Boosting) may limit the exploration of alternative approaches.

The evaluation of human forecasts is constrained by the availability and format of TAF data, potentially affecting the objectivity of comparisons. The introduction of an epsilon parameter to mitigate log-loss sensitivity introduces a degree of subjectivity. The evaluation of the TEMPO group in human forecasts was limited, and further investigation into the impacts of these groups on user interpretation is needed. Overall, formal processing of final code is just introduced, but further it will be investigated within separate research.

4.3. Recommendations

Future studies should expand the geographical scope to include a wider range of airports and climatic conditions. Investigating additional meteorological variables, such as cloud cover and precipitation, would provide a more comprehensive evaluation of TAF forecasting. Exploring advanced machine learning techniques, such as deep learning and ensemble methods, could further enhance prediction accuracy.

To improve the reliability of automated TAF forecasts, developing robust calibration methods is essential to address dataset imbalance and enhance probabilistic predictions. Incorporating real-time observational data and long-term climatological information can further improve model adaptability and accuracy. With sufficiently accurate and locally adjusted forecasts, such a system could serve as a benchmark for evaluating the predictability of challenging weather situations. This benchmark could also facilitate comparisons of forecast complexity across airports in diverse climatic regions, providing comparison metrics of the local weather patterns' inherent variability.

Further investigation into the impact of TEMPO groups and other change groups on forecast interpretation and user understanding is warranted. Additionally, developing standardized evaluation metrics for probabilistic TAFs, accounting for both accuracy and calibration, would facilitate objective comparisons and enhance forecast reliability. Finally, creating a live, automated TAF framework, incorporating the findings of this study, would allow for continual testing and refinement of the system.

Acknowledgements

This research work was supported by the Project for the Development of the Organization DZRO "Military autonomous and robotic systems" under Ministry of Defence and Armed Forces of Czech Republic.

References

Altinok, A., Kiran, R., Bue, B., & Bilimoria, K. (2018). Modeling key predictors of airport runway configurations using learning algorithms. In *2018 Aviation Technology, Integration, and Operations Conference*. Aerospace Research Central. https://doi.org/10.2514/6.2018-3673

Anggoro, M., Siregar, D. C., Ninggar, R., & Widomurti, L. (2019). Aerodrome warning verification using quality measurement of contingency table (Case study in Jakarta and Tanjungpinang). In *Proceedings of the 1st International Conference on Statistics and Analytics* (*ICSA 2019*), Bogor, Indonesia. EUDL. https://doi.org/10.4108/eai.2-8-2019.2290486

Ayhan, S., Comitz, P., Gerberick, G., & Wilson, I. (2013). ADTOS: Arrival departure tradeoff optimization system. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming* (pp. 50–57). ACM Digital Library. https://doi.org/10.1145/2534303.2534314

Boyd, D., & Guinn, T. (2021). A comparison of the Localized Aviation MOS Program (LAMP) and Terminal Aerodrome Forecast (TAF) accuracy for general aviation. *Journal of Aviation Technology and Engineering*, *10*(1). https://doi.org/10.7771/2159-6670.1230

Buxi, G., & Hansen, M. (2013). Generating day-of-operation probabilistic capacity scenarios from weather forecasts. *Transportation Research Part C*: *Emerging Technologies*, *33*, 153–166. https://doi.org/10.1016/j.trc.2012.12.006

Chase, R. J., Harrison, D. R., Burke, A., Lackmann, G. M., & McGovern, A. (2022). A machine learning tutorial for operational meteorology. Part I: Traditional machine learning. *Weather and Forecasting*, *37*(8), 1509–1529. https://doi.org/10.1175/WAF-D-22-0070.1

Dhal, R., Roy, S., Taylor, C., & Wan, Y. (2013). Forecasting weather-impacted airport capacities for flow contingency management: Advanced methods and integration. In *AIAA Aviation Forum 2013*. Aerospace Research Central. https://doi.org/10.2514/6.2013-4356

Everitt, B. J., Landau, S., & Leese, M. (1974). *Cluster analysis*. John Wiley & Sons. https://openlibrary.org/books/OL21531058M/Cluster_analysis

Famiglini, L., Campagner, A., & Cabitza, F. (2023). Towards a rigorous calibration assessment framework: Advancements in metrics, methods, and use. In *Frontiers in artificial intelligence and applications. ECAI* (Vol. 372, pp. 645–652). IOS Press. https://doi.org/10.3233/FAIA230327

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2008). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

Hubbard, K. G., Goddard, S., Sorensen, W., Wells, N. P., & Osugi, T. T. (2005). Performance of quality assurance procedures for an applied climate information system. *Journal of Atmospheric and Oceanic Technology*, *22*(1), 102–111. https://doi.org/10.1175/JTECH-1657.1

International Civil Aviation Organization. (2016). *International Standards and Recommended Practices*: *Annex 3 to the Convention on International Civil Aviation – Meteorological Service for International Air Navigation* (19th ed.). ICAO.

Jacobs, A., & Maat, N. (2005). Numerical guidance methods for decision support in aviation meteorological forecasting. *Weather and Forecasting*, *20*(1), 82–100. https://doi.org/10.1175/WAF-827.1

Jolliffe, I. T. (2005). Principal component analysis. In *Encyclopedia of statistics in behavioral science*. John Wiley & Sons. https://doi.org/10.1002/0470013192.bsa501

Kenny, D. A. (1979). *Correlation and causality*. Wiley. http://ci.nii.ac.jp/ncid/BA00969679

Kicinger, R., Chen, J., Steiner, M., & Pinto, J. (2016). Airport capacity prediction with explicit consideration of weather forecast uncertainty. *Journal of Air Transportation*, *24*(1), 18–28. https://doi.org/10.2514/1.D0017

Lanyon, A., Standen, J., & Buchanan, P. (2020). A new process for producing first-guess TAFs. In *EGU General Assembly 2020*. https://doi.org/10.5194/egusphere-egu2020-5154

Mason, S. J. (2004). On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Monthly Weather Review*, *132*(7), 1891–1895. https://doi.org/10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00021

National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce. (2015). *NCEP GFS 0.25-degree global forecast grids historical archive*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/D65D8PWK

Nielsen-Gammon, J. W., Hu, X. M., Zhang, F., & Pleim, J. E. (2010). Evaluation of planetary boundary layer scheme sensitivities for the purpose of parameter estimation. *Monthly Weather Review*, *138*(9), 3400–3417. https://doi.org/10.1175/2010MWR3292.1

Novotný, J., Dejmal, K., Repal, V., Gera, M., & Sladek, D. (2021). Assessment of TAF, METAR, and SPECI reports based on ICAO ANNEX 3 regulation. *Atmosphere*, *12*(2), Article 138. https://doi.org/10.3390/atmos12020138

Patriarca, R., Simone, F., Di Gravio, G., Bonafé, G. P., Gobbi, G. P., & Angelini, F. (2023). Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering. *Expert Systems with Applications*, *213*, Article 119210. https://doi.org/10.1016/j.eswa.2022.119210

Simone, F., Di Gravio, G., & Patriarca, R. (2022). Performance-based analysis of aerodrome weather forecasts. In *2022 New Trends in Civil Aviation* (*NTCA*) (pp. 27–33). IEEE. https://doi.org/10.23919/NTCA55899.2022.9934004

Sládek, D., Chalupníková, B., Schneidrová, A., & Roučková, L. (2024). Analyses of European terminal aerodrome weather forecasts in 2022 and 2023. *Aviation*, *28*(2), 100–114. https://doi.org/10.3846/aviation.2024.21690

Wong, W.-K., Lau, C.-S., & Chan, P.-W. (2013). Aviation model: A fine-scale Numerical Weather Prediction system for aviation applications at the Hong Kong International Airport. *Advances in Meteorology*, *2013*, 1–11. https://doi.org/10.1155/2013/532475