



A High-Resolution Global METAR Dataset of Aviation Weather Observations from Major Cities (2024–2026): Characterization, Adverse-Event Analysis, and Predictive Modeling

Alla A. Kaybiyaynen
Management and Entrepreneurship
Kazan National Research Technological University
Kazan, Russian Federation
alhen2@yandex.ru

ABSTRACT

Aviation operations are highly sensitive to rapidly changing meteorological conditions, yet existing research often focuses on isolated airports or specific weather hazards. This study addresses these gaps by developing and analyzing a high resolution, global Meteorological Aerodrome Report (METAR) dataset comprising 958,338 observations from 26 major cities across 15 countries between 2024 and 2026. By integrating meteorological characterization, adverse-weather identification, and machine learning, the study quantifies the temporal and geographical variability of aviation weather conditions. Results reveal a severe class imbalance, with adverse weather events accounting for only 4.29% of observations. While classification models achieved high overall accuracy (up to 98.1%), minority class precision and recall remained low, though fog detection demonstrated strong ranking ability (ROC-AUC = 0.860). Conversely, regression modeling for a continuous severity score proved highly challenging ($R^2 = -1.43$), indicating that temperature alone is insufficient for severity prediction. Spatiotemporal analyses highlighted distinct diurnal peaks in adverse weather (11:00–14:00 UTC) and significant geographic heterogeneity. Ultimately, this study provides a comprehensive, multi hazard framework that underscores the need for advanced feature engineering and cost sensitive learning to support risk based operational decision making in global aviation.

Keywords: Aviation Meteorology, METAR, Adverse Weather, Machine Learning, Predictive Modeling, Class Imbalance, Diurnal Cycle, Global Dataset

Received: 3 May 2026, **Revised:** 10 July 2026, **Accepted:** 18 July 2026

Copyright: DLINE

1. Introduction

Aviation operations depend heavily on accurate, timely meteorological information because adverse weather can affect flight safety, airport operations, and economic efficiency. Modern aviation systems generate large

volumes of high frequency meteorological observations, with commercial aircraft alone contributing more than 130,000 observations per day, including temperature and wind measurements [1]. In addition, Meteorological Aerodrome Reports (METAR) provide standardized, airport level observations that are particularly valuable for monitoring rapidly changing atmospheric conditions.

Severe weather events such as heavy rainfall, strong winds, hail, squall lines, fog, reduced visibility, and extreme temperatures can create substantial operational challenges for aviation [2, 3, 4]. Consequently, high resolution numerical weather prediction and observational systems have become increasingly important for detecting and forecasting aviation relevant weather conditions [8]. Recent studies have also demonstrated the value of METAR observations for understanding fog development, wind shifts, and extreme cold conditions [13].

At the same time, machine learning approaches have increasingly been applied to statistical weather forecasting, particularly for fog and extreme meteorological events. Studies have explored artificial neural networks, deep learning, random forests, decision trees, support vector machines, autoregressive models, Bayesian networks, and ensemble methods [17, 18, 19]. Ensemble approaches such as random forests and gradient boosting have shown particular promise for fog event classification, while XGBoost has been applied for its robustness and ability to handle missing values in operational METAR data [31]. However, existing research has largely focused on individual airports, specific weather hazards, or individual modelling approaches.

To address this limitation, the present study develops and analyzes a high resolution global METAR dataset covering major cities during 2024–2026. The study combines meteorological characterization, adverse-weather event analysis, and predictive modelling to examine the temporal and geographical variability of aviation weather conditions and assess their predictability. By integrating observational analysis with data-driven modelling, the study aims to provide a comprehensive framework for understanding adverse aviation weather and supporting more informed risk based operational decision-making.

2. Background and Related Studies

The increasing availability of high resolution aviation meteorological observations has created new opportunities for characterizing weather conditions and developing predictive models for aviation applications. Commercial aircraft currently provide more than 130,000 meteorological observations per day, including temperature and wind measurements and, in some cases, humidity, vertical wind gusts, and eddy dissipation rate associated with turbulence. Temperature and wind observations obtained from aircraft are incorporated into operational numerical weather prediction systems at the National Centers for Environmental Prediction (NCEP) and other forecasting centers worldwide. These observations are also important for representing atmospheric conditions along oceanic air routes, including at synoptic times. The development of dedicated data access systems has further facilitated the use of such observations for meteorological analysis and forecasting [1].

Severe weather represents a particularly important area of aviation meteorological research because rapidly developing atmospheric events can adversely affect flight operations, airport infrastructure, and passenger safety. Meso and microscale analyses of severe convective weather commonly focus on phenomena such as short duration heavy rainfall, strong or gusty winds, hail, and squall lines, which can have substantial destructive potential [2-6]. Such events are especially significant for aviation because their occurrence near airports and along flight routes can cause considerable economic losses and increase operational and safety risks [7]. Consequently, reliable identification and prediction of adverse meteorological conditions are essential components of aviation weather support.

High resolution numerical weather prediction models (HNMs) have become increasingly important for analysing and predicting severe convective weather [8]. In aviation applications, these models can provide physically based information and refined guidance products for convective weather conditions at airports. However, the

usefulness of HNMs depends strongly on their reliability and predictive skill. Researchers have therefore developed various verification approaches to evaluate their performance. Traditional verification methods include pixel versus pixel measures such as the threat score [9, 10], while more recent approaches incorporate spatial verification techniques that are better suited to evaluating the location, structure, and spatial characteristics of predicted weather phenomena [11, 12].

Within this broader context, Meteorological Aerodrome Reports (METAR) provide an important source of high frequency, airport based observations for investigating weather development and its implications for aviation operations. Shankar investigated the onset and formation mechanisms of fog using high resolution hourly METAR observations from an airport in India located in the core of the Indo Gangetic Plain (IGP). Using observations covering April 2014 to October 2024, the study examined pre-fog atmospheric conditions, with particular emphasis on wind regimes and thermodynamic profiles, to characterize how different weather types develop and dissipate. The study proposed a pre-fog diagnostic framework that improved the ability to identify and distinguish weather types during early development. This approach has direct operational relevance because improved early fog identification can support more precise and timely airport fog advisories, enhancing preparedness and risk mitigation in weather sensitive aviation operations [13].

Low temperature conditions also represent an important component of aviation weather risk. Ziyi [14] established a progressive low temperature support framework by integrating SMART-SF sub seasonal temperature forecasts. The framework was intended to improve the standardization and reliability of extreme cold weather flight testing and was also identified as having potential applications in meteorological support for civil aviation operations during winter [14]. These studies demonstrate that aviation weather support extends beyond severe convective events and fog to include a broader range of meteorological conditions that can influence flight operations and safety.

Accurate probabilistic forecasting is particularly important in operational aviation because decisions concerning flight operations and risk mitigation often depend not only on predicted weather conditions but also on the associated uncertainty. Sadak examined how changes in operational weather regimes affect probability calibration, specifically investigating how a predictive model trained under one regime performs when applied to another and how much target domain data are needed to restore or improve calibration. This work highlights the importance of considering differences between weather regimes when developing operational forecasting systems and demonstrates that predictive performance and probability calibration may not remain stable when models are transferred across meteorologically distinct environments [15].

METAR observations can also be complemented by other local meteorological data sources. Cunningham [16] investigated the characteristics and complementary use of METAR and Personal Weather Station (PWS) observations. METAR observations for four airports during 2012 were obtained from the National Weather Service, while PWS observations from stations located within 10 miles of Hartsfield Jackson Atlanta International Airport (ATL) were obtained from Weather Underground. The study examined differences in the duration of wind shift events among stations and demonstrated both the advantages and limitations of METAR and PWS observations for categorizing such events. These findings emphasize the importance of understanding the characteristics, spatial representativeness, and limitations of observational sources when constructing aviation weather datasets.

Alongside the development of high resolution observational datasets, machine learning (ML) methods have increasingly been applied to statistical weather forecasting, particularly for fog and other adverse meteorological phenomena. Previous studies have investigated a broad range of approaches, including artificial neural networks [17, 18, 19] deep learning [20] feed forward neural networks and tree based ensemble methods [21] [Bari, D], random forests [22] Benàèek, P], decision trees [23, 24], support vector machines [25] [Negishi, M.], autoregressive time-series methods [26] and Bayesian decision networks [27]. Altogether, these studies demonstrate the growing role of data driven approaches in identifying complex nonlinear relationships among meteorological variables and predicting adverse weather events.

A comprehensive review by Salcedo Sanz et al. [28] examined the application of ML methods to the statistical modelling of extreme meteorological phenomena, including fog. Among studies specifically focused on statistical fog forecasting, Castillo Botón et al. [29] provide an important benchmark through an extensive evaluation of different ML algorithms for both classification and regression problems. For classification tasks, particularly adverse weather event identification, their results indicated that ensemble based methods, especially random forest and gradient boosting techniques, were well suited to fog event classification. This effectiveness was attributed to tree based learners' ability to represent characteristics associated with fog events [30].

More recently, Chen employed Extreme Gradient Boosting (XGBoost) as the principal predictive model because of its robustness for tabular datasets, computational efficiency, and native capability to handle missing observations [31]. This latter characteristic is particularly relevant to operational METAR data, where missing or incomplete observations can occur because of the nature of continuous observational streams. The use of XGBoost therefore illustrates the increasing emphasis on computationally efficient and robust ML approaches for operational meteorological prediction.

Overall, the existing literature demonstrates the growing importance of high resolution meteorological observations, particularly METAR data, for aviation weather characterization, severe weather analysis, fog forecasting, low temperature support, and operational risk assessment. Previous studies have addressed individual weather phenomena, specific airports, alternative observational sources, and a wide range of statistical and ML forecasting techniques. However, the literature in the extract focuses mainly on individual weather hazards, specific locations, or particular forecasting problems. A high resolution, multi-city METAR dataset covering multiple major cities over an extended recent period can therefore provide a broader basis for characterizing aviation weather conditions, identifying adverse weather events, and evaluating predictive modelling approaches across geographically and meteorologically diverse environments. Such a framework provides the basis for systematically linking descriptive characterization of aviation weather with adverse-event analysis and predictive modelling.

3. Research Issues and Problem Statement

Aviation operations are highly sensitive to rapidly changing meteorological conditions, including fog, reduced visibility, strong winds, heavy precipitation, severe convective activity, extreme temperatures, and other adverse weather phenomena. Although modern numerical weather prediction systems and high resolution forecasting models provide increasingly sophisticated meteorological guidance, aviation weather forecasting effectiveness depends on reliable, high frequency, spatially representative observations. METAR observations are particularly valuable in this context because they provide standardized, airport based measurements that directly describe the atmospheric conditions experienced at or near aviation facilities.

Despite their operational importance, several challenges remain in systematically analysing METAR observations. First, much existing research has focused on individual airports or specific weather hazards. For example, previous work has examined fog formation at individual airports, low temperature support for winter aviation operations, wind shift characteristics, or particular severe weather phenomena. Such studies provide important insights but offer limited evidence concerning the consistency of aviation weather patterns across geographically diverse airports and cities.

Second, adverse aviation weather is inherently multidimensional. A single meteorological variable may not adequately represent operational weather severity. Visibility, wind, precipitation, temperature, and weather phenomena may interact to produce different levels of operational risk. Consequently, analytical frameworks need to integrate multiple meteorological observations rather than treating individual weather variables as isolated phenomena.

Third, aviation forecasting requires not only accurate predictions but also reliable risk information and probability estimates. As Sadak shows, predictive performance and probability calibration can suffer when

models trained under one operational weather regime are applied to another. This issue is particularly relevant when developing models intended for application across different cities or climatic regimes. Therefore, assess the generalizability of predictive models by accounting for geographical and meteorological heterogeneity.

Fourth, operational METAR streams may contain missing or incomplete observations. This creates an additional methodological challenge for statistical and machine learning models. The use of XGBoost by Chen illustrates the importance of selecting predictive algorithms that are computationally efficient and robust to missing values, particularly when working with operational METAR data.

Finally, although machine learning approaches have increasingly been applied to fog and extreme weather forecasting, including artificial neural networks, deep learning, random forests, decision trees, support vector machines, autoregressive models, Bayesian networks, and ensemble methods, comparatively less attention has been given to integrating broad scale METAR characterization, adverse event identification, and predictive modelling within a single analytical framework.

Accordingly, the central research problem addressed by this study is the lack of an integrated, high resolution framework for characterizing aviation weather conditions across multiple major cities, systematically identifying adverse meteorological events, and evaluating their predictability using data driven methods. The study therefore seeks to transform high frequency METAR observations into an analytical framework that describes temporal and meteorological variability, quantifies adverse weather occurrence, and supports predictive aviation weather assessment.

3.1 Research Gaps

The existing literature reveals several important gaps that motivate the present study.

Geographical gap. Existing METAR based investigations have frequently focused on individual airports or geographically restricted areas. The study by Shankar, for example, investigated fog development using observations from a single airport in the Indo Gangetic Plain. Similarly, Cunningham examined wind shift characteristics using observations from selected airports and nearby personal weather stations. While these studies provide detailed local information, their findings may not adequately represent aviation weather variability across multiple climatic and geographical environments.

Temporal and resolution gap. High frequency meteorological observations contain substantial information about short duration weather transitions and rapidly developing adverse conditions. However, previous studies have often focused on specific weather events or relatively short observational periods. A high-resolution dataset spanning 2024–2026 offers an opportunity to investigate recent aviation weather behaviour at a finer temporal scale and identify short term variability that lower resolution analyses may obscure.

Multi-hazard gap. Much of the existing literature emphasizes particular hazards, especially fog, severe convection, wind shifts, or extreme cold. Severe convective weather studies have considered heavy rainfall, gusty winds, hail, and squall lines, while other studies have concentrated on fog or low temperature conditions. An integrated framework that simultaneously characterizes multiple adverse weather indicators remains comparatively underdeveloped.

Integration gap. Existing research commonly addresses observation characterization, weather event analysis, forecasting, or model verification as separate analytical tasks. The literature therefore provides limited evidence of an integrated workflow that progresses from raw METAR observations to descriptive characterization, adverse event classification, and predictive modelling.

Predictive generalization gap. Machine learning methods have demonstrated considerable potential for weather event classification and prediction. Castillo Botón et al, for example, reported strong suitability of ensemble approaches such as random forest and gradient boosting for fog event classification. However, the transferability of predictive models across different aviation weather regimes remains an important concern, particularly given the probability calibration issues identified by Sadak. This creates a need to examine predictive

performance in heterogeneous aviation environments rather than assuming that a model performs uniformly across locations and weather regimes.

Operational data gap. Operational METAR streams may contain missing observations and other data quality challenges. Chen's application of XGBoost highlights the relevance of computationally efficient algorithms capable of handling missing values in such datasets. Nevertheless, the methodological implications of missingness, variable heterogeneity, and high frequency observations require greater consideration within an integrated aviation weather modelling framework.

Collectively, these gaps indicate the need for a comprehensive approach that combines high-resolution METAR observations, multi-hazard weather characterization, adverse event analysis, and predictive modelling across multiple major cities. This study addresses that need.

3.2 Research Design and Methods

3.2.1 Research Design

The study uses a quantitative, observational, data-driven research design based on high resolution METAR observations from major cities. The methodological framework is designed to progress from data preparation and quality assessment to exploratory characterization, adverse weather identification, statistical analysis, and predictive modelling. This sequential structure enables the study to establish the characteristics of the observational dataset before assessing the occurrence and predictability of aviation relevant weather conditions.

The overall analytical workflow consists of five interconnected stages: (i) acquisition and preparation of METAR observations, (ii) quality assessment and preprocessing, (iii) descriptive and temporal characterization of meteorological conditions, (iv) identification and analysis of adverse weather events, and (v) development and evaluation of predictive models.

3.3 METAR Data Preparation

METAR observations constitute the primary observational source for the analysis. The dataset captures high-frequency aviation meteorological conditions across major cities during 2024–2026. Relevant meteorological variables include atmospheric conditions reported through METAR observations, with particular attention to aviation weather risk variables such as temperature, wind, visibility, precipitation, and reported weather phenomena.

Prior to statistical analysis, observations should be chronologically ordered and checked for duplicate records, inconsistent timestamps, invalid measurements, and missing values. Variables reported using categorical METAR codes should be transformed into analytically meaningful weather event indicators where appropriate.

3.4 Data Quality Assessment

Data quality assessment is performed before modelling to determine the completeness and consistency of the observations. Missing observations are identified for each meteorological variable, and the extent of missingness is evaluated across locations and time periods. Where appropriate, missing observations may be retained, removed, or handled using methods compatible with the selected analytical model.

Particular attention should be given to the treatment of missing meteorological observations because incomplete operational weather streams can influence both statistical estimates and predictive performance. This consideration is consistent with the motivation for using algorithms such as XGBoost, which can natively accommodate missing values.

3.5 Meteorological Characterization

Descriptive statistical analysis establishes the overall characteristics of the METAR dataset. Continuous meteorological variables are summarized using measures of central tendency and dispersion, while categorical weather conditions are summarized using frequencies and proportions.

Temporal analysis then investigates changes in meteorological conditions across the study period. Where sufficient observations are available, the analysis can examine hourly, daily, monthly, and seasonal patterns. This approach is particularly important for identifying recurring patterns in temperature, wind, visibility, precipitation, fog, and other weather phenomena relevant to aviation operations.

3.6 Adverse-Weather Event Identification

An adverse-weather framework identifies conditions that may have operational implications for aviation. Rather than treating weather variables independently, observations can be classified into operationally meaningful categories based on combinations of meteorological conditions.

Potential indicators include reduced visibility, fog, precipitation, strong winds or gusts, extreme temperature conditions, and severe-weather observations. The resulting categories provide a basis for quantifying the frequency and temporal distribution of adverse weather conditions across the study locations.

3.7 Comparative Analysis Across Locations

This study compares the occurrence and characteristics of adverse-weather conditions across major cities to assess geographical heterogeneity. The study can examine differences in the frequency, duration, and intensity of identified weather conditions using appropriate descriptive and inferential statistical methods.

This comparative component is particularly important because previous studies have frequently examined individual airports or geographically limited settings. A multi-city framework provides an opportunity to determine whether weather event characteristics are consistent across locations or exhibit substantial spatial variation.

3.8 Statistical Analysis of Weather Conditions

Statistical analyses examine relationships among meteorological variables and adverse-weather indicators. Correlation analysis can assess associations among continuous variables, while appropriate categorical-data tests can determine whether adverse-weather occurrence differs significantly among locations or weather categories.

Where appropriate, effect-size measures and confidence intervals should accompany significance tests so that the magnitude and uncertainty of observed differences can be evaluated rather than relying exclusively on p-values.

3.9 Predictive Modelling

Machine learning models are used to investigate whether observed meteorological variables can predict adverse weather conditions. The choice of algorithms is informed by previous research demonstrating the application of neural networks, deep learning, tree based methods, random forests, support vector machines, autoregressive approaches, Bayesian networks, and ensemble learning to meteorological forecasting.

Particular consideration can be given to ensemble tree based approaches, including random forest and gradient boosting, because previous research has identified their suitability for fog-event classification. XGBoost may also be considered because of its computational efficiency, robustness for tabular data, and ability to accommodate missing observations.

3.10 Model Evaluation

Evaluate predictive models using measures appropriate to the prediction task. For classification, accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve can quantify predictive performance. For probabilistic predictions, also assess calibration because accurate discrimination does not necessarily imply reliable probability estimates.

The importance of calibration is supported by Sadak's investigation of how probability calibration changes when models are transferred between operational weather regimes [Sadak. Accordingly, model evaluation

should consider not only predictive discrimination but also the reliability and generalizability of predictions across locations and weather regimes.

3.11 Model Interpretation and Aviation Relevance

Model interpretation identifies the meteorological variables that contribute most strongly to adverse weather predictions. Feature importance analysis can provide an interpretable representation of the relationship between meteorological observations and predicted weather events.

The final stage of the analysis links statistical and predictive findings to aviation weather applications. The objective is not merely to maximize predictive accuracy but to determine whether the identified meteorological patterns and predictive relationships can contribute to improved recognition of adverse conditions and risk-informed operational decision making.

3.12 Data Collection and Processing

The dataset was compiled from the Iowa Environmental Mesonet (IEM) ASOS/METAR archive [32] a well-established repository of global surface observations distributed by NOAA and international meteorological agencies. We collected data via a systematic, one request per station year process, explicitly fetching both routine and special reports. We applied no interpolation or quality control algorithms to the primary observations, ensuring the data reflect precisely what each station's automated or human observer system issued. This approach maintains the integrity of the original data and provides transparency for downstream quality control by end users.

3.13 Data Description

The dataset consists of two primary Comma-Separated Value (CSV) files, detailed below.

The File contains observations, which are the core observational data. It comprises 958,338 unique observational records. The schema is as follows:

Colum Name	Data Type	Description
station	String	ICAO (International Civil Aviation Organization) station identifier for the reporting airport (e.g., KLGA, RJTT).
valid_utc	Timestamp	Observation timestamp in Coordinated Universal Time (UTC), formatted as YYYY-MM-DD HH:MM.
temp_c	Float	Air temperature parsed from the METAR report, recorded in degrees Celsius (°C) with a resolution of 0.1°C where the T-group is present.
is_speci	Integer	A binary flag indicating the report type. 1 denotes a SPECI (special) report, while 0 denotes a routine (hourly/half-hourly) observation.
metar	String	The complete, raw METAR or SPECI text string as originally issued by the station. This field enables the extraction of all supplementary meteorological variables (e.g., wind speed/direction, visibility, present weather phenomena, cloud layers, pressure, and remarks) using any standard METAR parsing library.

File 2. stations.csv

This file provides the station-level metadata necessary for geographical and urban categorization. It contains five columns: station (ICAO identifier), city (served city), station_name (airport name), and country.

3.14 Spatial and Temporal Coverage

The dataset provides broad geographical coverage, with 26 stations situated in 26 major cities across five continents. The selected stations represent a diverse range of climate zones and include: Toronto (CYYZ), Munich (EDDM), Helsinki (EFHK), London City (EGLC), Amsterdam (EHAM), Warsaw (EPWA), Atlanta (KATL), Austin (KAUS), Dallas (KDAL), Houston (KHOU), Los Angeles (KLAX), New York (KLGA), Miami (KMIA), Chicago (KORD), Seattle (KSEA), San Francisco (KSFO), Madrid (LEMD), Paris-Le Bourget (LFPB), Ankara (LTAC), Istanbul (LTFM), Tokyo Haneda (RJTT), Buenos Aires (SAEZ), Moscow Vnukovo (UUWW), Lucknow (VILK), Beijing (ZBAA), and Shanghai Pudong (ZSPD). The temporal range covers a continuous period from January 1, 2024, through August 1, 2026, providing a contemporary 2.5-year baseline for analysis. Observations per station range from 24,840 to 53,862, reflecting different routine reporting cadences and the frequency of special reports.

3.15 Potential Applications and Research Use Cases

This high resolution, globally distributed dataset is designed to support a multitude of research objectives:

1. **Sub-Hourly Forecasting:** The temporal resolution of the data, which can be as fine as minutes between reports, makes it highly suitable for developing and validating time-series forecasting models for temperature and other weather parameters.
2. **Extreme Event Analysis:** The inclusion of SPECI reports makes the dataset particularly powerful for studying the dynamics and frequency of extreme events like heatwaves, cold snaps, heavy precipitation, and strong winds, which are precisely the conditions that trigger special reports.
3. **Urban Climate Studies:** The data can be used for cross-city comparisons of urban heat island effects, seasonal cycles, and synoptic weather influences, leveraging a consistent observation protocol across diverse global cities.
4. **Benchmarking and Methodological Research:** The preserved raw text provides an ideal corpus for natural language processing (NLP) studies aimed at benchmarking METAR parsers or developing new algorithms for structured data extraction from unstructured meteorological text. Furthermore, the dataset is a ready-made resource for systematically investigating the impact of temporal sampling frequency on the calculation of daily, monthly, and seasonal climate statistics.
5. **Reanalysis Validation:** As a source of independent ground truth observations, this dataset can serve as a critical benchmark for evaluating the performance of global reanalysis products and regional climate models at key urban locations.

3.16 Limitations and Caveats

Users of this dataset should be aware of several important considerations:

- The temp_c column contains parsed temperature values. However, on rare occasions, stations may issue reports without a parseable temperature, and these rows are omitted to maintain data integrity.
- All timestamps are in UTC. Researchers must convert these to local time zones before computing diurnal cycles or “daily” statistics (e.g., daily maximum temperatures) for a specific city.
- The temporal coverage (through August 1, 2026) marks the end of the current data collection period.

3.17 Data Availability and Citation

The dataset is openly available for download on the Kaggle platform at [33]. The data are provided under a

license specified on the repository. Users are asked to acknowledge the original data source (Iowa Environmental Mesonet) and to cite this dataset descriptor in any resulting publications.

4. Results

Metric	Value	Percent
Observations	958338	100
Stations	26	0.002713
Countries	15	0.001565
Adverse-weather observations	41137	4.292536
Low-visibility observations	19327	2.016721
Snow observations	20174	2.105103
Fog observations	20143	2.101868
Rain observations	96928	10.11418
Thunderstorm observations	12033	1.255611

Table 1. Summary statistics of the METAR observational dataset

Table 1 provides an overview of the dataset composition and the prevalence of key weather phenomena extracted from the raw METAR/SPECI reports. The corpus comprises 958 338 unique observations from 26 stations across 15 countries. Adverse-weather observations (any report containing significant weather phenomena that may affect aviation operations) account for 41 137 records (4.29 %). Among specific phenomena, rain is the most frequent (96 928 observations, 10.11 %), followed by snow (20 174; 2.11 %), fog (20 143; 2.10 %), low visibility (19 327; 2.02 %), and thunderstorms (12 033; 1.26 %).

These frequencies confirm the strong class imbalance inherent in operational aviation weather data: most reports describe benign conditions, while the events of greatest operational interest remain comparatively rare. The imbalance has direct implications for downstream modelling (see Table 2) and underscores the value of SPECI reports, which preferentially capture rapidly changing or hazardous conditions.

4.1 Classification Model Results

target	train_n	test_n	accuracy	precision	recall	F1	ROC_AUC	PR_AUC
adverse	105470	144530	0.950412	0.060736	0.027772	0.038116	0.789725	0.203359
lowvis	105470	144530	0.980862	0.136264	0.025462	0.042907	0.838204	0.135424
snow	105470	144530	0.979368	0.069556	0.032425	0.044231	0.707039	0.027964
fog	105470	144530	0.976828	0.186788	0.145843	0.163795	0.859732	0.132875

Table 2. Binary classification performance for adverse-weather and related targets

Table 2 reports the performance of classification models trained to detect four binary targets adverse weather, low visibility, snow, and fog on a fixed train/test split (105 470 / 144 530 observations). Accuracy is high across all targets (0.950–0.981), as expected under severe class imbalance. Precision, recall and F1 scores are markedly lower, reflecting the difficulty of correctly identifying the minority class.

- Fog yields the strongest minority class metrics (precision 0.187, recall 0.146, F1 0.164) and the highest ROC-AUC (0.860).
- Low-visibility and adverse weather models achieve moderate ranking ability (ROC-AUC 0.838 and 0.790) but remain limited by low precision recall area under the curve (PR-AUC 0.135 and 0.203).
- Snow detection is the weakest (ROC-AUC 0.707, PR-AUC 0.028), consistent with the geographic distribution of the stations (many lie outside persistent snow climates).

Overall, the results demonstrate that simple models can rank the risk of adverse conditions reasonably well, yet operational decision thresholds would require careful calibration or cost-sensitive learning to achieve usable precision and recall.

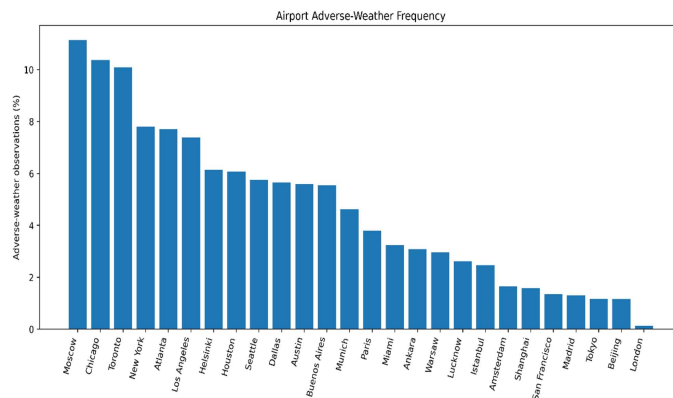


Figure 1. Airport-level frequency of adverse-weather observations

4.3 Airport-level frequency of adverse-weather observations

Figure 1 ranks the 26 airports by the percentage of reports classified as adverse weather. Moscow (UUWW), Chicago (KORD) and Toronto (CYYZ) exhibit the highest frequencies (approximately 10–11 %), followed by New York (KLGA), Atlanta (KATL) and Los Angeles (KLAX). At the opposite extreme, London City (EGLC), Beijing (ZBAA) and Tokyo (RJTT) show frequencies below 1.5 %. The ranking reflects both climatological differences (continental versus maritime climates) and local reporting practices, and provides a clear baseline for inter city comparative studies of urban weather risk.

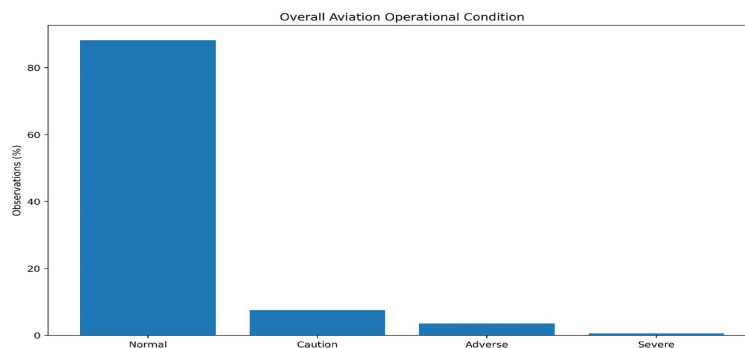


Figure 2. Overall distribution of aviation operational conditions

4.4 Distribution of aviation operational conditions

Figure 2 aggregates all observations into four operational categories: Normal ($\approx 88\%$), Caution ($\approx 8\%$), Adverse ($\approx 3.5\%$) and Severe ($\approx 0.5\%$). The overwhelming predominance of Normal conditions again highlights the class imbalance challenge, while the non negligible Caution and Adverse tails quantify the operational exposure that forecasting and decision-support systems must address.

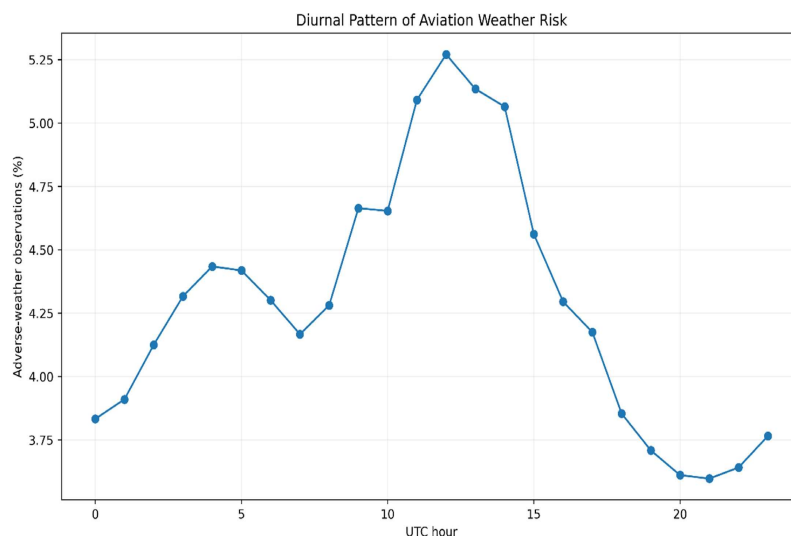


Figure 3. Diurnal cycle of adverse weather frequency

4.5 Diurnal cycle of adverse-weather frequency

Figure 3 displays the percentage of adverse weather reports as a function of UTC hour. A clear diurnal pattern emerges: frequency rises through the morning, reaches a broad maximum of approximately 5.25 % between 11:00 and 14:00 UTC, and declines steadily toward a nocturnal minimum near 3.6 % around 20:00–22:00 UTC. The afternoon peak is consistent with the typical timing of convective activity and the diurnal temperature cycle at many mid-latitude stations, offering a useful prior for nowcasting systems.

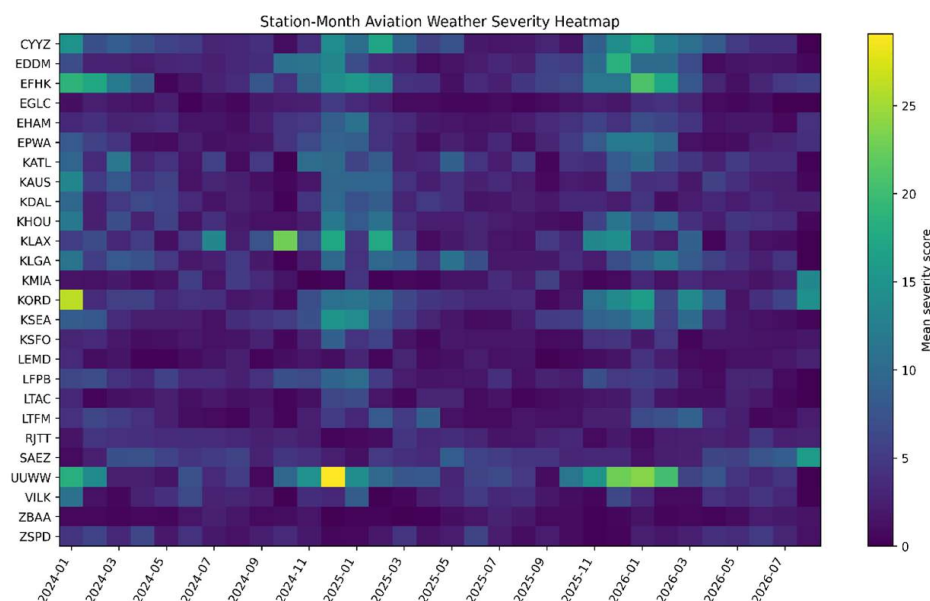


Figure 4. Station-month heatmap of mean aviation weather severity

Figure 4 presents a heatmap of mean severity score by station (rows) and calendar month (columns) from January 2024 to July 2026. Most station month cells are dark (low severity). Distinct bright bands appear at selected locations and seasons, most notably Chicago (KORD) in early 2024, Buenos Aires (SAEZ) in several winter periods, and Moscow (UUWW) during colder months, indicating episodic concentrations of elevated weather risk. The visualisation is valuable for identifying both climatological hotspots and anomalous months that merit case-study analysis.

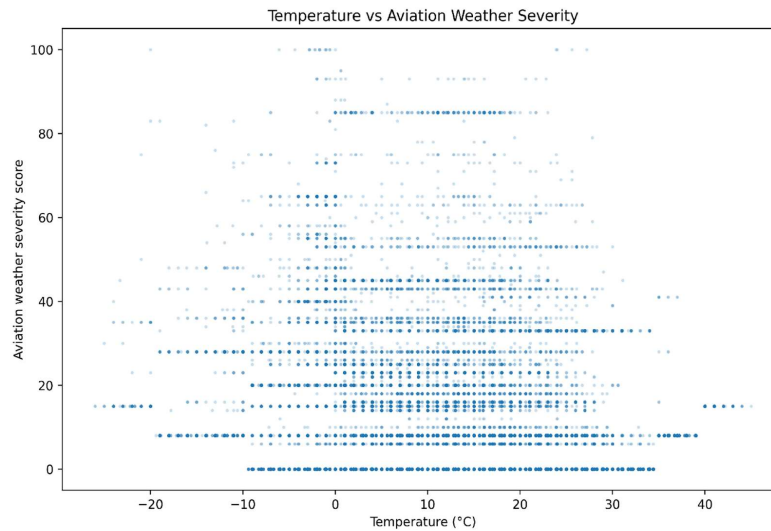


Figure 5. Relationship between air temperature and aviation weather severity

4.6 Relationship between air temperature and aviation weather severity

Figure 5 is a scatter plot of temperature (°C) against the continuous severity score. Severity values span nearly the full range at almost every temperature, yet a pronounced concentration of points near zero severity is visible across the entire temperature domain. Horizontal banding suggests that the severity score takes on a limited set of discrete or semi discrete values. Extreme cold (below -10°C) and extreme heat (above 30°C) both contain elevated severity observations, but no simple monotonic relationship is evident. The plot therefore supports the conclusion drawn from Table 3 that temperature alone is an insufficient predictor of aviation weather severity, and that additional meteorological variables (visibility, precipitation type, wind, etc.) must be incorporated.

Collectively, the tables and figures characterise a large, globally distributed, high temporal resolution METAR corpus, quantify the rarity of operationally critical events, document strong diurnal and geographic structure, and illustrate the modelling challenges posed by class imbalance and weak linear predictability of continuous severity. These results establish a quantitative foundation for subsequent forecasting, extreme event and urban climate analyses.

train_n	test_n	MAE_sm	RMSE_sm	R2
3003	37051	4.26279	5.114313	-1.42997

Table 3. Regression performance for continuous aviation weather severity

4.7 Regression performance for continuous aviation weather severity

Table 3 summarises a regression model that predicts a continuous aviation weather severity score. With only 3 003 training observations the model attains a mean absolute error of 4.26 and a root mean square error of 5.11 on a large held out set (37 051 observations). The coefficient of determination is strongly negative ($R^2 = -1.43$), indicating that the model performs worse than a trivial mean predictor. This outcome suggests either

that the chosen severity score is only weakly related to the available predictors under the current feature set, or that the small training sample is insufficient to capture the underlying relationships. Further feature engineering or larger training corpora would be required before the severity score can be used as a reliable continuous target.

5. Discussion

The present study successfully establishes an integrated, high resolution framework for characterizing aviation weather conditions across a diverse set of global urban environments. By analyzing nearly one million METAR and SPECI observations, the research addresses critical geographical and multi hazard gaps in the existing literature, which has traditionally been confined to single airports or isolated weather phenomena.

A primary finding of this study is the profound class imbalance inherent in operational aviation weather data. Adverse weather observations made up only 4.29% of the total dataset, while most reports described benign, “Normal” conditions. This imbalance directly impacts predictive modeling; while classification models achieved high accuracy (95.0% to 98.1%), their precision and recall for minority classes were markedly low. However, the models demonstrated varying degrees of success depending on the hazard. Fog classification yielded the strongest minority-class metrics and the highest ROC-AUC (0.860), corroborating previous literature that highlights the suitability of tree-based ensemble methods for fog event identification. In contrast, snow detection was the weakest, likely reflecting the geographic distribution of the selected stations, many of which are situated outside persistent snow climates.

Spatiotemporal analyses provided valuable operational insights. The identification of a distinct diurnal cycle peaking between 11:00 and 14:00 UTC aligns with the typical timing of convective activity and mid-latitude diurnal temperature cycles, offering a useful prior for nowcasting systems. Furthermore, the airport level frequency analysis highlighted significant geographic heterogeneity, with continental airports like Moscow (UUWW), Chicago (KORD), and Toronto (CYYZ) exhibiting the highest adverse-weather frequencies.

The failure of the continuous severity regression model ($R^2 = -1.43$) is a critical discussion point. It demonstrates that temperature alone is an insufficient predictor of aviation weather severity and that the current feature set and training sample size ($n=3,003$) are inadequate for capturing complex, non linear severity relationships. This underscores the necessity for multi-variable feature engineering incorporating visibility, precipitation type, and wind dynamics to accurately model continuous operational risk.

6. Limitations

While this study provides a robust foundation for global aviation weather analysis, several limitations must be acknowledged:

- 1. Class Imbalance and Model Calibration:** The extreme rarity of adverse events (e.g., thunderstorms, snow) inherently limits the precision and recall of standard classification models. The study did not employ advanced resampling techniques or cost sensitive learning, which are necessary to optimize decision thresholds for rare but high impact events.
- 2. Continuous Severity Predictability:** The regression model for continuous aviation weather severity performed worse than a trivial mean predictor ($R^2 = -1.43$). This indicates that the severity score is only weakly related to the available predictors under the current feature set, and the small training sample was insufficient to capture underlying relationships.
- 3. Feature Extraction Constraints:** The modeling relied heavily on parsed variables (such as temp_c) and basic extracted phenomena. The predictive phase did not fully leverage deeper natural language processing (NLP) of the raw metar text strings to extract complex cloud layers, precise wind gusts, or nuanced remarks, potentially limiting model performance.

4. Temporal and Processing Constraints: All timestamps in the dataset are recorded in UTC. While this ensures standardization, researchers must manually convert these to local time zones to accurately compute localized diurnal cycles or daily statistics.

5. Temporal Coverage: The dataset's temporal coverage concludes on August 1, 2026. Consequently, it does not capture future climate shifts or long-term multi-decadal trends, limiting its use strictly to contemporary baseline analysis.

6. Model Generalization: As noted in the background literature, transferring predictive models across distinct meteorological regimes can degrade probability calibration. This study evaluated models on a fixed train/test split but did not explicitly test cross-regime transferability across the 26 diverse global cities.

7. Conclusion

This study presents a comprehensive, data-driven framework for analyzing high resolution METAR observations across 26 major global cities from 2024 to 2026. By transforming nearly one million raw observational records into an integrated analytical workflow, the research successfully characterizes the multi hazard aviation weather landscape, quantifying both the rarity of operationally critical events and their distinct geographic and diurnal structures.

The findings highlight the significant challenge of class imbalance in operational meteorological data, showing that while machine learning models can effectively rank the risk of specific events like fog, achieving usable precision and recall for rare adverse conditions requires careful threshold calibration. Furthermore, the inability to accurately predict continuous weather severity using basic temperature metrics underscores the need for advanced feature engineering that integrates multiple meteorological variables.

Ultimately, this globally distributed, high temporal resolution dataset serves as a vital resource for the aviation meteorology community. It provides a quantitative foundation for sub hourly forecasting, extreme-event analysis, and urban climate studies, paving the way for more informed, risk based operational decision-making in global aviation. Future work will focus on enhancing feature extraction from raw METAR strings and applying cost sensitive learning to improve the predictability of rare, high impact adverse weather events.

References

- [1] Moninger, W. R., Mamrosh, R. D., & Pauley, P. M. (2003). Automated meteorological reports from commercial aircraft. *Bulletin of the American Meteorological Society*, 84 (2), 203-216.
- [2] Fu, L., Li, W. J., Zhang, P. Q., Zhang, Q., Gao, G. (2011). Inter-decadal change of hail events over China and causation analysis in north China in recent 50 years. *Meteorological Monographs*, 37(6), 669-676.
- [3] Wan, Z. W., Liu, X. H. (2018). Simulation analysis of severe convective weather in weakly forced system in East China with GRAPES-Meso. *Materials Science and Technology*, 46(2), 282-291.
- [4] Xue, X. Y., Ren, G. Y., Sun, X. B., Ren, Y. (2019). Climatological characteristics of meso scale and micro-scale strong convective weather events in China. *Climatic and Environmental Research*, 24(2), 199-213.
- [5] Yu, X. D., Zhou, X. G., Wang, X. M. (2012). The advances in the nowcasting techniques on thunderstorms and severe convection. *Acta Meteorologica Sinica*, 70 (3), 311-337.
- [6] Zhang, Q. H., Ni, X., Zhang, F. Q. (2017). Decreasing trend in severe weather occurrence over China during the past 50 years. *Scientific Reports*, 7, 42310. <https://doi.org/10.1038/srep42310>.
- [7] Wan, Z., Liu, X., Xu, C. (2021). Multi-source observations and high-resolution numerical model applied on the

analysis of a severe convective weather affecting the airport. *Meteorological Applications*, 28(4), e2012. <https://doi.org/10.1002/met.2012>.

[8] Qi, L. B. (2015). Operational progress of high resolution numerical model on severe convective weather warning. *Meteorological Monthly*, 41(6), 661–673.

[9] Huang, L. P., Chen, D. H., Deng, L. T., Xu, F. Z., Yu, F., Jiang, Y., et al. (2017). Main technical improvements of GRAPES_Meso V4.0 and verification. *Journal of Applied Meteorological Science*, 28(1), 25–37.

[10] Zhou, B. B., Du, J., Liu, S., DiMego, G. (2011). Verifications of simulated radar reflectivity and echo-top forecasts at NCEP. In *24th Conference on Weather and Forecasting, 91st Annual Meeting* (p. 90–102). Seattle, WA.

[11] Tang, W. Y., Zheng, Y. G., Zhang, X. W. (2018). Fss-based evaluation on convective weather forecasts in North China from high resolution models. *Journal of Applied Meteorological Science*, 29(5), 513–523.

[12] Zhang, X. W., Tang, W. Y., Zheng, Y. G., Sheng, J., Zhu, W. J. (2020). Comprehensive evaluation of GRAPES_3km numerical model in forecasting convective storms using various verification methods. *Meteorological Monographs*, 46(3), 367–380.

[13] Shankar, A., Sahana, B. C., Kumar, A. (2026). Visibility challenges: Characterizing fog events in the Indo-Gangetic plain for enhanced aviation operations. *Meteorology and Atmospheric Physics*, 138, 8. <https://doi.org/10.1007/s00703-025-01115-6>.

[14] Ziyi, W., Yangjinxi, G., Zhenyu, C., Junhui, D., Wenting, Z., Mo, J. (2026). Statistical analysis and forecast comparison of meteorological conditions for commercial aircraft's extreme cold weather flight testing in China. In *Proceedings of the 8th China Aeronautical Science and Technology Conference (CASTC 2025)*, Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-95-3019-9_34.

[15] Sadak, M. S. (2026). Data-efficient calibration transfer across operational weather regimes for probabilistic airport hazard prediction. *International Journal of Data Science and Analytics*, 22, 224. <https://doi.org/10.1007/s41060-026-01189-4>.

[16] Cunningham, J., Asencio, M., Lent, S., Brinton, C. (2014). The categorization of wind shift events using METAR and personal weather station data. In *2014 Integrated Communications, Navigation and Surveillance Conference (ICNS) Conference Proceedings* (p. S3-1–S3-12). Herndon, VA. <https://doi.org/10.1109/ICNSurv.2014.6820019>.

[17] Colabone, R. O., Ferrari, A. L., Vecchia, F. A. S., Tech, A. R. B. (2015). Application of artificial neural networks for fog forecast. *Journal of Aerospace Technology and Management*, 7, 240–246.

[18] Durán-Rosal, A. M., Fernández, J. C., Casanova-Mateo, C., Sanz-Justo, J., Salcedo-Sanz, S., Hervás-Martínez, C. (2018). Efficient fog prediction with multi-objective evolutionary neural networks. *Applied Soft Computing*, 70, 347–358.

[19] Cornejo-Bueno, S., Casillas-Pérez, D., Cornejo-Bueno, L., Chidean, M. I., Caamaño, A. J., Cerro-Prada, E., Casanova-Mateo, C., Salcedo-Sanz, S. (2021). Statistical analysis and machine learning prediction of fog-caused low-visibility events at A-8 motor-road in Spain. *Atmosphere*, 12, 679.

[20] Zhu, L., Zhu, G., Han, L., Wang, N. (2017). The application of deep learning in airport visibility forecast. *Atmospheric and Climate Sciences*, 7, 314–322.

[21] Bari, D., Ouagabi, A. (2020). Machine-learning regression applied to diagnose horizontal visibility from mesoscale NWP model forecasts. *SN Applied Sciences*, 2, 556.

- [22] Benáèek, P., Borovièka, P. (2021). Application of machine learning in fog prediction in the Czech Republic. *Meteorological Bulletin*, 74, 141–148.
- [23] Dutta, D., Chaudhuri, S. (2015). Nowcasting visibility during wintertime fog over the airport of a metropolis of India: Decision tree algorithm and artificial neural network approach. *Natural Hazards*, 75, 1349–1368.
- [24] Bartoková, I., Bott, A., Bartok, J., Gera, M. (2015). Fog prediction for road traffic safety in a coastal desert region: Improvement of nowcasting skills by the machine-learning approach. *Boundary Layer Meteorology*, 157, 501–516.
- [25] Negishi, M., Kusaka, H. (2022). Development of statistical and machine learning models to predict the occurrence of radiation fog in Japan. *Meteorological Applications*, 29, e2048.
- [26] Guijo-Rubio, D., Gutiérrez, P. A., Casanova-Mateo, C., Sanz-Justo, J., Salcedo-Sanz, S., Hervás-Martínez, C. (2018). Prediction of low-visibility events due to fog using ordinal classification. *Atmospheric Research*, 214, 64–73.
- [27] Boneh, T., Weymouth, G. T., Newham, P., Potts, R., Bally, J., Nicholson, A. E., Korb, K. B. (2015). Fog forecasting for Melbourne Airport using a Bayesian decision network. *Weather and Forecasting*, 30, 1218–1233.
- [28] Salcedo-Sanz, S., Pérez-Aracil, J., Ascenso, G., Del Ser, J., Casillas-Pérez, D., Kadow, C., Fister, D., Barriopedro, D., García-Herrera, R., Restelli, M. (2022). Analysis, characterization, prediction and attribution of extreme atmospheric events with machine learning: A review. *arXiv*, arXiv:2207.07580.
- [29] Castillo-Botón, C., Casillas-Pérez, D., Casanova-Mateo, C., Ghimire, S., Cerro-Prada, E., Gutierrez, P. A., Deo, R. C., Salcedo-Sanz, S. (2022). Machine learning regression and classification methods for fog events prediction. *Atmospheric Research*, 272, 106157.
- [30] Bartok, J., Šišán, P., Ivica, L., Bartoková, I., Malkin Ondík, I., Gaál, L. (2022). Machine learning based fog nowcasting for aviation with the aid of camera observations. *Atmosphere*, 13(10), 1684.
- [31] Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- [32] Hertzmann (2017). mesonet.agron.iastate.edu
- [33] <https://www.kaggle.com/datasets/rngrng/metar-weather-26-global-cities-2024-2026>