Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

TITLE PAGE

Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering

Riccardo Patriarca [1]*; Francesco Simone* [2]*; Giulio Di Gravio* [3]

[1] Department of Mechanical and Aerospace Engineering, Sapienza University of Rome (Italy). Mail: riccardo.patriarca@uniroma1.it

[2] Department of Mechanical and Aerospace Engineering, Sapienza University of Rome (Italy). Mail: francesco.simone@uniroma1.it

[3] Department of Mechanical and Aerospace Engineering, Sapienza University of Rome (Italy). Mail: giulio.digravio@uniroma1.it

*corresponding author: Riccardo Patriarca. Department of Mechanical and Aerospace Engineering, Sapienza University of Rome (Italy), Via Eudossiana, 18 – 00184 Rome (Italy). Tel: +390644585252; Mail: riccardo.patriarca@uniroma1.it

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

Abstract

Weather forecasting is a critical factor for aerodrome and enroute flight operations. Airport decision-makers rely on assessments made by forecasters to ensure operations safety and optimize flight schedule despite potential adverse weather conditions. This manuscript suggests a novel methodology based on Machine Learning to detect forecasting anomalies in historic data, and to rely on them for anticipating potential threats in aerodrome future forecasts. The methodology is fed with historic bulletins from radars and with previous forecasts, which are then processed via an anomaly detection algorithm, and a hierarchical clustering algorithm. While the former algorithm spots anomalous data points, the latter is used to group sets of similar forecasts. The joint usage of the results allows calculating an error propensity metric, which can predict the expected tendency of a certain forecast to be inaccurate. The methodology is meant to enhance decision makers in managing aerodrome weather forecasting, understanding criticalities related to their accuracy levels.

Keywords

Artificial Intelligence; Decision making; Hierarchical clustering; Anomaly detection; Weather forecasting

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

Acronyms

ANSP	Air Navigation Service Provider
AI	Artificial Intelligence
AIRMET	AIRman's METeorological Information
ANN	Artificial Neural Network
BECMG	BECoMinG indicator for change group in TAF
CNN	Convolutional Neural Network
CSI	Critical Success Index
DL-FC	Deep Learning Fully-Connected
ETL	Extraction Transformation Loading
FAR	False Alarm Ratio
FBI	Frequency Bias Index
FM	FroM indicator for change group in TAF
HC	Hierarchical Clustering
ICAO	International Civil Aviation Organization
KPI	Key Performance Indicator
J48	Decision Tree classification
METAR	METeorological Aerodrome Report
ML	Machine Learning
MLP	MultiLayer Perceptron classifier
NsNsNs	Cloud type (i.e. cloudiness) weather element
PC	Proportion Correct
POD	Probability of Detection
PROB	PROBability indicator for change group in TAF
RBF	Radial Basis Function classifier
RF	Random Forest
SGD	Stochastic Gradient Descendent
SIGMET	Significant Meteorological Information
SPECI	Special meteorological aerodrome Report
SR	Spectral Residual
TAF	Terminal Aerodrome Forecast
TEMPO	TEMPOrary indicator for change group in TAF
TT	Temperature weather element
VVVV	Visibility weather element
ddd	Wind direction weather element
ff	Wind intensity (or velocity) weather element
nsnsns	Cloud celing (i.e. height) weather element
ww	Weather phenomena and precipitations weather element

Notation

$\mathcal{A}$	Set of anomalous TAFs
$\mathbb{AD}$	Set of AD aerodromes
ΔT	Time step for accuracy analysis
$\Delta T'$	Time step for anomaly detection algorithm
$\mathcal{F}(x)$	Fourier transform operator of a function x
$\mathcal{F}^{-1}(x)$	Inverse Fourier transform operator of a function x
α_C, β, γ	Parameters for Lance-Williams recursive algorithm referred to a cluster C
δ	Scale factor to obtain $\Delta T'$ from ΔT
η_C	Error propensity metric for TAFs belonging to cluster C
θ	Angle between two observation vectors
μ	Mean value of time series points included in the sliding window of the anomaly detection algorithm
σ	Variance of time series points included in the sliding window of the anomaly detection algorithm
φ	Silhouette score of a cluster
$\bar{\varphi}$	Average silhouette score of a set of clusters
A_o	Mean distance between an observation o and other observations in its cluster
AD	Aerodrome referred to TAFs
$ALA(f)$	Averaged logarithmic amplitude spectrum operator for a function f
B_o	Mean distance between an observation o and other observation in other clusters
C	Cluster obtained from hierarchical clustering algorithm
H_{CORR}	Number of ΔT in which the TAF is correct during TEMPO group validity
H_M	Number of ΔT in which the main forecast is correct during TEMPO group validity
H_T	Number of ΔT in which the TEMPO forecast is correct during TEMPO group validity
H_{TnotM}	Number of ΔT in which the TEMPO forecast is correct, and the main forecast is not correct during TEMPO group validity
$LA(f)$	Logaritmic amplitude spectrum operator for a function f
M	Total observation features
Nc	Total number of clusters
N_u	Total number of ΔT time steps within TAF validity
$P(f)$	Phase operator for a function f
$\overline{POD}$	Average POD value time series for a single aerodrome
$\overline{\overline{POD}}$	Average POD value time series for a set of aerodromes
R	TAF richness, i.e., how many analyzed weather attributes it contains
$S(f)$	Saliency function of a function f
$SR(f)$	Spectral residual for a function f
T	Analysis end time
T^*	Sets of anomalous time steps identified by anomaly detection algorithm
TAF	u-th element of the set of TAF

113	U	Set of TAFs, whose time includes t
114	V^e	TAF validity end time
115	V^s	TAF validity start time
116	a	Hit score
117	b	Correct rejection score
118	c	Miss score
119	d	False alarm score
120	$d_{\vec{o}_i\vec{o}_j}$	Distance between generic observation i and generic observation j
121	$d_{C_IC_J}$	Distance between generic cluster I and generic cluster J
122	f	Fourier transform of time series x
123	$h_n(f)$	Convolution matrix of a function f
124	m	Index for m-th observation feature, $m = 1, \dots, M$
125	n	Index for n-th TAF validity time step, $n = 1, \dots, N_u$
126	$\vec{o}$	Observation vector for hierarchical clustering algorithm
127	r	Index for r-th TAF element, $r = 1, \dots, R$
128	s_C	Size of a cluster C
129	t	Time index for POD aggregation referred to accuracy analysis, $t = 0, \dots, T$ with
130		increment ΔT
131	t'	Time index for anomaly detection time series, $t' = 0, \dots, T$ with increment $\Delta T'$
132	t^*	Anomalous time step contained in T^*
133	u	Index for u-th TAF
134	x	Input time series for anomaly detection algorithm of length T
135	$\bar{x}$	Average of time series point for time series x
136	y	Output of anomaly detection algorithm of length T

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

1. Introduction

The economic growth of the aviation sector is largely determined on the optimization of flights schedule (Atay et al., 2021). For example, adverse weather conditions nearby airports, or enroute, may lead to the interruption of the scheduled plan, i.e., unpredictable flight delays, on-air holding, or even flight diversions. There is also an increasing probability of incidents if weather conditions are not correctly anticipated and managed (Schultz et al., 2021; Zhang & Mahadevan, 2019). These events have the potential to jeopardize both safety and efficiency. Weather forecasting represents a fundamental aspect both for airlines and Air Navigation Service Providers (ANSPs), being these latter responsible for ensuring a safe and smooth air traffic management. A correct functioning of airport-related meteorological services allows optimizing airports daily operations as well as supporting decision-making regarding flight routing and planning. These statements are well documented in literature. For example, Von Gruenigen, Willemse, & Frei (2014) proposed a case study at Zurich Airport about the economic benefits of accurate weather forecasting for airlines. Klein et al. (2009) suggested a metric to measure the Weather Impacted Traffic Index Forecast Accuracy (WITI-FA) which was used to evaluate the impact of weather forecasts on the scheduling of an air traffic system from the ANSP perspective.

The International Civil Aviation Organization (ICAO) standards for air navigation meteorological services (ICAO, 2018) are the baseline for airport-related weather management. The use of two main drivers is suggested: observations of actual weather scenarios; and forecasts for future weather conditions. Observations (or measures) are taken with a fixed frequency and called METereological Aerodrome Reports (METARs). METARs are complemented by SPECIs, i.e., special reports that can be emitted at any time. For example, METARs can report weather elements on an hourly basis, but a SPECI can be emitted in between if some weather elements are considered worth to be reported. The notion of weather elements is used to refer to those variables that can be observed (or measured) and can be used to represent weather conditions. Accordingly, METARs and SPECIs can carry information about wind, visibility, meteorological phenomena, clouding, temperature, and pressure, among others. These data are then used to produce the Terminal Aerodrome Forecasts (TAFs) which are previsions for future weather conditions. A TAF shall be issued at a specified time, and it shall consist of a concise statement of the expected meteorological conditions for a specified period (e.g., 4 hours, 8 hours, or even 24 hours), where this latter represents the TAF validity time. The TAFs accuracy, expressed by Key Performance Indicators (KPIs) is a critical measure to assess aerodrome systems performance in making weather forecasting. A KPI compares a TAF (i.e., expected weather) with the actual weather registered during its validity period (i.e., corresponding METARs). These KPIs shall be monitored continuously to gain aerodromes performance. Considering increasing digitalization and turbulent market conditions, KPI monitoring becomes paramount to support decision-makers at addressing both operational and tactical actions.

Besides descriptive historic data analysis, KPIs can be further used to reveal hidden patterns related to the ability of an aerodrome system to forecast accurately weather conditions. While the usage of Artificial Intelligence (AI) techniques has raised an increasing interest over recent years for weather forecasting, limited evidence is available on the usage of AI techniques to support decision-making based on systematic weather KPIs analysis (Gujanatti et al., 2021).

On this path, a novel methodology is presented in this paper, integrating two Machine Learning (ML) algorithms that take advantage of KPIs data. The proposed methodology aims to identify common sets of weather data, to isolate negative performance dynamically through an anomaly detection algorithm, and to inform decision-makers on future estimated accuracy levels basing upon the calculation of an error propensity metric related to clusters of TAFs.

The remainder of this paper is organized as follows. Section 2 reviews available approaches related to the usage of ML for weather forecasting and approaches for the calculation of TAF accuracy and weather KPIs. Section 3 presents the ML methodology, which is then tested into a real operating scenario, described in Section 4. Section 5 adds discussion on obtained results and comments on the methodology application for decision-making. Inherent limitations and perspectives for future research are finally suggested in Section 6.

2. Background

ML benefits are nowadays widespread in multiple domains with impactful consequences. Nonetheless, the literature review of this paper aims to explore research contributions in weather forecasting mainly about the usage of ML techniques in meteorological services. The retrieved contributions are twofold: (i) firstly, documents proposing the usage of ML in the generation of weather elements; (ii) secondly, contributions detailing the usage of ML to assess the impact of weather forecasts at aerodromes in terms of their accuracy and the related losses on airport operations. In addition, considering the scope of the paper, a review of different approaches for forecast accuracy assessment is provided, even if not directly linked to ML. Accordingly, this section contains four subsections detailing the different streams of research in the field of weather forecasting: Section 2.1 identifies relevant ML approaches used to predict weather elements; Section 2.2 reports previous research on assessing the impact of forecasts at aerodromes; Section 2.3 reviews different approaches to evaluate forecast accuracy; Section 2.4. locates this work in previously identified literature.

This manuscript promotes the usage of ML into weather accuracy analysis. The contribution advances an under-developed stream of literature about the implementation of ML decision support systems for accuracy analysis in meteorological services (Gujanatti et al., 2021).

2.1. ML to generate weather elements

The usage of ML solutions for weather forecasting can be categorized in terms of algorithms to be applied, or based upon the set of weather elements to be considered (Jaseena & Kovoor, 2020). Applications can be mainly found in the generation of forecasts: automated systems generate previsions which are subsequently validated by forecasters (or decision makers at different levels) who then emit a weather bulletin (i.e., a TAF). Forecast generators mostly rely on numerical methods that have been integrated recently with ML solutions (Weyn et al., 2021). These approaches constitute the major stream of ML research in weather related problems. Some examples are provided in the following lines concerning the usage of ML to improve specific elements forecasting. Murugan Bhagavathi et al. (2021) suggested a short-term forecast model to integrate numerical weather predictions with ML decision tree and clustering algorithms. Similarly, an hourly temperature prediction tool based on Artificial Neural Network (ANN) was presented by Astsatryan et al. (2021). The ANN is fed with measures from meteorological stations. Combined satellite, lighting, and radar observations were used instead as inputs for a random forest model to predict severe storms by Mecikalski et al. (2021). Another application of random forests was proposed by A. Wang et al. (2021) who used a ML algorithm to adjust numerical wind predictions. A Multiple Discriminant Analysis prediction tool to interpret METARs and generate more accurate TAFs was proposed by Montpetit et al. (2002), yet in very short-range intervals. Almeida, França, & Campos Velho (2020) tested six ML algorithms (i.e. Random Forest (RF); Decision Tree (J48); Multilayer Perceptron (MLP) classifier; Radial Basis Function (RBF) classifier; ensemble of RF, J48, and custom MLP plus RDF classifiers; Deep Learning fully-connected (DL-FC)) to predict storms occurrence and severity from atmospheric discharge data. A deep learning dense algorithm turned out to be the most effective one in terms of weather KPIs, yet for one location at a time. An algorithm to nowcast the occurrence or absence of certain visibility levels and cloud ceiling values for the next hour was presented by Cordeiro, França, Neto, & Gultepe (2021), being trained on Rio de Janeiro airport data. All these examples confirm the benefits in terms of forecasting accuracy: numerical methods are not sufficient to predict weather conditions since they have highest chances to miss complicated patterns and non-linear behaviours (Hennayake et al., 2021).

2.2. ML to assess the impact of forecasts at aerodromes

A second stream of ML research for weather elements refers to the development of methodologies to support the forecaster in deciding about the reliability of the numerical (or ML-driven) bulletins, providing decision support systems to perform the forecasts with higher confidence. When coming to this stream of research, i.e. decision support methodologies, ML has been used so far only to a minor extent. For example, Cristani, Domenichini, Olivieri, Tomazzoli, & Zorzi, (2018) presented a software that updates its output dynamically based on historic data to support decision makers in TAF bulletin generation. Complementarily, ML has been

used to classify impacts of weather forecasting on aerodrome performance (Schultz et al., 2021) in terms of losses related to airport operations. A similar logic has been applied via ML regression trees to predict disruption on airport's arrivals with respect to adverse weather conditions (Y. Wang, 2017). A ML technique based on ANN in conjunction with existing numerical models has been proposed to assist detecting turbulences and weather anomalies (Cai et al., 2019). In a larger management context, Mangortey et al. (2020) proposed a ML based solution to support airport operations management: the approach relies on the prediction of airlines ground stops due to adverse weather and investigation of key factors which contribute to their occurrence.

2.3. Accuracy assessment of weather forecasts

In parallel to the recent introduction of ML, other weather-related research refers to the development of refined performance analysis systems to compare TAFs over their related METARs (i.e., expected weather vs. actual weather). Starting from ICAO directives (ICAO, 2018), several approaches exist in this regard (Sladek, 2021). A relevant example is the work from Austro Control (Mahringer, 2008), who proposed a method based on a twofold verification: the most favourable and the most adverse observed values should be used to assess TAF's score through certain fixed intervals. Starting from there, Sharpe et al. (2016) defined a novel reliability table to measure TAF performance including probabilistic information, and compare it with the deterministic multicategory approach by Mahringer (2008), only for visibility values. Recently, Sladek (2021) focused on setting criteria to enhance previous methods on TAFs weather and cloudiness. A larger scope review on TAFs accuracy assessment and directives for weather forecasting is provided by Sládek (2019).

Nevertheless, in order to measure systemic forecasting performance, it is necessary to move the scope of analysis from the difference between a single TAF and its corresponding set of METARs, to the accuracy of a set of TAFs. For example, Karel Dejmal, Novotny, & Hudec (2015) evaluated the TAF accuracy looking at the values of TAF wind speed, wind direction and some selected meteorological phenomena on a 24 hours interval. The results showed correlation between the successfulness of predictions performance and time, also highlighting the occurrence of many formal errors in TAF strings at specific hours of the day. Similarly, Novotny et al. (2021) instantiated data pre-processing and accuracy calculation on 5 Czech airports. Analysing weather performance in Czech Republic, K. Dejmal & Novotný (2018) developed an algorithm to assess TAF reliability based on the numerosity of errors and their main features (e.g., which weather element is not correct, TAF structure, time discrepancy). The study was not limited to a single day of operations, but TAFs emitted in Czech stations from 2011 to 2017 were considered, showing major criticalities and hidden correlations between errors made in TAF and the period of emission.

2.4. Locating this study

Figure 1 summarizes the discussed literature, and places the contribution of this work into a simplified functional mapping of the weather forecasting process. Traditionally, an automatic forecast generator system uses METARs and SPECIs to propose previsions for future weather. These previsions are directed to a forecaster who has to decide whether the proposed TAF can be emitted or not, possibly modifying it. Aerodrome operations are planned based on these forecast bulletins. The three main streams of research previously documented from literature (see Section 2.1-2.3) aim to improve this traditional process in different ways. Some works propose to integrate or substitute numerical forecasting generation approaches with ML solutions (blue path in Figure 1, discussed in Section 2.1). Consequently, the process improves by limiting errors of the decision maker delivering an already precise prevision. The second research stream instead focuses on the development of decision support systems to use weather information strategically for supporting TAF bulletin generation (green path in Figure 1, Section 2.2). The third group of papers suggests innovative methods to evaluate forecasts accuracy and guide their redaction by improving the knowledge on overall system performances (block with red border in Figure 1, Section 2.3).

The current manuscript aims to contribute to the definition of decision support systems for weather forecasts which are based on performance evaluations metrics by designing a ML-based methodology that leverages on historical data and provide systematic indexes for decision making at forecaster level.

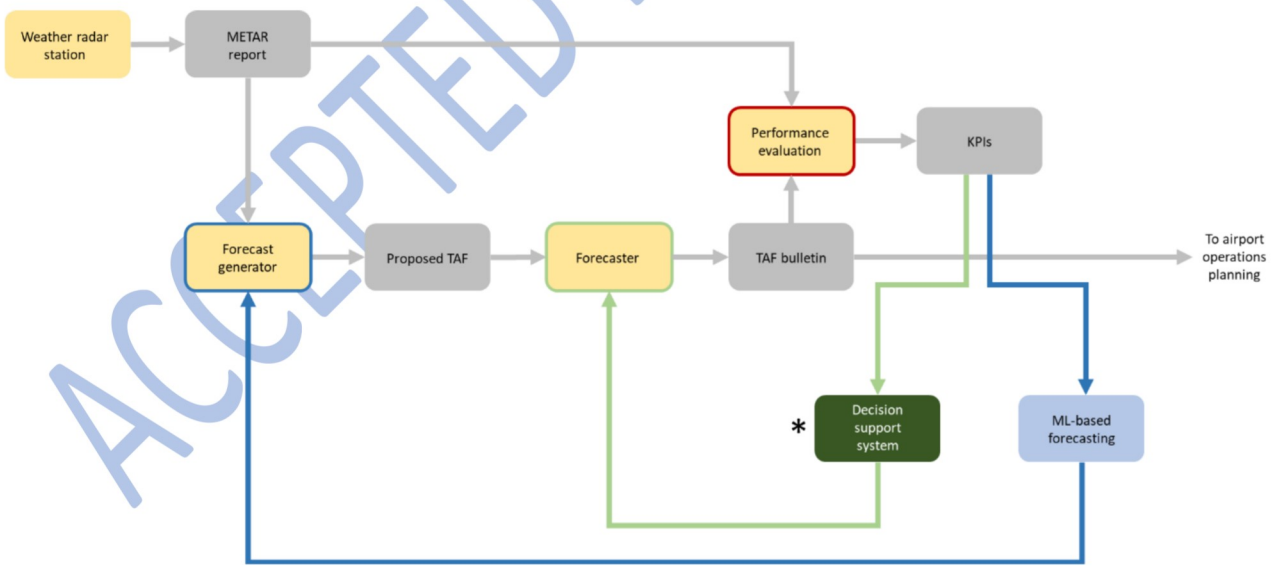


Figure 1. Summary of traditional process for weather forecasting, and improvements suggested in literature: blue elements refer to ML used to support elements generation, green elements refer to ML used to support decision-making through a supporting reasoner, and orange elements refer to approaches used to improve accuracy calculation. The contribution in this manuscript suggests a novel decision support system based on ML to improve forecaster decision capacities, marked with the (*).

3. Materials and methods

ML focuses mainly on three main approaches: (i) descriptive analyses to transform data into information, (ii) predictive analyses to transform information into decisions, (iii) prescriptive analyses to transform decision into actions (Nakhal A et al., 2021a). The proposed methodology spans over descriptive and predictive analytics, as summarized in Figure 2.

Historic data are firstly pre-processed in order to enable descriptive analysis: METARs and TAFs must be prepared to be comparable for subsequent accuracy evaluation. These processed data constitute a data mart which serves as an input for ML algorithms. Two ML algorithms are applied in the methodology. At first, the time series of a selected KPI feeds an anomaly detection algorithm to spot abnormal performance. Secondly, a parallel study on TAF records is conducted to define clusters them via a dedicated ML algorithm considering forecasted weather elements as clustering features. At a final stage, the anomalies are used to predict the potential for forecasting mishaps via the definition of an error propensity metric that consider clusters. The comparison between the results of the anomaly detection and the clustering algorithm suggests the propensity of certain types of TAF (both historic, and future) to be characterized by performance anomalies. These results can be used in a predictive perspective to support the forecasting process.

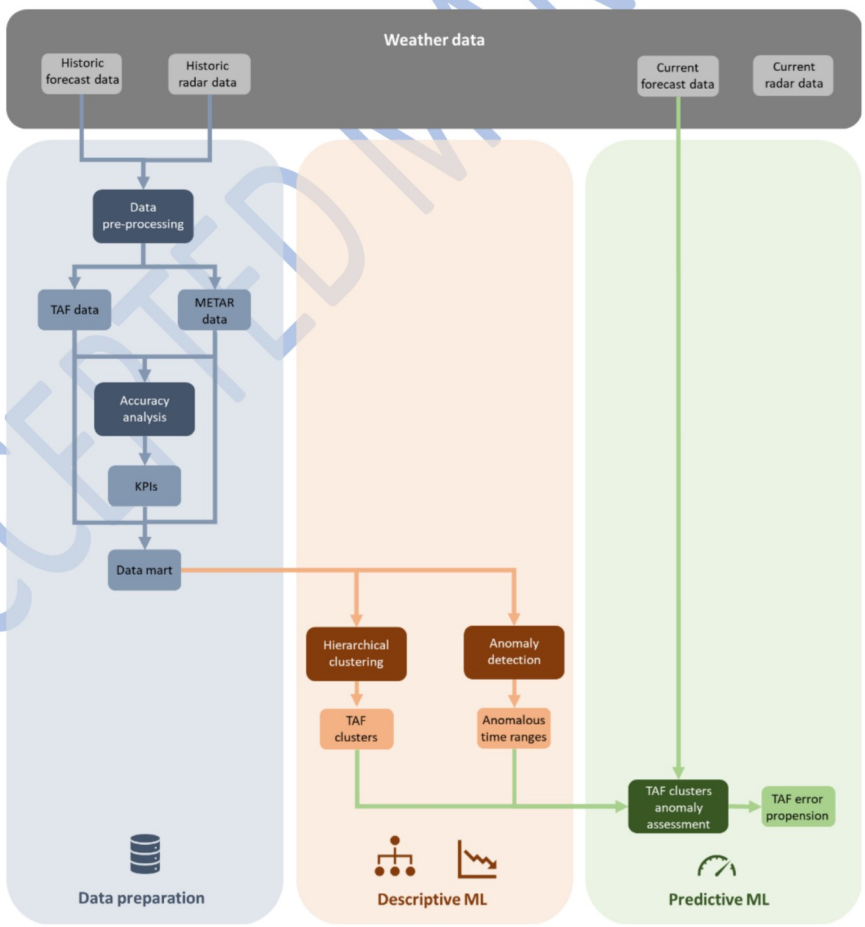


Figure 2. Decision support methodology for aerodrome weather forecasting.

3.1. Data Pre-Processing

The final goal of the descriptive step is to extract information from available data and store them into a data model, which is the core for all the subsequent analyses. The resulting data model comprehends data relationships, dimensions, and measures within the processed data, obtained via Extraction-Transformation-Loading (ETL) (Nakhal A et al., 2021b).

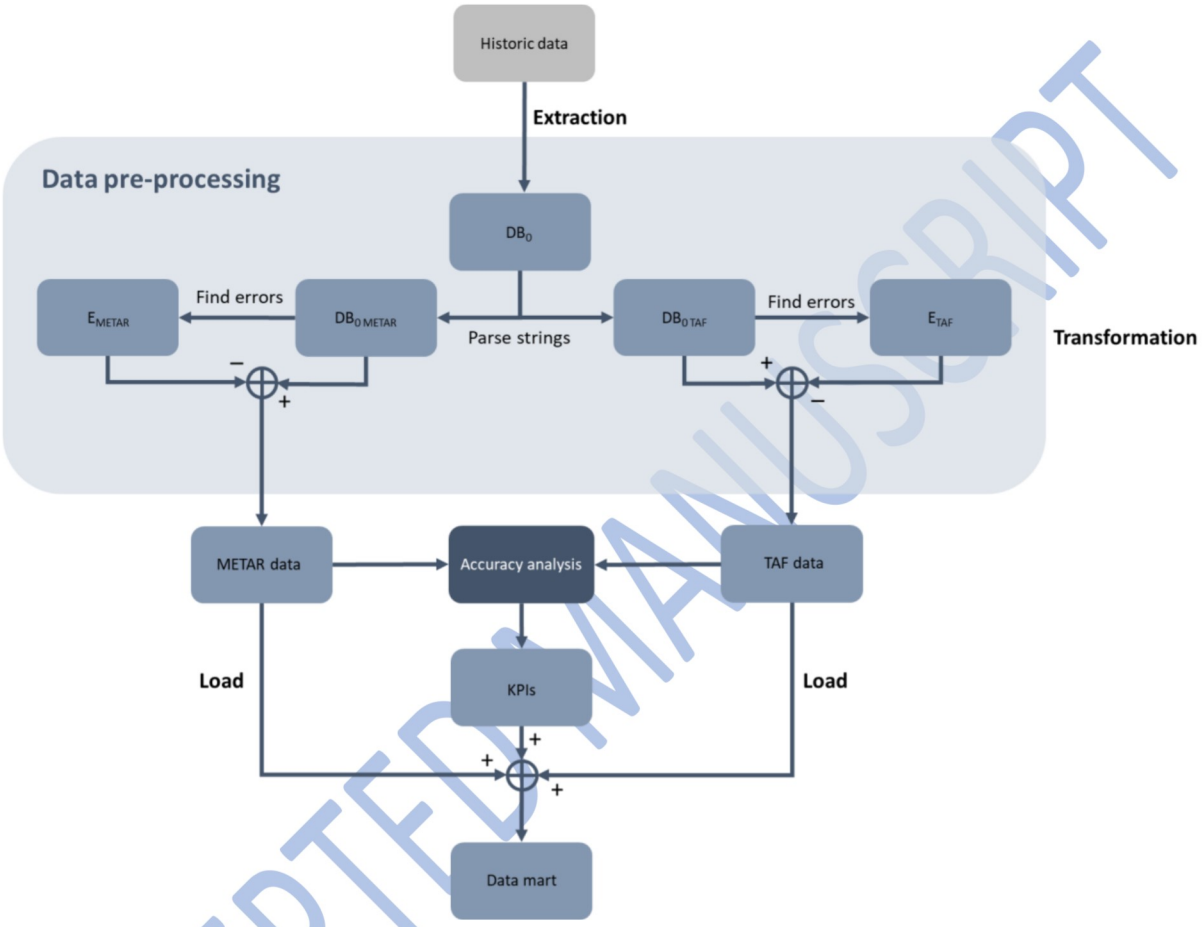


Figure 3. Data pre-processing steps (ETL).

3.1.1. Extraction

Extraction refers to the actions to acquire data from the systems that collect them, e.g. sensors on aerodrome measuring stations and forecast bulletins collection. The weather forecasting process makes use of data taken from measuring stations to produce TAFs. METARs are produced on an hourly or semi-hourly base and made available to forecasters in addition to SPECI, i.e., reports emitted under special circumstances. On the other hand, TAFs can be produced at different time steps, (e.g.) every 4 hours or 8 hours. Both METARs and TAFs are stored in a dedicated database, whose main fields are reported in Table 1. With reference to Figure 3, DB_0 represents the database obtained after the extraction of data from the historic records.

Table 1. Weather database structure.

Field	Description
ID	Unique identification code for the record
KIND	Type of record (METAR, SPECI, TAF)
AD	Location at which the record was registered (aerodrome)
TEXT	Alphanumerical string coded as (ICAO, 2018) to carry information about weather
TIME STAMP	Date and time at which the record is stored in the database
NOTE	Optional additional notes

3.1.2. Transformation

Transformation refers to all the actions to transform data through queries. These can involve: (e.g.) re-organizing data in a new format, remove duplicate records, joining data from multiple sources, aggregating or disaggregating records, etc. The data mart is obtained through a set of queries applied to DB_0 database (cf. Figure 3). Since METARs and TAFs contain all weather information into a single string field, parsing actions are needed to split each string and isolate the relevant elements. In Figure 4 an exemplary parsed string is shown along with the input and output data structures of the pre-processing step. The parsing process allows obtaining two subsets of DB_0 (Figure 3): $DB_{0\,TAF}$ containing forecasts; $DB_{0\,METAR}$ containing observations (both METARs and SPECIs). Custom queries are developed to spot syntactical errors (Dejmal & Novotný, 2018) of reports in both $DB_{0\,TAF}$ and $DB_{0\,METAR}$ and isolate the corresponding strings, then stored in E_{TAF} and E_{METAR}. These latter are subtracted respectively from $DB_{0\,TAF}$ and $DB_{0\,METAR}$ to obtain the TAF data and the METAR data to be used for further analyses, which contain only data with no errors. After the data pre-processing stage, the data table has a structure as the one presented in Figure 4. Specifically, every record in the input table (among the ones which have not been detected as errors) contain the "TEXT" field split to isolate information inside it consisting of observed/forecasted weather elements and additional information (e.g., record type, airport of collection). In this regard, some important fields to be discussed are the ones related to the validity time of both observations (METARs and SPECIs) and forecasts (TAFs): through the "From day", "From hour", "To date", and "To hour" fields each row from the data table can be decomposed at a fixed time granularity enabling comparison between records.

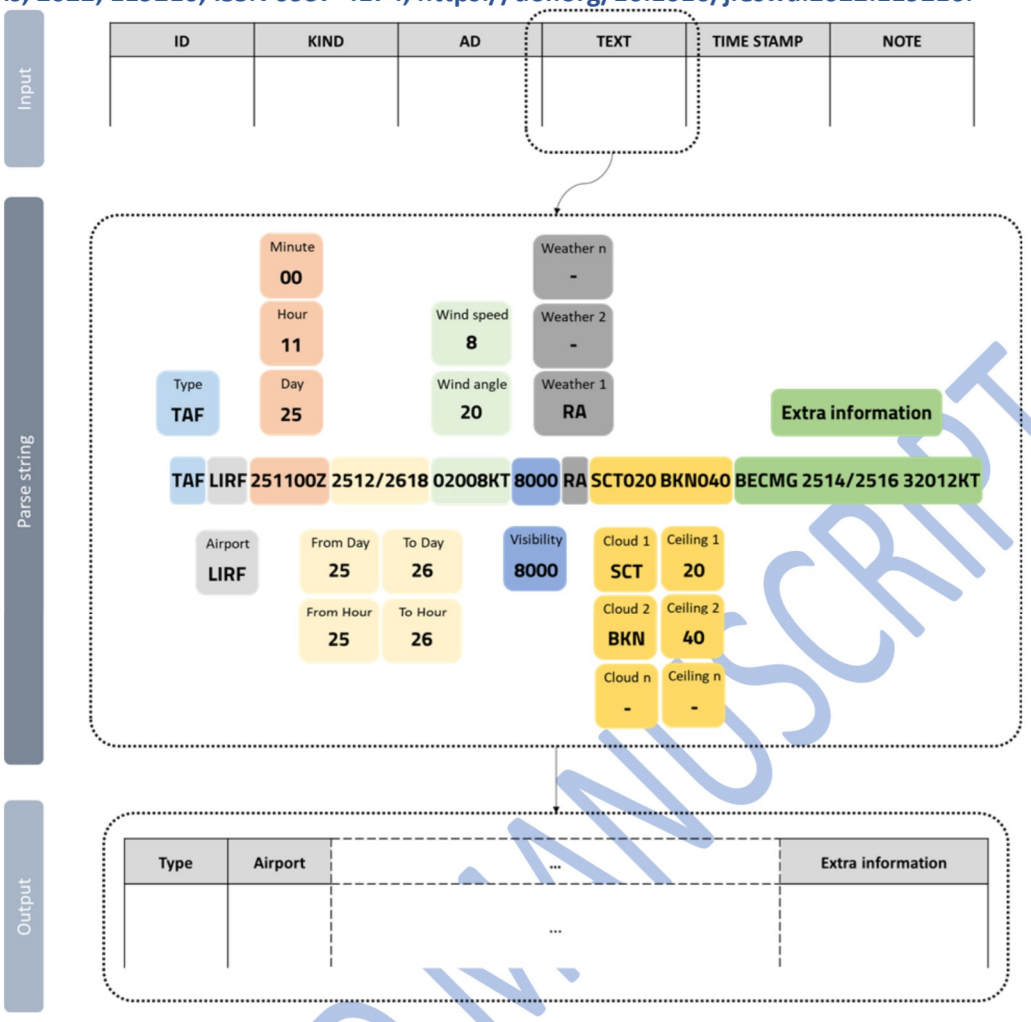


Figure 4. Example of a string data parser for an exemplary TAF.

3.1.3. Loading

Loading consists in importing the data mart into the target system to proceed with reporting and advanced analytics. Pre-processed data are loaded into the data mart which represents the starting point for following analyses.

3.2. Accuracy analysis

The accuracy analysis is used to define and compute KPIs to compare TAFs over METARs and SPECIs. ICAO recommends seven TAF elements to be compared with corresponding METARs and SPECIs ones (ICAO, 2018): wind direction (*ddd*), wind speed (*ff*), visibility (*VVVV*), precipitations and weather phenomena (*ww*) clouds' type (*NsNsNs*), clouds' ceiling (*nsnsns*), and air temperature (*TT*). The KPIs rely on diverse binary contingency matrices (i.e., presence or not presence, in or out a fixed threshold) as in Table 2.

Table 2. Contingency table for binary weather elements.

		Event observed		
		Yes	No	Marginal totals
Event forecasted	Yes	*a*	*b*	*a* + *b*
	No	*c*	*d*	*c* + *d*
	Marginal totals	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

With reference to Table 2, *a* represents the occurrences in which a forecasted event has been observed (Hit), *b* the ones in which an event has been forecasted but not observed (False alarm), *c* the ones in which an event has been observed but not forecasted (Miss), and *d* the ones in which an event has not been observed neither forecasted (Correct rejection). When contextualizing contingency matrices for weather KPIs, it is not necessary meaningful to consider all combinations sketched in Table 2, (e.g.) false alarm and correct rejection make no sense for wind direction or wind speed evaluations (ICAO, 2018). The individual contingency criteria recall ICAO Annex 3 regulation and are described in depth in Annex A of this manuscript, Table 1A, whose main aspects are reported below:

- Wind direction accuracy is calculated based on the difference between the forecasted direction and the observed one compared to a certain threshold (20°).
- Wind intensity accuracy is calculated based on the difference between the forecasted intensity and the observed one compared to a certain threshold (5 kt).
- Horizontal visibility values are categorized in two classes. The accuracy control differs depending on the horizontal visibility: if is less than 800 m or above.
- Weather phenomena and precipitations are reported through codes (e.g., RA for rain or TS for thunderstorm) and then evaluated only by their occurrence or not.
- Predictions on clouds are evaluated by considering the height of the bottom clouds layer. Moreover, this latter has to refer to broken clouds (5 to 7 okta) or overcast clouds (equal or more than 8 okta) with a ceiling less than 1500 m. If the forecast does not satisfy these hypotheses, it is always considered correct. In all other cases, the predicted value is controlled through its occurrence or not.
- Clouds ceiling (i.e., height) is evaluated, as for visibility, in two classes. The accuracy control differs depending on the weather forecasted ceiling: if is less than 300 m or above.

- Temperature evaluation is based on the difference between its forecasted value and the observed one. Temperature KPIs rely on binary contingency matrices with a threshold of ±1° admitted variation. Temperature accuracy verification is suggested only for very long-term previsions (Chan & Li, 2003), and it will not be considered in this paper in following KPIs analyses.

3.2.1.Management of TAF change groups

KPIs are calculated by reporting METAR, SPECI, and TAF data at a certain time resolution related to a specific time step ΔT. This discretization allows comparing forecasted weather elements against observations (METARs and SPECIs). From an operational perspective, TAF elements remain valid for the declared overall time validity of the TAF itself. Nevertheless, these elements can be further refined, adding extra forecasting group(s) in a TAF string to document expected significant changes. These groups are called "change groups" and they can be of four different types (World Meteorological Organization, 2017):

- From group (FM): it is used to set a change in weather element acting from a specific time moment, until the end of the TAF validity. A change in elements inserted in the FM group completely substitute the ones in the main forecast.
- Becoming group (BECMG): it is used to insert a transition period within the TAF. The change in one element inserted in the BECMG group coexists with the ones in the main forecast. Once the validity of the BECMG group expires, the elements in the BECMG group substitute the ones in the main forecast.
- Temporary group (TEMPO): it is used to indicate temporary fluctuations in the TAF. Elements in the TEMPO group are valid together with the ones in the main forecast only for the time validity of the TEMPO period.
- Probability group (PROB): it is used to assign probabilities of a change in weather elements. The PROB indicator is considered out of scope in this work.

Further information about the management of TAF change groups are available in Annex B.

3.2.2.KPI definition

KPIs are calculated relying on a, b, c, d frequency as obtained from binary and multi-variate contingency matrixes. Common KPIs that are widely used to evaluate aerodrome forecasts are summarized in Table 3 (Roebber, 2009).

Table 3. Main KPIs for weather forecasting accuracy.

KPI	Acronym	Analytical expression
Frequency Bias Index	FBI	$\text{FBI} = \frac{a+b}{a+c}$ (1)
Proportion Correct	PC	$\text{PC} = \frac{a+d}{a+b+c+d}$ (2)
Critical Success Index	CSI	$\text{CSI} = \frac{a}{a+b+c}$ (3)
Probability Of Detection	POD	$\text{POD} = \frac{a}{a+c}$ (4)
False Alarm Ratio	FAR	$\text{FAR} = \frac{b}{a+b}$ (5)

KPIs in Table 3 can be calculated referring at any contingency matrix since they are not dependent from the weather element under analysis, and also they can be updated following changes related to change groups.

The Frequency Bias Index (FBI) is the ratio between the total number of events forecasted ($a + b$) over total number of events observed ($a + c$). FBI evaluates whether an event has been overestimated or underestimated: FBI = 1 is the perfect score, FBI < 1 depicts underestimation, FBI > 1 depicts overestimation.

The Proportion Correct (PC) is the ratio between total number of correct forecasted events (also considering correct rejections) over total number of events observed and forecasted. It ranges between 0 to 1 and PC = 1 represents the perfect score. It is highly affected by the presence of a non-forecasted/non-observed common event.

The Critical Success Index (CSI) is the ratio between the number of correct forecasted events over total number of events observed and forecasted (without considering correct rejections). It ranges between 0 to 1 with CSI = 1 perfect score.

The Probability of Detection (POD) is the ratio between the number of correct forecasted events over the total number of events observed. False alarms are not considered in POD. POD ranges between 0 to 1 with POD = 1 perfect score.

The False Alarm Ratio (FAR) is the ratio between the number of forecasted events that are not observed (false alarms) over total number of events being forecasted. It ranges between 0 to 1 with FAR = 0 perfect score.

The KPIs calculated at atomic level (i.e., for each time step ΔT) are loaded into the data mart increasing the pool of data to be used in the subsequent ML-driven analysis (cf. Figure 1).

3.3. ML solutions for the calculation of the TAFs error propensity metric

As per the aim of this paper, two ML algorithms are involved to support aerodrome weather forecasting decision-making: anomaly detection (section 3.3.1), and clustering (section 3.3.2). The algorithms have been selected to reproduce the functional workflow of a systemic accuracy analysis: firstly, to detect anomalies in historic forecast accuracy levels, then to find commonalities in the data contributing to these anomalies, and lastly, to integrate these findings for anticipating future anomalies.

3.3.1. Anomaly detection

An anomaly detection algorithm permits to isolate abnormal values within a time series. Given a sequence of real values $x = x_{t'} = (x_0, x_1, \dots, x_T)$ with $t' \in [0, T]$; an anomaly detection algorithm aims to produce an output sequence of corresponding values $y = y_{t'} = (y_0, y_1, \dots, y_T)$ with each $y_{t'} \in [0, 1]$ that denotes whether the corresponding $x_{t'}$ is an anomaly point or not (Ren et al., 2019), i.e. $y_{t'} = 1$ denotes anomaly in $x_{t'}$, $y_{t'} = 0$ depicts non-anomaly in $x_{t'}$. As real-world data can generate many different types of time series with different characteristics (e.g. seasonal, stable, etc.), it is hard to develop a generalized algorithm that deals with all these situations efficiently (Ren et al., 2019). Saliency, i.e., the property by which something stands out, can represent a potential solution for these problems. Accordingly, the proposed algorithm makes use of Spectral Residual (SR) to quantify the difference between data points in the frequency domain, and subsequently it provides a function for saliency in the spatial domain (Hou & Zhang, 2007). A convolutional neural network (CNN) is then applied on the results produced by the SR to dynamically define a threshold rule to decide whether a point should be considered anomalous of not (Zhao et al., 2015). The input of the algorithm is a time series x. While the time series to feed the algorithm can be any of the KPIs, the POD is selected since it is the only KPI that can be calculated for any weather elements (for example, *ddd* and *ff* only have Hit and Miss values from their respective accuracy rules, cf. Annex A).

The KPI represents a key point to apply the proposed methodology as it permits to spot time frames in which anomalous operations happened. In this regard, the following lines presents some steps to aggregate the KPI values of multiple weather elements in a unique time series to be processed by the anomaly detection algorithm.

Every TAF can be defined by four dimensions:

$$TAF(V^s, V^e, R, AD) \tag{6}$$

where V^s and V^e are respectively TAF validity start time and TAF validity end time, R represents the richness of the string in terms of how many different types of weather elements it contains, and AD is TAF emission location, i.e., the aerodrome for which the forecast applies. $TAF_{u_{AD_t}}$ represents a TAF referred to a time step t included in a period of analysis T, from an emission location AD.

A time resolution $\Delta T'$ for the time series represents the time period between two recorded points of x. Notice that $\Delta T'$ can be set equal to ΔT, or differs from it, since ΔT represents the time resolution of the accuracy analysis:

$$\Delta T' = \delta \cdot \Delta T, \quad \delta \in \mathbb{R} \tag{7}$$

For example, a yearly analysis will imply setting T = 365 days = 8,760 hours. In this case, one may want to have a daily anomaly detection analysis ($\Delta T' = 1\ day = 24\ hours$) even though METAR and TAF data are collected with an hourly frequency ($\Delta T = 1\ hour$). Accordingly, δ = 24 so that the desired time series has 365 data points rather than 8760. All calculations for accuracy analysis are firstly run on the more granular time step, i.e., ΔT, and eventually aggregated to follow the larger time step, i.e., $\Delta T'$.

Once set the time resolution $\Delta T'$, It is possible to define the number of time units $TAF_{u_{AD_t}}$ covers as:

$$N_{u_{AD_t}} = \frac{\left[V^e_{u_{AD_t}} - V^s_{u_{AD_t}}\right]}{\Delta T} \tag{8}$$

Note that $N_{u_{AD_t}}$ is expressed in terms of how many ΔT the TAF covers. Accordingly, for each of the $N_{u_{AD_t}}$ time step of extension ΔT, a METAR will be valid, too, permitting the calculation of the KPI (POD in this case). Consequently, $N_{u_{AD_t}}$ is a measure of how many values have been obtained from the accuracy analysis. The POD index of a $TAF_{u_{AD_t}}$ can be indicated in formal terms (left), and with simplified notation (right) as:

$$POD_{AD_t,\, u_{AD_t},\, r_{u_{AD_t}},\, n_{r_{u_{AD_t}}}} \rightarrow POD_{AD,u,r,n} \tag{9}$$

where:

- AD_t identifies an aerodrome location. It is a time dependent index since the number of TAFs, their elements, and their validity change over time with a time resolution equal to ΔT (t increments every ΔT). In the simplified notation, hereafter it will be referred as AD.
- u_{AD_t} identifies a TAF for a time t (t increments every ΔT), among the ones emitted per a certain AD. In the simplified notation, hereafter it will be referred as u.
- $r_{u_{AD_t}}$ identifies one of TAF weather elements. This index depends on the TAF to be examined and, obviously, on the corresponding AD and time t. In the simplified notation, hereafter it will be referred as r.
- $n_{r_{u_{AD_t}}}$ identifies a time step within a certain $TAF_{u_{AD_t}}$ validity. This index depends on each TAF element to be examined within a $TAF_{u_{AD_t}}$ for a time t, among the ones emitted per a certain AD. In the simplified notation, hereafter it will be referred as n.

The average value of POD of each element for a time interval N_u covered by a TAF_u can be computed as:

$$POD_{AD,u,r} = \frac{\sum_{n=1}^{N_u} POD_{AD,u,r,n}}{N_u} \tag{10}$$

Six $POD_{AD,u,r}$ can be calculated with reference to *ddd*, *ff*, *VVVV*, *ww*, *NsNsNs*, and *nsnsns*. POD of an entire TAF forecast can be computed as their average value:

$$POD_{AD,u} = \frac{\sum_{r=1}^{R} POD_{AD,u,r}}{R} \tag{11}$$

Since *ww*, *NsNsNs* and *nsnsns* are not mandatory information to be contained in a TAF, $R \in \{3,4,5,6\}$. In this paper, TAF accuracy is calculated through mandatory elements, i.e., $R = 3$.

A graphical representation of $POD_{AD,u}$ is presented in Figure 5a. Notice that multiple TAFs can co-exist at the same time, as they are progressively released by the forecaster without any prescription on the emission frequency. Accordingly, the POD index for a given time interval (based on fixed time granularity ΔT) aggregating multiple $POD_{AD,u}$ is defined as:

$$POD_{AD} = \frac{\sum_{u \in U_{AD}} POD_{AD,u}}{|U_{AD}|}; \quad U_{AD} : \{t \mid t \geq V_u^s \ \wedge \ t \leq V_u^e\} \tag{12}$$

where U_{AD} represents a set of TAFs that are valid in any given time step t, $|U_{AD}|$ is the cardinality of the set, and V_u^s and V_u^e represent the simplified notation for respectively $V_{u_{AD_t}}^s$ and $V_{u_{AD_t}}^e$. An exemplary representation of POD_{AD} is proposed in Figure 5b for $t \in [0, \ T]$. Notice that both U_{AD} and $|U_{AD}|$ change over time but they are referred to a time resolution ΔT. A final step is needed to transform POD_{AD} to the time scale desired for the anomaly detection analysis $\Delta T'$.

$$\overline{POD}_{t'} = \frac{1}{\delta} \sum_{t=\delta t'-1}^{\delta(t'+\Delta T)-1} POD_{AD}, \quad t' \in [0, T] \tag{13}$$

This last step enables a wider view of the anomaly occurrence by reporting the anomalous time moment and labelling as anomalous also nearest POD values. $\overline{POD}$ timeseries (cf. Figure 5c) represents the input time series x for the anomaly detection algorithm.

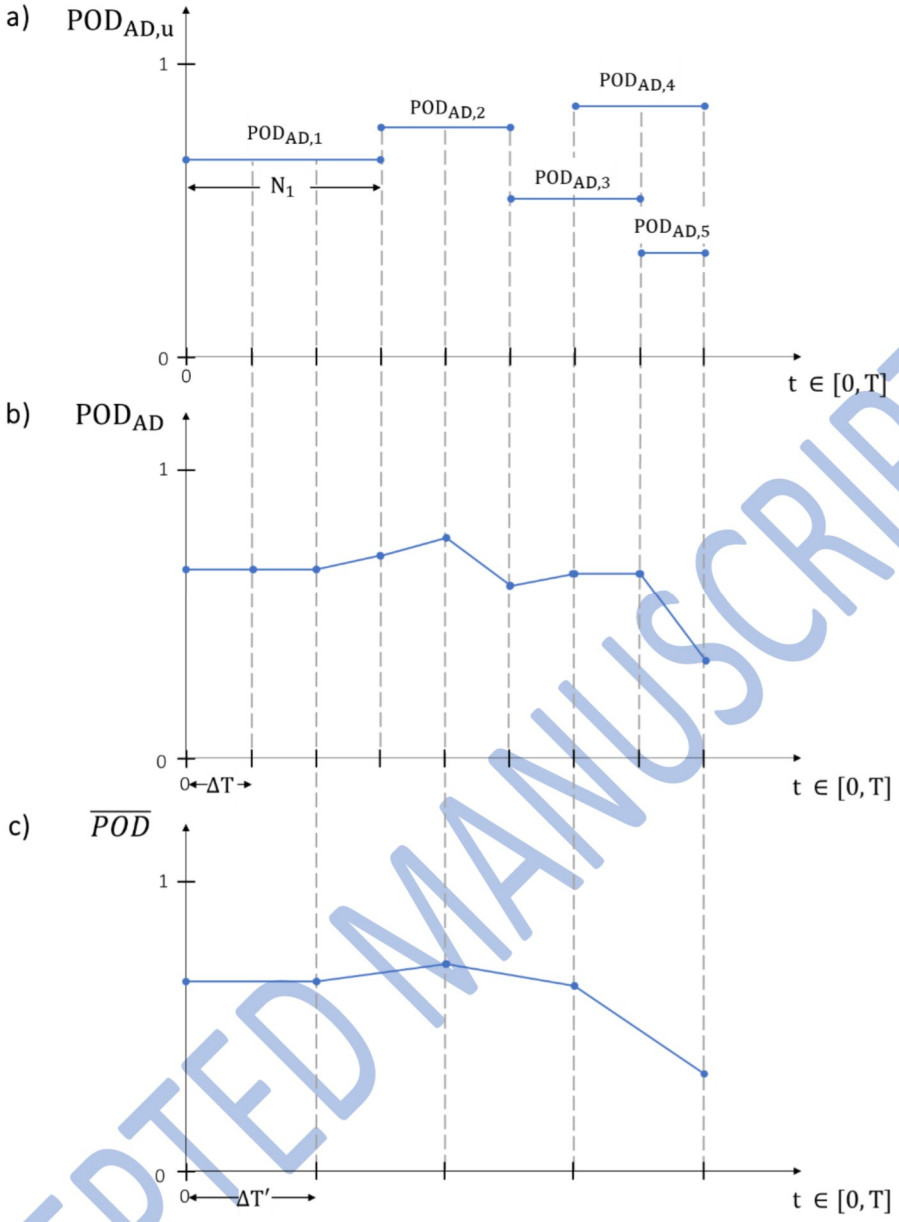


Figure 5 a) POD value for multiple TAFs over time. b) Exemplary aggregated POD time series with accuracy analysis time resolution. c) Exemplary aggregated POD time series with anomaly detection algorithm time resolution.

On the input time series POD_t, , SR is applied by: (i) computing the Fourier Transform $\mathcal{F}$ of the series to then get the logarithmic amplitude spectrum $LA(f)$, and the averaged logarithmic spectrum $ALA(f)$; (ii) calculating the spectral residual $SR(f)$, (iii) computing the Inverse Fourier Transform $\mathcal{F}^{-1}$ to return the sequence back to spatial domain and to obtain the saliency function $S(x)$. The following variables are defined to proceed with the algorithm:

$$f = \mathcal{F}(x) \tag{14}$$

$$LA(f) = \log\left(Amplitude(f)\right) \tag{15}$$

$$P(f) = Phase(f) \tag{16}$$

$$ALA(f) = h_n(f) \cdot A(f) \tag{17}$$

where $h_n(f)$ is a $T \times T$ matrix (T is the length of x, i.e., $\overline{POD}_{t\prime}$) to convolute the input sequence x and it is defined as:

$$h_n(f) = \frac{1}{T^2}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \tag{18}$$

Then the spectral residual is calculated:

$$SR(f) = LA(f) - ALA(f) \tag{19}$$

Finally, the saliency function is computed:

$$S(x) = \left\| \mathcal{F}^{-1}(\exp(SR(f) + \sqrt{-1} \cdot P(f))) \right\| \tag{20}$$

The obtained transformed time series shows more significant anomaly points since the input is now normalized to behave as a stable variable. At this stage, by fixing a threshold-based rule, it is possible to compute the output sequence in terms of the time series y. The rule is verified within a sliding window of the time series comparing points with their neighbours. The rule remains however fixed for the entire time series, ignoring potential localized trends. To enhance this process, a CNN is applied on the saliency function to dynamically establish a more sophisticated decision rule to modify the traditional single threshold adopted by the SR solution. A discriminative model is trained on synthetic data that are generated by injecting new anomaly points (not included in the evaluated data) into the saliency map of $\overline{POD}_{t\prime}$ (i.e., $S(x)$). The injected points are labelled as anomalies while the others are labelled as normal, resolving the problem of the availability of large-scale labelled data. In practice, the CNN model selects a set of points within the time series. Based on them, it calculates the injection value, and then it gets the saliency function. The values of anomaly points are calculated by:

$$x_{t\prime} = (\bar{x} + \mu) \cdot (1 + \sigma(x)) \cdot r + x_{t\prime}$$

where $\bar{x}$ is the average of the points preceding the generic point $x_{t\prime}$, μ and $\sigma(x)$ are respectively the mean and the variance of all the points within the sliding window, and r is a randomly sampled value that can be equal to 0 or equal to 1. This approach permits the anomaly detector to be adaptive to the changes in time series distribution, without needing any manually labelled data. The CNN architecture consists of: (i) two 1-D convolutional layers with filter size equal to the size of the sliding window, and channel size equal to the size of the sliding window (for the first layer) and the double of the size of the sliding window (for the second

layer); (ii) two fully connected layers stacked before the Sigmoid output. As loss function to be minimized to improve model accuracy, it has been used the cross entropy. The training process is based on a Stochastic Gradient Descendent (SGD) optimizer (Sun et al., 2010).

3.3.2. *Hierarchical clustering*

Hierarchical Clustering (HC) is a method to compute clusters of data following a hierarchical representation. Data to be clustered can be seen as a set of M-dimensional observation vectors $\overrightarrow{o_u}$:

$$\vec{o} = \overrightarrow{o_u} = (o_{u1}, o_{u2}, \dots, o_{uM})$$

where $o_{u1}, \dots, o_{uM}$ represent the M coordinates (i.e. TAF weather elements) of each observation $\overrightarrow{o_u}$ (i.e. a TAF) used to cluster multiple $\overrightarrow{o_u}, u \in U_{AD_T}$, where U_{AD_T} represents the total number of TAFs emitted in a location AD for the whole period of analysis T.

In this work, M = 10 to map the weather elements described in Annex C, Table 1C. It is worth noticing that some coordinates $o_{um}, m \in [1,\ M]$ can be null, since (e.g.) wind gust, precipitations, or clouds can be omitted in a TAF (World Meteorological Organization, 2017). These null values are filled with zeros for numerical values (e.g., *fmfm*) and with "N/A" for categorical ones (e.g., *ww*). These substitutions are actioned to allow clustering data also in case of no information. For example, if no weather phenomena *ww* has been reported in the TAF, it is implicitly said that weather is expected to be calm (e.g. no thunderstorm, snow, rain), and this information becomes a valuable feature to consider when clustering TAFs by similarity.

Subsequently, categorical variables are pre-processed with one-hot encoding to make them continuous. Data are normalized to improve efficiency of the clustering algorithm (Ah-Pine, 2010).

The hierarchy dendrogram is obtained computing the matrix of distances between normalized observations in terms of cosine distance. This latter is equal to the complement of the angle cosine between observations vectors in the M-dimensional space. With normalized vectors:

$$d_{\overrightarrow{o_i}\,\overrightarrow{o_j}} = 1 - \cos\left(\theta_{\overrightarrow{o_i}\,\overrightarrow{o_j}}\right) = 1 - \frac{\overrightarrow{o_i} \cdot \overrightarrow{o_j}}{\|\overrightarrow{o_i}\| \, \|\overrightarrow{o_j}\|} = 1 - \frac{\sum_{m=1}^{M} o_{im} o_{jm}}{\sqrt{\sum_{m=1}^{M} {o_{im}}^2 \cdot \sum_{m=1}^{M} {o_{jm}}^2}} \quad (21)$$

$$i, j \mid \{i, j\} \in U_{AD_T}, i \neq j$$

Clusters have been agglomerated following a bottom-up approach: every data is initially put in its own cluster, and then clusters are merged progressively moving up into the hierarchy. To achieve this target, once the distance matrix is computed, the distance between two or more observations is calculated through the Ward linkage criterion which is a popular yet general criterion to perform hierarchical clustering (Horne et al., 2020). The Ward linkage criterion is implemented via the Lance-Williams recursive algorithm (Murtagh & Legendre, 2014). At the first iteration, the two clusters characterized by minimum distance are merged (C_I, C_J). Then, at

each subsequent iteration the algorithm's goal is to minimize the distance variance within clusters. Setting C_K as a third cluster to be merged, the updated cluster distance between $C_I \cup C_J$ and C_K is computed recursively through:

$$d_{(C_I \cup C_J)C_K} = \alpha_{C_I} d_{C_I C_K} + \alpha_{C_J} d_{C_J C_K} + \beta d_{C_I C_J} + \gamma \left| d_{C_I C_K} - d_{C_J C_K} \right| \quad (22)$$

with:

$$\alpha_{C_I} = \frac{s_{C_I} + s_{C_K}}{s_{C_I} + s_{C_J} + s_{C_K}}; \quad \alpha_{C_J} = \frac{s_{C_J} + s_{C_K}}{s_{C_I} + s_{C_J} + s_{C_K}}; \quad \beta = \frac{-s_{C_K}}{s_{C_I} + s_{C_J} + s_{C_k}}; \quad \gamma = 0 \quad (23)$$

where s_{C_I} , s_{C_J} and s_{C_K} refer to the respective number of observations included in clusters *I*, *J* and *K*, i.e. their size.

A critical parameter for HC is the number of clusters to be used as a stop criterion. The average silhouette score is frequently used for this purpose (Lin et al., 2022; Nakhal A et al., 2021b). This latter is a measure of how much an object is resembling in its own cluster in comparison with the others. Silhouette score ranges between –1 and +1 (being this latter the optimal value) and it is computed as:

$$\varphi(\overrightarrow{o_i}) = \begin{cases} 1 - A_{\overrightarrow{o_i}}/B_{\overrightarrow{o_i}}, & if\ A_{\overrightarrow{o_i}} < B_{\overrightarrow{o_i}} \\ 0, & if\ A_{\overrightarrow{o_i}} = B_{\overrightarrow{o_i}} \\ B_{\overrightarrow{o_i}}/A_{\overrightarrow{o_i}} - 1, & if\ A_{\overrightarrow{o_i}} > B_{\overrightarrow{o_i}} \end{cases} \quad (24)$$

with $A_{\overrightarrow{o_i}}$ being the mean distance between observation $\overrightarrow{o_i}$ and all other observations in its own cluster C_I:

$$A_{\overrightarrow{o_i}} = \frac{1}{s_{C_I} - 1} \sum_{j \in C_I, i \neq j} d_{\overrightarrow{o_i}\,\overrightarrow{o_j}} \quad (25)$$

and $B_{\overrightarrow{o_i}}$ being the minimum of the mean distances between observation $\overrightarrow{o_i}$ and all other observations in each cluster except its own:

$$B_{\overrightarrow{o_i}} = \min_{C_{I'} \neq C_I} \frac{1}{s_{C_{I'}}} \sum_{j \in C_{I'}} d_{\overrightarrow{o_i}\,\overrightarrow{o_j}} \ \forall\, I' \neq I \quad (26)$$

The optimal number of clusters *Nc* is the one that maximizes $\bar{\varphi}(Nc)$, which is the average of the silhouette score $\varphi(\overrightarrow{o_i})$ of all observations $\overrightarrow{o_i}$ for a given number of clusters *Nc*.

The obtained clusters permit to classify (as per the calculated likelihood) past and future TAFs on the basis of their weather elements. This classification enables to identify groups of similar TAFs in terms of structure and content.

3.4. TAF clusters anomaly assessment

The two ML outputs (i.e., anomalous time steps, and clusters of similar TAFs) are combined to assess the tendency which characterizes a TAF to generate an anomaly in the POD index.

Specifically, from the anomaly detection algorithm (cf. Section 3.3.1), the set of anomalous time steps of the POD time series can be identified as:

$$T^* = \{t^* \mid y_{t*} = 1\} \tag{27}$$

Accordingly, the set of all TAF_u valid in anomalous time steps can be isolated:

$$\mathcal{A} = \{ TAF_u : u = u_{t*}, t^* \mid V_u^e \leq t^* \leq V_u^s \} \tag{28}$$

In parallel, the hierarchical clustering algorithm (cf. Section 3.3.2) defines specific ranges of weather elements to differentiate TAFs according to their belonging cluster. This information allows defining a number of $\mathcal{A}$ subsets, which ranges between 1 (in case all TAFs referred to an anomalous t^* belong to a single TAF cluster) and *Nc* (in case at least one TAF referred to an anomalous t^* is assigned to each cluster).

$$\mathcal{A}_{C_i} = \{ TAF_u \mid TAF_u \in \mathcal{A} \wedge TAF_u \; is \; classified \; in \; C_i \} \qquad i = 1, \dots, Nc \tag{29}$$

Figure 6 sketches the outcome of this process. For exemplary purposes, only two anomalous time steps are defined in T^*, i.e., ${t_1}^*$ and ${t_2}^*$, which refer to seven TAFs constituting $\mathcal{A}$. This latter is then re-organized in three sub-sets.

An error propensity metric to assess the tendency of TAFs to generate an anomalous value of POD can be calculated for each cluster as:

$$\eta_{C_i} = \frac{\left|\mathcal{A}_{C_i}\right|}{s_{C_i}} \qquad i = 1, \dots, Nc \tag{30}$$

where $\left|\mathcal{A}_{C_i}\right|$ is the cardinality of C (i.e., number of anomalous TAFs in the *i*-th cluster); and s_{C_i} is the size of C_i based on the whole set of historic TAFs directly obtained from the HC algorithm.

The metric η_{C_i} ranges between 0 and 1 and can be interpreted as the propensity of future TAFs respecting the inclusion criteria in C_i to generate anomalies in POD: the higher η_{C_i}, the higher the chances a TAF belonging to the C_i cluster might be inaccurate. For example, TAFs belonging to a generic cluster C_i with $\eta_{C_i} = 0.20$ will have double the chances to be incorrect with respect to TAFs belonging to another cluster C_j with $\eta_{C_j} = 0.10$.

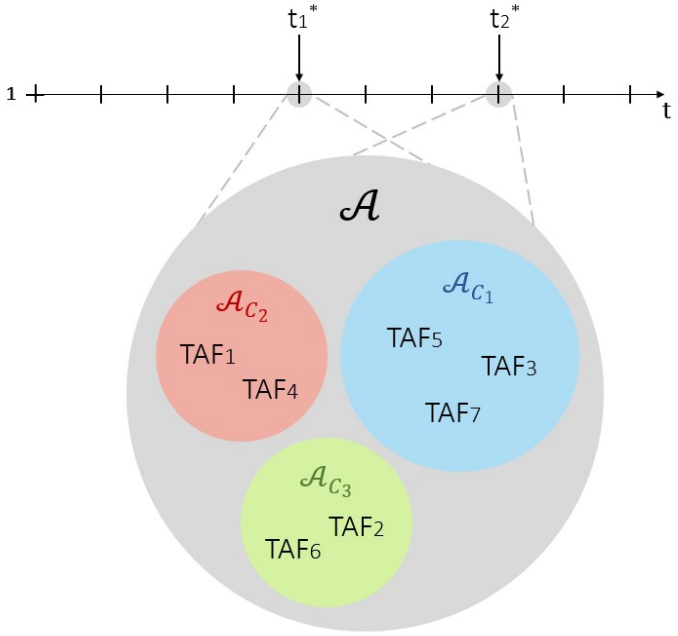


Figure 6. Schematic representation of anomalous TAFs clustering.

Considering the sample size in terms of numerosity of records and respective dimensions, the ML models are usually trained on 70% of the data mart and then tested with the remaining 30%, where samples are randomly chosen to avoid seasonality issues (Boutaba et al., 2018). The testing is performed based on η_{C_i}, with the aim to ensure $\left|{\eta_{C_i}}^1 - {\eta_{C_i}}^0\right| < 5\%$, where ${\eta_{C_i}}^1$ is error propensity for training dataset, and ${\eta_{C_i}}^0$ is the error propensity metric for testing dataset.

4. Results

The proposed approach has been instantiated over a yearly dataset including METARs, SPECIs and TAFs. The input database counts for about 500,000 METARs/SPECIs, and about 50,000 TAFs. The records are referred to 40 aerodromes. The two set of data (i.e., observations and forecasts) have different size since observation are usually made on an hourly base, while forecasts instead have longer time validity. Section 4.1 shows sample results of the descriptive ML solution, complemented with the predictive one presented in Section 4.2.

4.1. Descriptive ML results

Following the theoretical approach described in Section 3.3, the anomaly detection and the hierarchical clustering algorithms is instantiated for a single airport, from now on referred as Airport 1. Data for Airport 1 contains 16,620 METARs/SPECIs and 1,517 TAFs.

Similar results for two additional airports are provided in Annex D, i.e., Airport 2 and Airport 3.

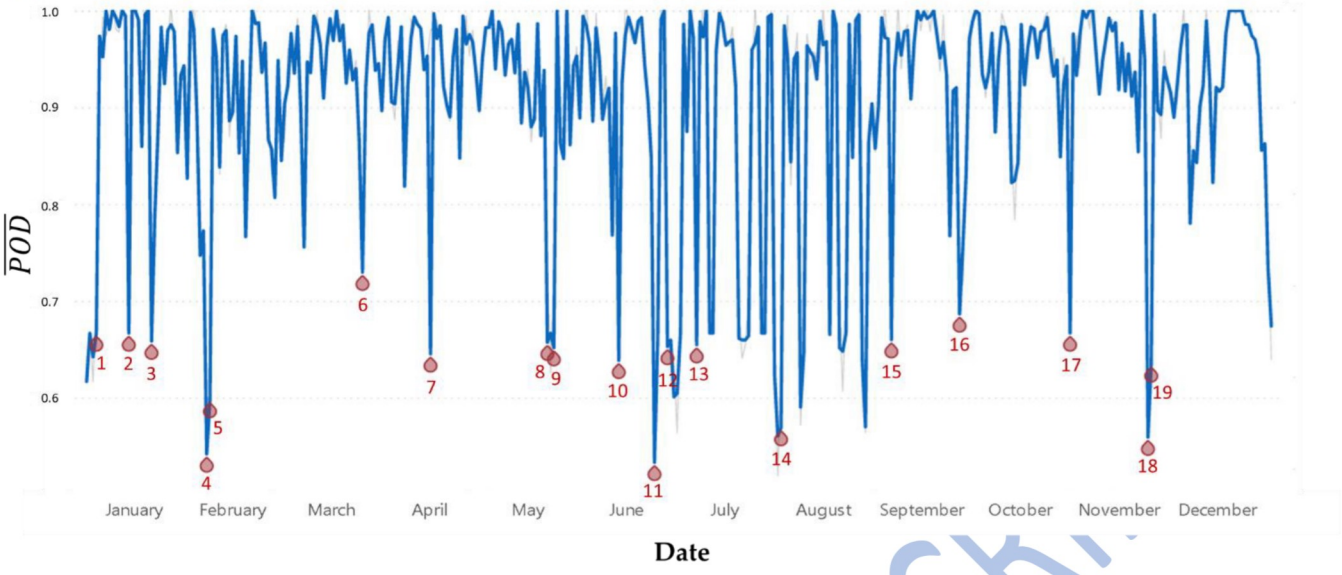
632 *4.1.1.Anomaly detection results*

633 Figure 7 presents the exemplary POD anomalies at Airport 1. Time resolutions are set as $\Delta T = 0.5\ hours$
634 and $\Delta T' = 24\ hours$. A threshold accuracy of 95% between actual and expected $\overline{POD}$ values is established,
635 subsequently minimum and maximum expected values are identified. Data covers a year of forecasts and
636 observations at a fixed airport, during the analysed period 19 anomalous time frames are highlighted. Table 4
637 presents numerical results for the 19 anomalous points out of the 365 under consideration.

638 *Table 4. Anomaly detection algorithm outputs.*

Anomaly point	$\overline{POD}$	Expected $\overline{POD}$	Min $\overline{POD}$	Max $\overline{POD}$	Loss on expected
1	0.67	0.72	0.68	0.76	6.94%
2	0.67	0.76	0.72	0.80	11.84%
3	0.66	0.71	0.67	0.75	7.04%
4	0.54	0.59	0.56	0.62	8.47%
5	0.60	0.64	0.61	0.67	6.25%
6	0.73	0.77	0.73	0.81	5.19%
7	0.65	0.98	0.93	1.00	33.67%
8	0.66	0.74	0.70	0.78	10.81%
9	0.65	0.73	0.69	0.77	10.96%
10	0.64	0.74	0.70	0.78	13.51%
11	0.53	0.57	0.54	0.60	7.02%
12	0.65	0.69	0.66	0.72	5.80%
13	0.65	0.75	0.71	0.79	13.33%
14	0.57	0.64	0.61	0.67	10.94%
15	0.66	0.73	0.69	0.77	9.59%
16	0.69	0.73	0.69	0.77	5.48%
17	0.67	0.74	0.70	0.78	9.46%
18	0.56	0.59	0.56	0.62	5.08%
19	0.61	0.65	0.62	0.68	6.15%

639



640

641 *Figure 7. POD timeseries for Airport 1 (exemplary) with highlighted anomalies time steps. Sensitive data have been removed.*

642 In the provided examples, anomalies have different aetiologies, which have been reconstructed *ex post*, once
643 the ML pipeline emphasized them. Some days referred to anomalous functioning of sensors, (e.g.) days
644 referred to Point 8-9 presented low POD values due to the malfunctioning of an anemometer; others referred
645 to unexpected transient weather conditions, (e.g.) Point 18 was linked to unexpected, localised clouds and
646 precipitation phenomena, caused by smaller-scale weather features difficult to anticipate.

647

648 *4.1.2.Hierarchical clustering results*

649 For exemplary purposes, hierarchical clustering algorithm has been applied on the same number of
650 TAFs described in 4.1.1, where data have been normalized replacing weather elements values with their
651 standardized values (i.e., subtracting the average and dividing by the standard deviation). To set the optimal
652 number of clusters, an average silhouette score $\bar{\varphi}(Nc)$ for $Nc = 2$ to $Nc = 10$ has been calculated, identifying the
653 number of clusters which results in the maximum values of $\bar{\varphi}(Nc)$, i.e., $Nc = 3$ as shown in Figure 8.

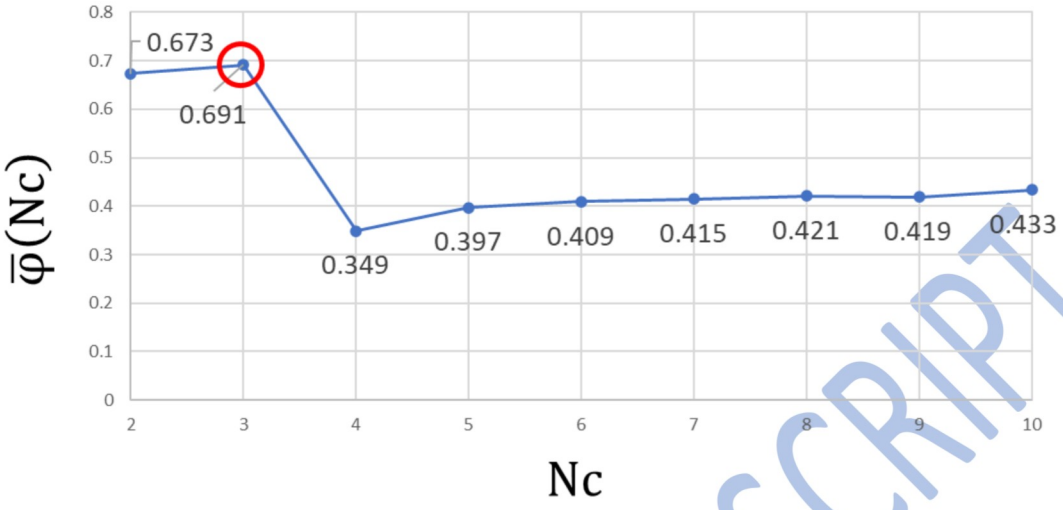


Figure 8. Average silhouette score $\bar{\varphi}$ per number of clusters Nc.

Figure 9 shows the exact silhouette values for all TAFs reorganized in the three identified clusters. Only few elements of C_3 register a slightly negative silhouette score, guaranteeing an overall good quality. Overlapping between clusters is shown to be minimum, this will permit to classify new emitted TAFs in a specific cluster without excessive uncertainty. Computed sizes for each cluster are: $s_{C_1} = 870$, i.e., 33% of data; $s_{C_2} = 580$, i.e., 22% of data; $s_{C_3} = 1{,}166$, i.e., 45% of data.

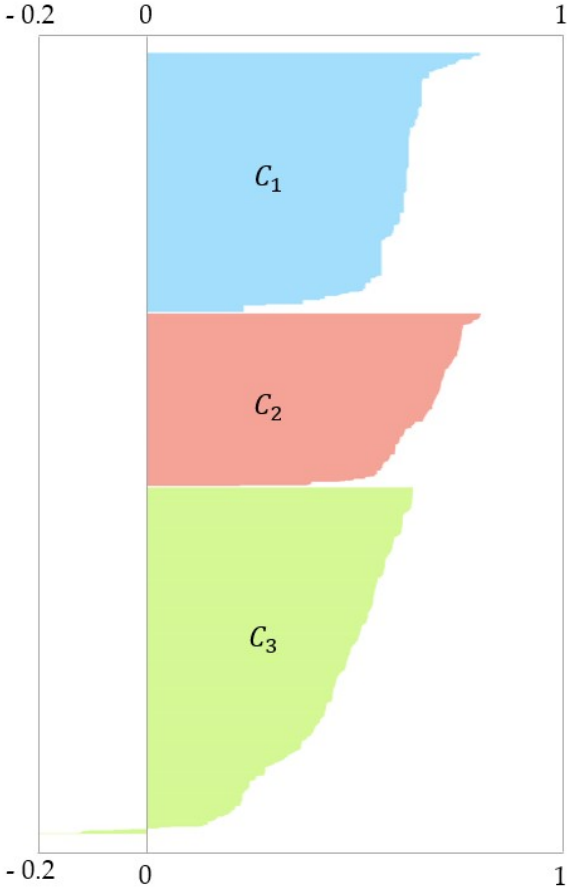


Figure 9 Silhouette scores for observations with Nc = 3.

4.2. *Predictive ML results*

The anomalous time frames highlighted in 4.1.1 are investigated by analysing all the TAFs emitted in their intervals. Each cluster identified in 4.1.2 will count a certain amount of anomaly-generating TAFs, as per the results summarized in Table 5:

Table 5. Number of anomaly-generating TAFs per each cluster.

Anomaly point	C_1	C_2	C_3
1	2	5	0
2	8	0	0
3	0	8	0
4	0	6	1
5	2	5	0
6	0	0	7
7	1	7	0
8	0	8	0
9	4	4	0
10	4	4	0
11	1	7	0

12	4	4	0
13	5	2	0
14	2	3	0
15	3	4	0
16	0	0	7
17	3	5	0
18	1	7	0
19	3	5	0

671

672 At this stage, these results can be combined with clusters' size information to compute the metric η_{C_i} with i =
673 1, 2, 3 as described in 3.4:

$$\eta_{C_1} = \frac{|\mathcal{A}_{C_1}|}{s_{C_1}} = \frac{43}{870} = 0.049 \rightarrow 4.9\% \quad (31)$$

$$\eta_{C_2} = \frac{|\mathcal{A}_{C_2}|}{s_{C_2}} = \frac{84}{580} = 0.145 \rightarrow 14.5\% \quad (32)$$

$$\eta_{C_3} = \frac{|\mathcal{A}_{C_3}|}{s_{C_3}} = \frac{15}{1{,}166} = 0.013 \rightarrow 1.3\% \quad (33)$$

674 These results depict the tendency of each TAF clusters to include TAFs that may generate anomalous POD
675 values, as summarized graphically in Figure 10. Cluster C_2 is the most critical with an error propensity score
676 equal to 14.5%, i.e., there is an expected anomaly in about one TAF every seven among them which are
677 described by the weather elements values characterizing this cluster. More reassuring results are obtained for
678 the remaining two clusters: a forecast failure every fifty forecasts are expected for C_1; almost a forecast failure
679 for a hundred forecasts is expected for C_3.

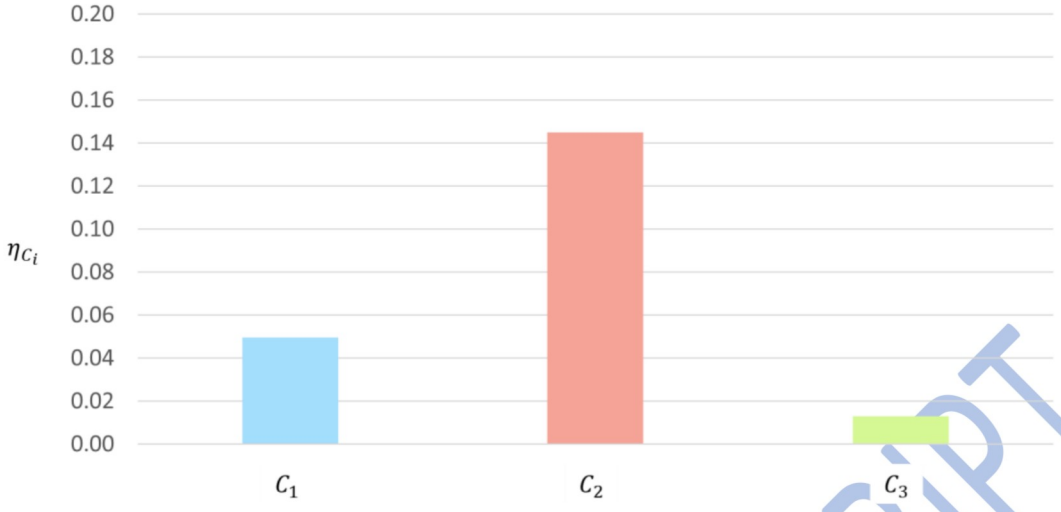


Figure 10. Error propensity of the three clusters linked to Airport 1.

This assessment has practical implications for predictive capacities related to TAFs emission. Every time a new TAF is emitted, its elements can be checked over each cluster typical weather elements to assess where the new TAF belongs. This assessment implies that the new TAF has an error propensity, which is based on the inaccuracy similar TAFs (i.e., belonging to the same cluster) historically had. The interpretation of these results from a decision-making perspective is provided in Section 5.

5. Discussion

Forecasters use data collected by METARs to make prevision on future weather, but no expertise can avoid errors at all. The proposed approach aims to minimize errors on aerodrome forecasting processes through a ML analysis based on historic weather data and weather KPIs.

Section 4 instantiated the proposed approach with historic data from an exemplary airport. Obtained results demonstrate the feasibility of the methodological solution to assess a TAF error propensity in terms of POD index. To confirm reproducibility, the ML pipeline has been tested for two additional cases (c.f. Annex D), showing its validity even for different settings. Finer results are expected to be obtained by implementing a system with incremental refresh of data, considering that η_{C_i} indexes will then depict updated forecasting ability of the aerodrome system. This updating frequency should not be too short to avoid being biased by temporary or seasonal phenomena, i.e., a yearly update can be recommended.

The methodology outputs can be of interest for diverse management perspectives. Decision-makers can benefit of prepared data (cf. Section 3.1) to monitor system's performance at different granularity levels and in terms of different KPIs (cf. Section 3.2). Setting up the data mart opens almost limitless possibilities for BA reporting using knowledge extracted from strings of METARs, SPECIs and TAFs. The obtained scores can support punctual improvement interventions or sharing of best practices among diverse forecasters. At more

operational level, data from the data mart can also support the forecasts generation phase enabling visualizing information in a clearer and more user-friendly way. Forecasters can retrieve historic data and compare it with current status, (e.g.) historical temperature ranges in a specific period and associated TAFs accuracy. Similarly, the anomaly detection output (cf. Section 3.3.1) has been demonstrated to be capable of highlighting time frames with lower performance, motivating deeper investigations. For example, anomalous KPIs could be linked to an erratic failure of a radar system, or to more systematic management errors of certain elements.

The method used to calculate POD among records with different validity, and the aggregation of different *POD* values, represents an insight for future development. Finest method may be experimented, and the different results can be compared to find the most effective solution. Also, in this paper, both descriptive and predictive ML have been instantiated into one airport at a time. However, the same steps remain significant for a set of selected aerodromes $\mathbb{AD}$. Accordingly, Equation (11) can be generalized at regional level:

$$\overline{\overline{POD}}_{t'} = \frac{\sum_{AD \in \mathbb{AD}} \overline{POD}_{t'}}{|\mathbb{AD}|} \tag{34}$$

where $|\mathbb{AD}|$ represent the numerosity in terms of aerodromes which belong to the region under analysis. Accordingly, POD indexes from multiple locations covered by the ANSP can be aggregated to show the mean $\overline{\overline{POD}}_t$, to gain overall understanding of company performance. Locations which majorly contributes at lower values of $\overline{\overline{POD}}$ can be then identified to successfully proceed with the analysis in Section 4.

Furthermore, the error propensity metric η_{C_i} (cf. Section 3.4) becomes a support to refine procedures in certain locations. More specifically, any new TAF generated by a numerical or ML-driven forecast generator, and then assigned to a TAF cluster, shall be subjected to a formal verification of its associated error propensity value. If the propensity is larger than a certain threshold, then it should be recommended to reduce the time interval of such TAF and emit new ones with higher frequency. From a decision-making perspective, these localized actions support resource allocation via the increment of forecasting resolution (and efforts) only when necessary. The error propensity can become a decision support tool for forecasters themselves, who may be more cautious when dealing with generated TAFs with lower values.

The methodology has been tested with an input database containing records of one year of observations and forecasts, but it can be enlarged to longer time intervals and larger sets of airports. The steps of the methodology can be customized for different elements, or KPIs, even custom, or for different bulletin types. A wider experimentation should be made to confirm the positive performance of the proposed solution over other alternative methods to document their pros and cons, and to spot possible areas of improvement (e.g., experimenting various ways of isolating anomalies, or different linkage criteria in hierarchical clustering encompassing time-dependent analyses).

The proposed methodology can be further specialized also considering the management of changes in weather elements, as introduced by group of type PROB (ICAO, 2018). These probability indicators (PROB) outline the

probability of occurrence of alternative values for defined weather elements. Even if out of the scope of this paper, these forecasts can be integrated through specific rules, as previously defined by Sharpe et al. (2016). Similarly, other types of bulletins can be implemented through dedicated pre-processing and ML logic, as needed for (e.g.) SIGMET and AIRMET.

6. Conclusion

In this paper a ML-driven methodology has been presented to support decision making in aerodrome systems with respect to weather forecasting. The proposed approach to deal with such under-investigated research area, required the development of a data pre-processing logic and the design of a specific descriptive and predictive ML pipeline based on anomaly detection and data clustering. This pipeline allowed the definition of an error propensity metric for TAFs to be used both at tactical and operational level based on two ML algorithms: anomaly detection trough spectral residual, and hierarchical clustering.

The definition of this metric, as the outcome of a systematic ML approach, represents the main contribution of this work to the literature. The novelty of the approach is indeed the capability of encompassing systematically historic data to augment the ability of a weather forecast expert in identifying anomalous behaviour and anticipating error-prone forecasts. While it has been acknowledged the absence one-size-fits-all ML algorithms, from a computational perspective, the selection of the anomaly detection algorithm (X. Xu et al., 2019) and the clustering algorithm (R. Xu & Wunsch II, 2005) could be further refined assessing the performance of other approaches. Specifically in terms of clustering, an additional time-based clustering to encompass a time dependent dimension in the generation of TAFs clusters (Paparrizos & Gravano, 2015).

Future studies may investigate the application of the proposed decision support system in other domains. For example, in industrial operations, a warehouse management system may benefit of similar solutions by analysing warehouse picking operations to spot anomalous behaviours, then building clusters for picking orders, and finally compute the error propensity metric to highlight critical orders and re-organizing the facility to better respond to customer needs (e.g., by designing a new layout to better allocate critical products).

Given the early development of this type of studies, a further perspective of improvement should evaluate the cost effectiveness of such solution (Schultz et al., 2018). An assessment of the savings should at least consider: (i) meteorological services operators which would spend fewer working hours in evaluating forecasts accuracy, (ii) resources saved for unnecessary aircraft trajectory deviation or turnarounds, and (iii) intangible assets in the short run, such as higher safety levels.

Overall, the promising results obtained in the study foster the design and development of a real-time automated tool to make the application of this methodology feasible.

Declaration of Interest: None

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

769

Bibliography

Ah-Pine, J. (2010). Normalized kernels as similarity indices. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6119 LNAI*(PART 2), 362–373. https://doi.org/10.1007/978-3-642-13672-6_36

Almeida, V. A. de, França, G. B., & Campos Velho, H. F. de. (2020). Short-range forecasting system for meteorological convective events in Rio de Janeiro using remote sensing of atmospheric discharges. *International Journal of Remote Sensing*, *41*(11), 4372–4388. https://doi.org/10.1080/01431161.2020.1717669

Astsatryan, H., Grigoryan, H., Poghosyan, A., Abrahamyan, R., Asmaryan, S., Muradyan, V., Tepanosyan, G., Guigoz, Y., & Giuliani, G. (2021). Air temperature forecasting using artificial neural network for Ararat valley. *Earth Science Informatics*, *14*(2), 711–722. https://doi.org/10.1007/s12145-021-00583-9

Atay, M., Eroğlu, Y., & Secxkiner, S. U. (2021). Investigation of breaking points in the airline industry with airline optimization studies through text mining before the covid-19 pandemic. *Transportation Research Record*, *2675*(5), 301–313. https://doi.org/10.1177/0361198120987238

Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., & Caicedo, O. M. (2018). A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *Journal of Internet Services and Applications*, *9*(1). https://doi.org/10.1186/s13174-018-0087-2

Cai, J., Zhang, Y., Doviak, R. J., Shrestha, Y., & Chan, P. W. (2019). Diagnosis and Classification of Typhoon-Associated Low-Altitude Turbulence Using HKO-TDWR Radar Observations and Machine Learning. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(6), 3633–3648. https://doi.org/10.1109/TGRS.2018.2886070

Chan, S. T., & Li, L. O. (2003). *Technical Note No . 105 Verification of weather forecasts for the aerodrome of the Hong Kong International Airport* (Issue 105).

Cordeiro, F. M., França, G. B., Neto, F. L. de A., & Gultepe, I. (2021). Visibility and ceiling nowcasting using artificial intelligence techniques for aviation applications. *Atmosphere*, *12*(12), 1–15. https://doi.org/10.3390/atmos12121657

Cristani, M., Domenichini, F., Olivieri, F., Tomazzoli, C., & Zorzi, M. (2018). It could rain: Weather forecasting as a reasoning process. *Procedia Computer Science*, *126*, 850–859. https://doi.org/10.1016/j.procS.2018.08.019

Dejmal, K., & Novotný, J. (2018). Usability and credibility of Czech TAF reports. In *New Trends in Civil Aviation* (pp. 43–47). https://doi.org/10.1201/9781351238649-8

Dejmal, K., Novotny, J., & Hudec, F. (2015). Assessment optimization of weather forecast: Terminal Aerodrome Forecast (TAF) - For 24 hours. *ICMT 2015 - International Conference on Military Technologies 2015*, 58–61. https://doi.org/10.1109/MILTECHS.2015.7153756

Gujanatti, R. B., Vijapur, N., Jadhav, S. S., Manage, P., & Konnur, A. (2021). Machine learning approaches used for weather attributes forecasting. *2021 2nd International Conference for Emerging Technology, INCET 2021*, 4–8. https://doi.org/10.1109/INCET51464.2021.9456291

Hennayake, K. M. S. A., Dinalankara, R., & Mudunkotuwa, D. Y. (2021). Machine Learning Based Weather Prediction Model for Short Term Weather Prediction in Sri Lanka. *2021 10th International Conference on Information and Automation for Sustainability, ICIAfS 2021*, 299–304. https://doi.org/10.1109/ICIAfS52090.2021.9606077

Horne, E., Tibble, H., Sheikh, A., & Tsanas, A. (2020). Challenges of clustering multimodal clinical data: Review of applications in asthma subtyping. *JMIR Medical Informatics*, *8*(5). https://doi.org/10.2196/16452

Hou, X., & Zhang, L. (2007). Saliency Detection: A Spectral Residual Approach. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference On*, *800*, 1–8.

ICAO. (2018). Annex 3, Meteorological Service for International Air Navigation. *International Civil Aviation Organization - International Standards and Recommended Practices*, *July*, 218.

Jaseena, K. U., & Kovoor, B. C. (2020). Deterministic weather forecasting models based on intelligent predictors: A survey. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2020.09.009

Klein, A., Macphail, T., Kavoussi, S., Hickman, D., Phaneuf, M., Lee, R. S., & Simenauer, D. (2009). Nas Weather Index : Quantifying Impact of Actual and Forecast En-Route and Surface Weather on Air Traffic. *14th Conference on Aviation, Range and Aerospace Meteorology*, *January*, 1–13.

Lin, Z., Laska, E., & Siegel, C. (2022). A general iterative clustering algorithm. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. https://doi.org/10.1002/sam.11573

Mahringer, G. (2008). Terminal aerodrome forecast verification in Austro Control using time windows and ranges of forecast conditions. *Meteorological Applications*, *15*(1), 113–123. https://doi.org/10.1002/met.62

Mangortey, E., Puranik, T. G., Pinon, O. J., & Mavris, D. N. (2020). Prediction and analysis of ground stops with machine learning. *AIAA Scitech 2020 Forum*, *1 PartF*(January), 1–20. https://doi.org/10.2514/6.2020-1684

Mecikalski, J. R., Sandmal, T. N., Murillo, E. M., Homeyer, C. R., Bedka, K. M., Apke, J. M., & Jewett, C. P. (2021). A random-forest model to assess predictor importance and nowcast severe storms using high-resolution radar goes satellite lightning observations. *Monthly Weather Review*, *149*(6), 1725–1746. https://doi.org/10.1175/MWR-D-19-0274.1

Montpetit, J., Bourgouin, P., Wilson, L., & Verret, R. (2002). *TAFTOOLS: Development of Objective TAF guidance for Canada and results*.

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, *31*(3), 274–295. https://doi.org/10.1007/s00357-014-9161-z

Murugan Bhagavathi, S., Thavasimuthu, A., Murugesan, A., George Rajendran, C. P. L., Vijay, A., Raja, L., & Thavasimuthu, R. (2021). Weather forecasting and prediction using hybrid C5.0 machine learning algorithm. *International Journal of Communication Systems*, *34*(10), 1–14. https://doi.org/10.1002/dac.4805

Nakhal A, A. J., Patriarca, R., Di Gravio, G., Antonioni, G., & Paltrinieri, N. (2021a). Business intelligence for the analysis of industrial accidents based on MHIDAS database. *Chemical Engineering Transactions*, *86*(January), 229–234. https://doi.org/10.3303/CET2186039

Nakhal A, A. J., Patriarca, R., Di Gravio, G., Antonioni, G., & Paltrinieri, N. (2021b). Investigating occupational and operational industrial safety data through Business Intelligence and Machine Learning. *Journal of Loss Prevention in the Process Industries*, *73*(June), 104608. https://doi.org/10.1016/j.jlp.2021.104608

Novotny, 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), 1–22. https://doi.org/10.3390/atmos12020138

Paparrizos, J., & Gravano, L. (2015). K-shape: Efficient and accurate clustering of time series. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, *2015-May*, 1855–1870. https://doi.org/10.1145/2723372.2737793

Ren, H., Xu, B., Wang, Y., Yi, C., Huang, C., Kou, X., Xing, T., Yang, M., Tong, J., & Zhang, Q. (2019). Time-Series Anomaly Detection Service at Microsoft. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 3009–3017.

Roebber, P. J. (2009). Visualizing multiple measures of forecast quality. *Weather and Forecasting*, *24*(2), 601–608. https://doi.org/10.1175/2008WAF2222159.1

Schultz, M., Lorenz, S., Schmitz, R., & Delgado, L. (2018). Weather Impact on Airport Performance. *Aerospace*, *5*(4), 1–19. https://doi.org/10.3390/aerospace5040109

Schultz, M., Reitmann, S., & Alam, S. (2021). Predictive classification and understanding of weather impact on airport performance through machine learning. *Transportation Research Part C: Emerging Technologies*, *131*(August 2020), 103119. https://doi.org/10.1016/j.trc.2021.103119

Sharpe, M. A., Bysouth, C. E., & Trueman, M. (2016). Towards an improved analysis of Terminal Aerodrome Forecasts. *Meteorological Applications*, *23*(4), 698–704. https://doi.org/10.1002/met.1593

Sladek, D. (2021). Weather phenomena and cloudiness accuracy assessment in TAF forecasts. *2021 8th International Conference on Military Technologies, ICMT 2021 - Proceedings*, 1–6. https://doi.org/10.1109/icmt52455.2021.9502819

Sládek, D. (2019). Attitudes comparison of TAF forecast quality assessment. *ICMT 2019 - 7th International Conference on Military Technologies, Proceedings*. https://doi.org/10.1109/MILTECHS.2019.8870081

Sun, X., Kashima, H., Matsuzaki, T., & Ueda, N. (2010). Averaged stochastic gradient descent with feedback: An accurate, robust, and fast training method. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 1067–1072. https://doi.org/10.1109/ICDM.2010.26

Von Gruenigen, S., Willemse, S., & Frei, T. (2014). Economic value of meteorological services to switzerland's airlines: The case of taf at zurich airport. *Weather, Climate, and Society*, *6*(2), 264–272. https://doi.org/10.1175/WCAS-D-12-00042.1

Wang, A., Xu, L., Li, Y., Xing, J., Chen, X., Liu, K., Liang, Y., & Zhou, Z. (2021). Random-forest based adjusting method for wind forecast of WRF model. *Computers and Geosciences*, *155*. https://doi.org/10.1016/j.cageo.2021.104842

Wang, Y. (2017). Weather impact on airport arrival meter fix throughput. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, *2017-Septe*. https://doi.org/10.1109/DASC.2017.8102133

Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. (2021). Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models. *Journal of Advances in Modeling Earth Systems*, *13*(7). https://doi.org/10.1029/2021MS002502

World Meteorological Organization. (2017). *Manual on Codes, International Codes, VOL. I.1: Vol. I* (Issue WMO-No. 306).

Xu, R., & Wunsch II, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645 – 678. https://doi.org/10.1109/TNN.2005.845141

Xu, X., Liu, H., & Yao, M. (2019). Recent Progress of Anomaly Detection. *Complexity*, *2019*. https://doi.org/10.1155/2019/2686378

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

892 Zhang, X., & Mahadevan, S. (2019). Ensemble machine learning models for aviation incident risk prediction.
893 *Decision Support Systems*, *116*(September 2018), 48–63. https://doi.org/10.1016/j.dss.2018.10.009

894 Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning.
895 *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*,
896 1265–1274. https://doi.org/10.1109/CVPR.2015.7298731

897

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

Annex A

Table 1A. Criteria for TAF elements accuracy evaluation with respect to corresponding METAR elements. The proposed elements refer to ICAO Annex 3 (ICAO, 2018).

Weather Element	Symbol	Rule
Wind direction	*ddd*	if $\lvert ddd_{TAF} - ddd_{METAR}\rvert \geq 20° \rightarrow Miss$ else $\lvert ddd_{TAF} - ddd_{METAR}\rvert < 20° \rightarrow Hit$
Wind speed (intensity)	*ff*	if $\lvert ff_{TAF} - ff_{METAR}\rvert \geq 5\ kt \rightarrow Miss$ else if $\lvert ff_{TAF} - ff_{METAR}\rvert < 5\ kt \rightarrow Hit$
Visibility	*VVVV*	if $VVVV_{TAF} \leq 800\ m \rightarrow \begin{cases} \text{if } \lvert VVVV_{TAF} - VVVV_{METAR}\rvert \leq 200m \rightarrow Hit \\ \text{else} \rightarrow Miss \end{cases}$ else if $VVVV_{TAF} > 800\ m \rightarrow \begin{cases} \text{if } 0.7 \cdot VVVV_{TAF} \leq VVVV_{METAR} \leq 1.3 \cdot VVVV_{TAF} \rightarrow Hit \\ \text{else} \rightarrow Miss \end{cases}$
Weather phenomena	*ww*	ww_{METAR}: Yes / No ww_{TAF} = Yes: *Hit* / *False alarm* ww_{TAF} = No: *Miss* / *Correct rejection*

<table>
<tr>
<td>Clouds type</td>
<td>NsNsNs</td>
<td>if $NsNsNs_{TAF} < 5\ okta$ and $nsnsns_{TAF} \geq 1'500$ m $\rightarrow$ Hit
else if $NsNsNs_{TAF} \geq 5\ okta \rightarrow$
<table>
<tr><td></td><td></td><td colspan="2">$NsNsNs_{METAR}$</td></tr>
<tr><td></td><td></td><td>Yes</td><td>No</td></tr>
<tr><td rowspan="2">$NsNsNS_{TAF}$</td><td>Yes</td><td>Hit</td><td>False alarm</td></tr>
<tr><td>No</td><td>Miss</td><td>Correct rejection</td></tr>
</table>
</td>
</tr>
<tr>
<td>Ceiling</td>
<td>nsnsns</td>
<td>if $nsnsns_{TAF} \leq 300\ m \rightarrow \begin{cases} \text{if } |nsnsns_{TAF} - nsnsns_{METAR}| \leq 30m \rightarrow Hit \\ \text{else} \rightarrow Miss \end{cases}$
else if $nsnsns_{TAF} > 300\ m$
$\rightarrow \begin{cases} \text{if } 0.7 \cdot nsnsns_{TAF} \leq nsnsns_{METAR} \leq 1.3 \cdot nsnsns_{TAF} \rightarrow Hit \\ \text{else} \rightarrow Miss \end{cases}$</td>
</tr>
<tr>
<td>Temperature</td>
<td>TT</td>
<td>if $|TT_{TAF} - TT_{METAR}| \geq 1° \rightarrow Miss$
else $|TT_{TAF} - TT_{METAR}| < 1° \rightarrow Hit$</td>
</tr>
</table>

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

Annex B

This annex reports information about how TAF change groups are managed in this work:

- From indicator (FM)

The FM indicator describes changes in one or more weather elements that apply from a specific time moment until the end of TAF validity. The values introduced after the FM indicator overwrite the ones declared in the main part of a TAF.

For example, if the main TAF forecasts no rain phenomenon, and in the FM group there is a mention to it, then rain is expected starting from the FM start validity time, i.e., the time from which the FM groups is valid. A TAF containing the FM change group is represented in Figure 1Ba. The FM validity time starts at 07:00 of the 1st of January, while the main forecast is valid from 01:00 to 12:00. It is possible to notice that at 07:00 the FM forecast (which predicts rain) completely substitutes the main one.

- Becoming indicator (BECMG)

The BECMG indicator describes intervals where weather elements are expected to reach or pass through specified thresholds. The validity interval of a BECMG is interpreted as a transition period, during which both the main and the BECMG weather elements are considered valid. Afterwards, i.e. outside the BECMG validity, the BECMG elements overwrite the ones declared in the main part of a TAF.

Imagine a string where the main TAF does not forecast rain phenomena, but rain is indicated in the BECMG group: in this case during the transition time, the presence or absence of rain are both allowed; at the end of BECMG validity, rain occurrence is considered to be forecasted since the BECMG group overwrite the main TAF forecast. A TAF containing the BECMG change group is represented in Figure 1Bb. The BECMG validity time starts at 07:00 of the 1st of January and ends at 08:00, while the main forecast is valid from 01:00 to 12:00. Accordingly, from 07:00 to 08:00 both rain and no rain are considered to be correct, while from 08:00 the BECMG prevision completely substitutes the main one expecting rain.

- Temporary indicator (TEMPO)

The TEMPO indicator describes temporary fluctuations of certain weather elements. During the TEMPO validity, both the main and the extra elements are considered valid. Outside the TEMPO interval, the main elements apply. Note however that the expected fluctuations should last less than one half of the time period of the TEMPO group, as per ICAO recommendations (ICAO, 2018). This situation differs from BECMG, where during the extra group validity, both elements are always considered valid.

In terms of KPIs, the number of time steps correctly forecasted in a TEMPO group cannot exceed the half of the total TEMPO interval. A penalty is imposed if the TEMPO condition is not observed at all; the non-occurrence lowers the score up to one-third of the total number of ΔT time steps in the TEMPO interval. In

terms of accuracy, the contribution of a TEMPO group which lasts τ time steps can be resumed as (Chan & Li, 2003):

$$H_{CORR} = \begin{cases} H_M + min(\tau/2, H_{TnotM}) & if\ H_T > 0 \\ max(0, H_M - \tau/3) & if\ H_T = 0 \end{cases} \quad (35)$$

where H_{CORR} is the number of time steps in which the TAF is considered correct; H_M is the number of time steps in which the main forecast is correct; H_T is the number of time steps in which the TEMPO forecast condition is correct; H_{TnotM} is the number of time steps in which the TEMPO forecast condition is correct and the main forecast is not correct.

For example, if considering a string where the main TAF forecast any rain phenomenon, but rain is indicated in a TEMPO group, the accuracy analysis has to be made as follows. During the TEMPO validity, both rain and no rain must occur to obtain the highest score (1). Even though the presence and absence of rain are both allowed, they may generate penalties on final accuracy scoring. At the end of the TEMPO validity, no rain is expected since the TEMPO group indicates a temporary phenomenon, after which the main TAF forecast returns to be effective. A TAF containing the TEMPO change group is represented in Figure 1Bc. The TEMPO validity time starts at 07:00 of the 1st of January and ends at 08:00, while the main forecast is valid from 01:00 to 12:00. Accordingly, from 07:00 to 08:00 both rain and no rain are considered correct for accuracy analysis following the corrections in (35). From 08:00 the main TAF forecast return to be effective.

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

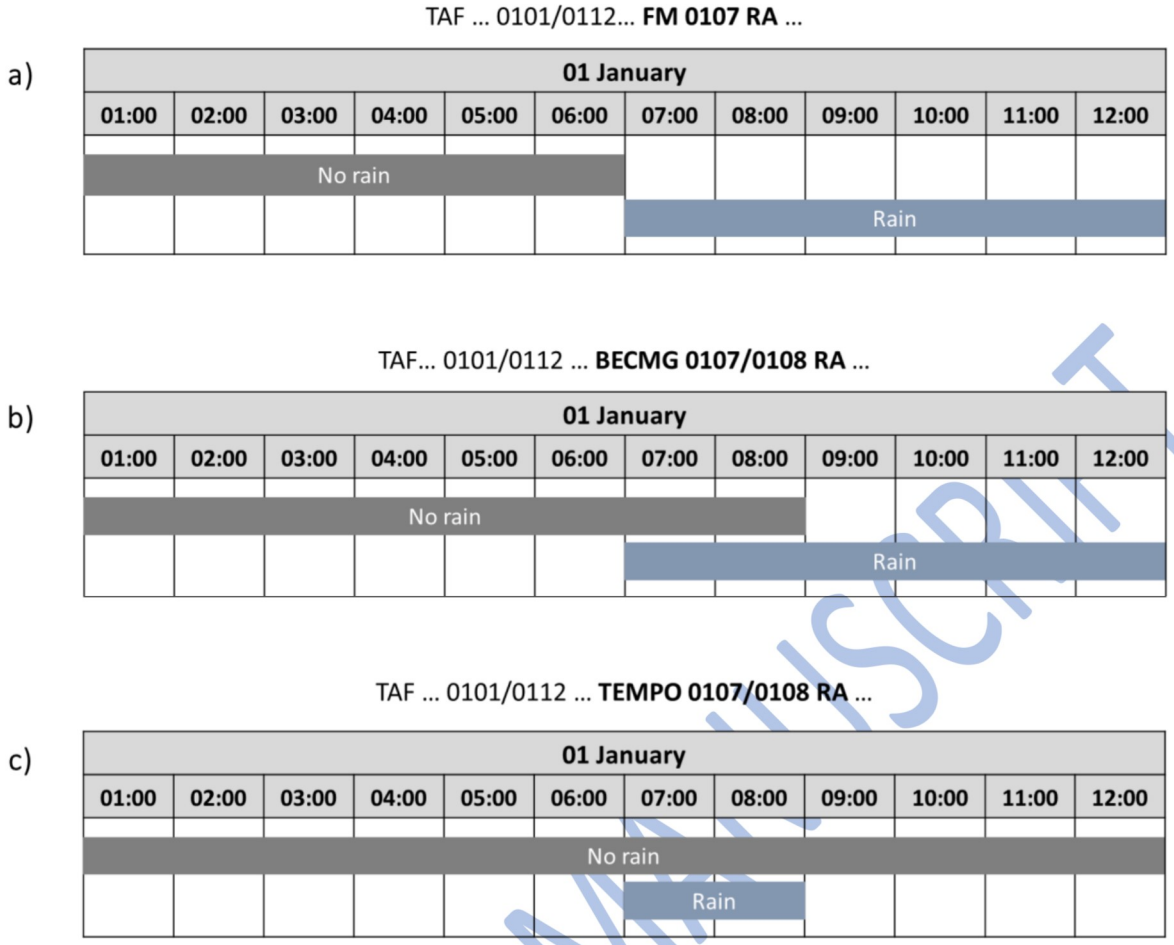


Figure 1B. a) Representation of forecast containing a change in rain acting from 1 January at 07:00 of type FM; b) Representation of forecast containing a change in rain acting from 1 January at 07:00 to 1 January at 08:00 of type BECMG; c) Representation of forecast containing a change in rain acting from 1 January at 07:00 to 1 January at 08:00 of type TEMPO.

955 **Annex C**

956 *Table 1C. TAF features used for clustering.*

Feature	Description	Variable type
TAF validity duration	Difference between V^e and V^s in hours. Mandatory value, no need to fill voids.	Numeric
Wind direction	Value of *ddd*. Mandatory element, no need to fill voids.	Numeric
Wind speed	Value of *ff*. Mandatory element, no need to fill voids.	Numeric
Wind gust	Value of *fmfm*. Voids are filled setting value to zero, i.e. no gust.	Numeric
Visibility	Value of *VVVV*. Mandatory element, no need to fill voids.	Numeric
Weather phenomena	Value of *ww*. Up to three weather phenomena groups can be available. They are treated as three different features. Voids are filled by "N/A" feature in one-hot encoding.	Categorical
Cloud type and ceiling (up to four)	Value of *NsNsNs* concatenated to value of *nsnsns*. Up to four clouds groups can be available. They are treated as four different features. Voids are filled by "N/A" feature in one-hot encoding.	Categorical
Pressure	Value of pressure indicated in TAF (no notation is used in this work). Voids are filled with weighted average value of pressure.	Numeric
Temperature	Value of *TT*. Voids are filled with weighted average value of temperature.	Numeric
Change groups number	Count of change groups reported in the TAF string. Zero value depicts no extra group in TAF, no need to fill voids.	Numeric

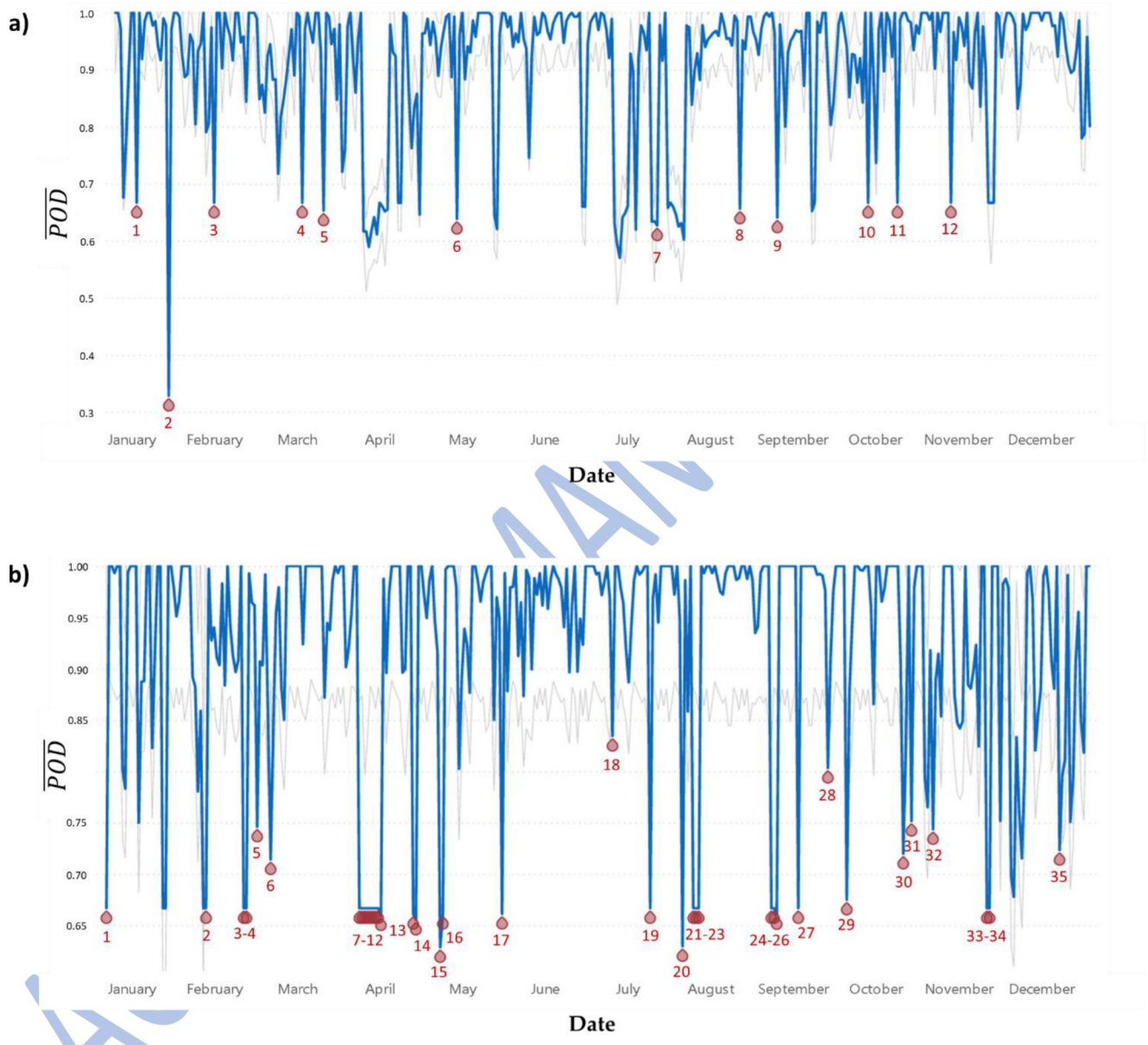
957

958

959 **Annex D**

960 Note that Airport 2 data include 15'470 METARs/SPECIs and 1,495 TAFs while Airport 3 data include 16,450
961 METARs/SPECIs and 1,489 TAFs.

962



963

964 *Figure 1D. a) POD timeseries for Airport 2 with highlighted anomalies time steps. Algorithm accuracy at 90%: 12 anomalous points out of 365 are*
965 *identified. b) POD timeseries for Airport 3 with highlighted anomalies time steps. Algorithm accuracy at 85%: 35 anomalous points out of 365 are*
966 *identified. Sensitive data have been removed.*

967

968

969

970

971

972 *Table 1D. Anomaly detection algorithm outputs for Airport 2.*

Anomaly point (a)	$\overline{POD}$	Expected $\overline{POD}$	Min $\overline{POD}$	Max $\overline{POD}$	Loss on expected
1	0.67	0.75	0.68	0.81	10.67%
2	0.33	0.98	0.91	1.00	66.33%
3	0.67	0.75	0.68	0.81	10.67%
4	0.67	0.74	0.68	0.80	9.46%
5	0.65	0.74	0.68	0.80	12.16%
6	0.64	0.73	0.66	0.79	12.33%
7	0.63	0.72	0.65	0.78	12.50%
8	0.66	0.73	0.67	0.80	9.59%
9	0.64	0.71	0.64	0.77	9.86%
10	0.67	0.77	0.71	0.84	12.99%
11	0.67	0.76	0.69	0.82	11.84%
12	0.67	0.75	0.68	0.81	10.67%

973

974

975 *Table 2D. Anomaly detection algorithm outputs for Airport 3.*

Anomaly point (b)	$\overline{POD}$	Expected $\overline{POD}$	Min $\overline{POD}$	Max $\overline{POD}$	Loss on expected
1	0.67	0.98	0.87	1.00	31.63%
2	0.67	0.94	0.83	1.00	28.72%
3	0.67	0.91	0.80	1.00	26.37%
4	0.67	0.94	0.83	1.00	28.72%
5	0.75	0.99	0.88	1.00	24.24%
6	0.71	1.00	0.89	1.00	29.00%
7	0.67	0.98	0.87	1.00	31.63%
8	0.67	0.98	0.87	1.00	31.63%
9	0.67	0.91	0.87	1.00	26.37%
10	0.67	0.94	0.80	1.00	28.72%
11	0.67	0.99	0.83	1.00	32.32%
12	0.66	0.98	0.88	1.00	32.65%
13	0.66	0.99	0.87	1.00	33.33%
14	0.66	0.97	0.86	1.00	31.96%
15	0.63	0.98	0.87	1.00	35.71%

16	0.66	0.98	0.87	1.00	32.65%
17	0.66	0.98	0.86	1.00	32.65%
18	0.83	0.99	0.88	1.00	16.16%
19	0.67	0.94	0.83	1.00	28.72%
20	0.63	0.98	0.87	1.00	35.71%
21	0.67	0.99	0.88	1.00	32.32%
22	0.67	0.99	0.87	1.00	32.32%
23	0.67	0.97	0.86	1.00	30.93%
24	0.67	0.94	0.83	1.00	28.72%
25	0.67	0.99	0.88	1.00	32.32%
26	0.66	0.99	0.87	1.00	33.33%
27	0.67	0.99	0.88	1.00	32.32%
28	0.80	0.99	0.88	1.00	19.19%
29	0.67	0.98	0.87	1.00	31.63%
30	0.72	0.99	0.88	1.00	27.27%
31	0.75	0.96	0.85	1.00	21.88%
32	0.74	0.94	0.83	1.00	21.28%
33	0.67	0.98	0.87	1.00	31.63%
34	0.67	0.99	0.88	1.00	32.32%
35	0.72	0.99	0.88	1.00	27.27%

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

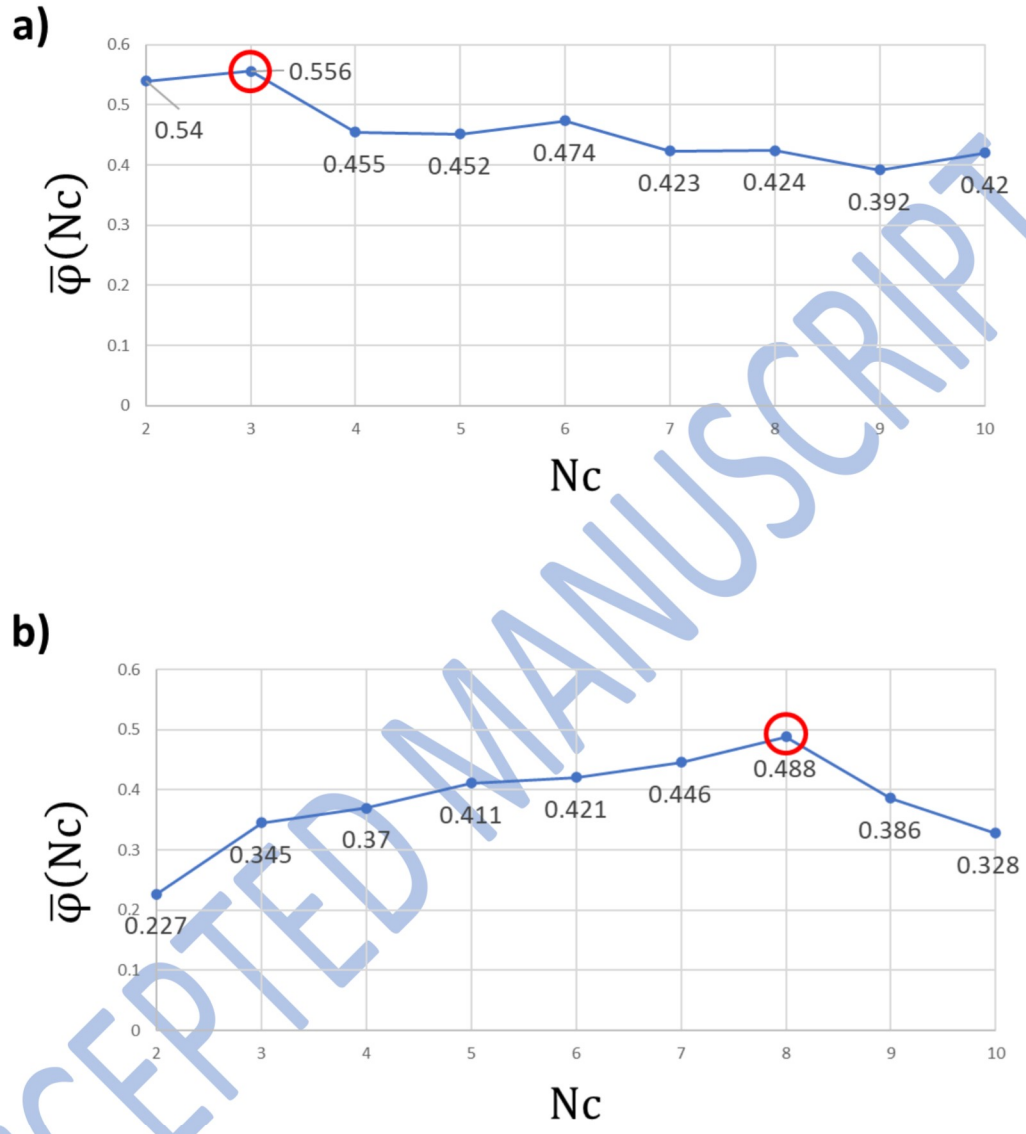


Figure 2D. a) Average silhouette score $\bar{\varphi}$ per number of clusters Nc for Airport 2. b) Average silhouette score $\bar{\varphi}$ per number of clusters Nc for Airport 3.

Table 3D. Number of anomaly-generating TAFs per each cluster for Airport 2.

Anomaly point	C_1	C_2	C_3
1	3	0	4
2	0	1	6
3	0	1	5
4	0	4	3
5	1	2	4
6	0	5	0
7	1	5	1
8	2	5	0
9	0	5	2
10	6	0	1
11	1	6	0
12	1	3	2

Table 4D. Number of anomaly-generating TAFs per each cluster for Airport 3.

Anomaly point	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
1	0	0	0	1	0	2	0	0
2	0	0	0	3	0	3	0	2
3	0	0	0	0	0	8	0	0
4	0	0	2	1	0	5	0	0
5	2	1	0	0	2	2	0	1
6	1	0	0	1	0	3	0	1
7	0	0	0	0	0	7	0	0
8	0	0	0	0	0	7	0	0
9	0	0	0	0	0	8	0	0
10	0	0	0	0	0	8	0	0
11	0	0	0	0	0	8	0	0
12	0	0	0	0	0	8	0	0
13	0	0	0	0	0	7	0	0
14	0	0	0	0	0	3	0	4
15	0	0	0	0	0	1	0	7
16	2	3	1	0	0	0	1	1
17	4	1	0	0	0	0	1	2
18	3	1	2	0	0	0	0	0
19	4	1	2	0	0	0	0	0

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

20	5	0	0	0	0	0	0	3
21	0	0	3	0	0	0	0	4
22	0	0	0	0	0	4	0	4
23	0	0	0	0	0	6	0	0
24	0	0	0	0	0	3	0	5
25	0	0	0	0	0	2	0	5
26	0	0	3	0	0	0	0	4
27	0	0	0	1	0	2	1	4
28	0	0	0	0	0	5	0	3
29	0	0	0	1	0	4	0	2
30	0	0	2	0	0	1	0	5
31	2	0	1	0	0	2	0	3
32	1	0	0	1	0	3	0	2
33	0	0	1	2	1	0	1	3
34	2	2	2	0	2	0	0	0
35	6	0	0	0	0	0	0	0

1011

1012

1013

1014

1015

1016

Riccardo Patriarca, Francesco Simone, Giulio Di Gravio, Supporting weather forecasting performance management at aerodromes through anomaly detection and hierarchical clustering, Expert Systems with Applications, 2022, 119210, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.119210.

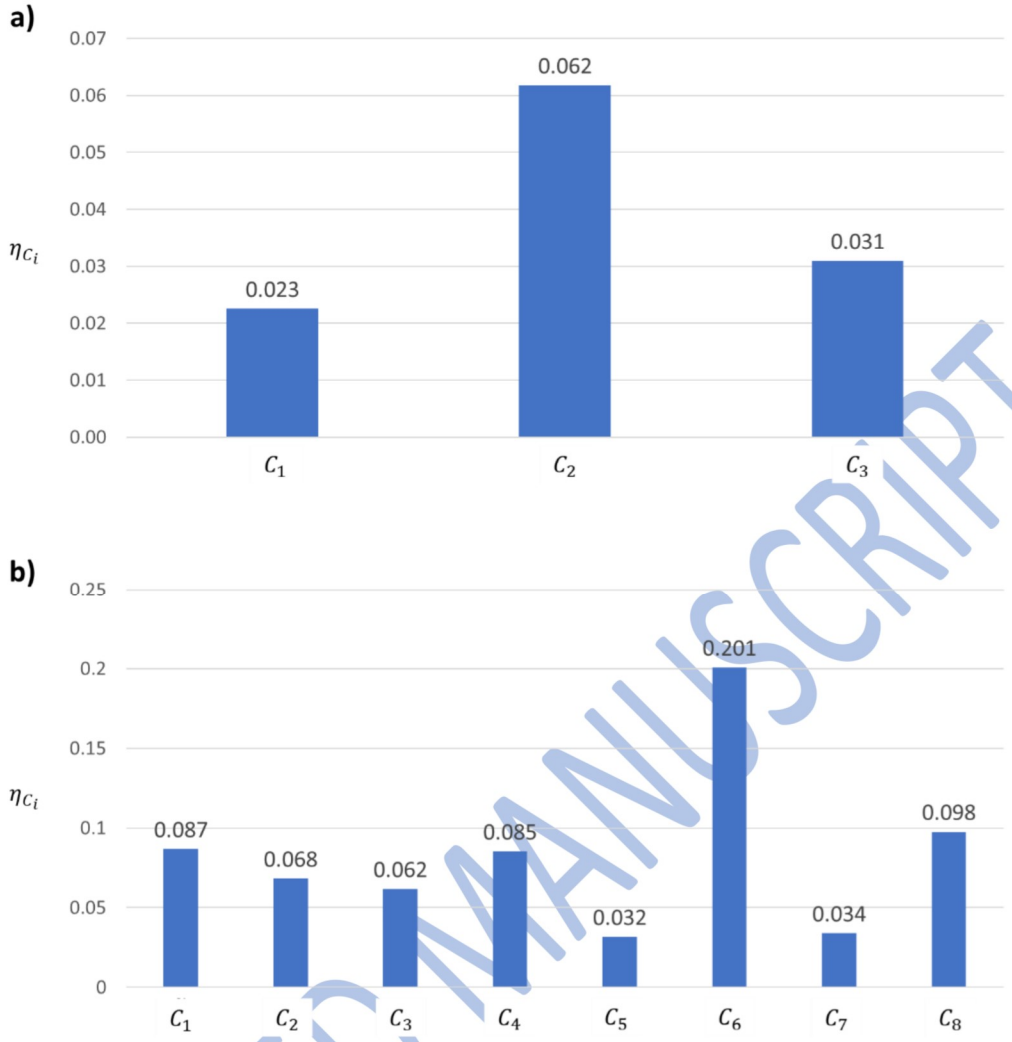


Figure 3D. a) Error propensity of the three clusters linked to Airport 2. b) Error propensity of the eight clusters linked to Airport 3.