Issue 25(1), 2025, pp. 133-159
ISSN: 1567-7141
TU Delft OPEN Publishing

EJTIR

SI: Sesar Innovation Days 2023
https://ejtir.tudelft.nl
DOI: 10.59490/ejtir.2025.25.1.7487

A Collection of Machine Learning Models for Improved Airport Operations Amidst Adverse Weather Conditions

Ramon Dalmau[1], Jonathan Attia[2]

[1] corresponding author ramon.dalmau-codina@eurocontrol.int, EGSD/INO/ENG, EUROCONTROL, France; 0000-0003-3587-7331
[2] EGSD/INO/ENG, EUROCONTROL, France

Keywords
Machine learning
Adverse weather
Airport operations
Air traffic flow management

Publishing history
Submitted: 27 March 2024
Revised date(s): 27 June 2024
Accepted: 14 October 2024
Published: 5 February 2025

Cite as
Dalmau, R., & Attia, J. (2025). A collection of machine learning models for improved airport operations amidst adverse weather conditions. *European Journal of Transport and Infrastructure Research, 25*(1), 133-159.

Abstract

In the face of escalating climate change, airports worldwide are finding themselves at the mercy of extreme weather events. This research paper presents a comprehensive system that models key indicators, aiding airport management during such challenging weather conditions. The system adopts an integrated approach, combining various machine learning models to provide a detailed projection of an airport's future state, drawing from past occurrences. The heart of the system is a model that focuses on the airport's peak service rate. This model meticulously correlates weather conditions and runway configurations with the 99th percentile of observed throughput from the training dataset. As such, the peak service rate model provides an estimate of the airport's capacity, which is essential for effective planning and resource allocation. Moreover, the system includes a predictive model that assesses the likelihood of air traffic flow management regulations based on weather data and calendar information. The robustness of this model against noise and uncertainty in the training dataset is fortified by the application of confident learning techniques and the inclusion of monotonic constraints. The system further enhances its capabilities by forecasting the potential entry rate of regulations, expressed in hourly arrivals, providing valuable insights that can guide proactive decision-making. By seamlessly integrating these three models, the system serves as an effective tool for airport operators and airlines. It enables operational optimisation and the development of strategic plans to mitigate the effects of increasing weather-related disruptions.

1 Introduction

Airport capacity is defined as the number of hourly arrivals and departures that can be accommodated. Both the runway configuration and weather conditions affect capacity. Air traffic flow management (ATFM) regulations are frequently used by flow managers when the expected traffic demand exceeds the predicted capacity of the airport. Flights subject to ATFM regulations are assigned ground delays (also known as ATFM delays) to smooth traffic demand and avoid overloads that could lead to safety issues or increased flight consumption due to airborne holding.

ATFM regulations are occasionally implemented at European airports. From the resurgence of air traffic following the lifting of COVID-19 pandemic restrictions on June 15th, 2021, until May 31st,2023, a total of 1.3K ATFM regulations were implemented at airports within the European Civil Aviation Conference (ECAC) region, generating 430 K minutes of delay. Notably, 13% of these ATFM regulations were caused by adverse weather conditions (e.g., snow). As the effects of climate change intensify (Furtak, 2023), airports are increasingly contending with extreme weather events (Voskaki, 2023; De Vivo, 2022) and the incidence of weather-related ATFM regulations may become more significant.

ATFM regulations caused by adverse weather conditions are typically implemented in advance, using predictions of traffic demand and weather forecasts to anticipate potential negative impacts on airport capacity. Precise and accurate estimation of airport capacity is essential for ensuring the effective and efficient implementation of ATFM regulations. Overestimating airport capacity may force arriving flights to wait in holding stacks (resulting in airborne delays), whereas underestimating airport capacity may result in excessive ATFM (i.e., ground) delays.

This study proposes a decision-support system to improve airport operations under adverse weather conditions. The system uses machine learning techniques to model airport capacity and evaluate the effectiveness of ATFM regulations. The core component of the system is a model that predicts the peak service rate of an airport, representing the 99thpercentile of arrival and departure throughput, conditioned on weather conditions and runway configuration. This model was initially introduced in a previous publication by the authors (Dalmau, 2023a). The system also includes a component that predicts whether ATFM regulations will be implemented in response to adverse weather conditions in addition to high traffic demand. This model, originally proposed by (Dalmau, 2023b), utilizes confident learning techniques to reduce label noise in the dataset, ensuring reliable predictions even amid data uncertainty. Furthermore, the application of monotonic constraints during model training enhanced performance and interpretability. Finally, the system includes a model that predicts the potential entry rate of ATFM regulations and provides information about the expected severity of ATFM measures if implemented. By integrating all these models, the system serves as a comprehensive tool for airport operators and airlines, assisting in the optimization of operations and anticipation of weather-related disruptions.

The remainder of this paper is organized as follows. Section 2 provides an in-depth discussion of previous works that defined the peak service rate and ATFM regulation probability models. Section 3 provides a literature review. Sections 4 and 5 describe the system architecture and machine learning implementation, respectively. Finally, Section 6 presents the performance of the various models, and Section 7 presents the main conclusions and plans for future work.

2 Background

The following sections provide a brief but concise overview of previous works that define the peak service rate and ATFM regulation probability models.

2.1 Peak service rate prediction:

The model presented in (Dalmau, 2023a) was developed based on the belief that machine learning can effectively learn mapping from weather conditions (visibility, wind speed, etc.) and runway configuration to airport capacity from historical data. However, to train the machine-learning model, both predictors (i.e., weather conditions and runway configuration) and targets (i.e., capacity) are required. Meteorological aerodrome reports (METARs) can be effectively used to characterize past weather conditions, and obtaining runway configuration and airport capacity data poses a challenge.

To address the first issue, the authors proposed a method inspired by multilabel classification problems to determine the active runway configuration at an airport from traffic data. In simple terms, the matching process was performed by treating the official runway configurations defined in the Airport Corner[1]. Accessed on March 7th, 2024. as predictions generated by a hypothetical classifier, and the observed runways in use as the ground truth. Each observation was then assigned the most similar prediction (i.e., runway configuration) in relation to the observed runway combination.

To address the second issue, the authors capitalized on the fact that when traffic demand exceeds airport capacity, the observed throughput (i.e., the number of movements per hour) becomes a reliable indicator of the capacity itself. This principle underpins the concept of the *peak service rate*, which represents the 99th percentile of the observed throughput, and acts as a capacity proxy for airports operating at or near capacity. Accordingly, the focus of the study was not on learning from the declared capacities, but on unraveling the highest sustainable throughput the airport can achieve, conditioned on the runway configuration and weather conditions.

To illustrate the concept of the peak service rate, Fig. 1 shows the cumulative proportion of arrival throughput at Zurich airport when the visibility was above and below 800 m (i.e., Cat II/IIIA/IIIB precision approach conditions), regardless of the runway configuration. The dashed vertical line represents the peak service rate.

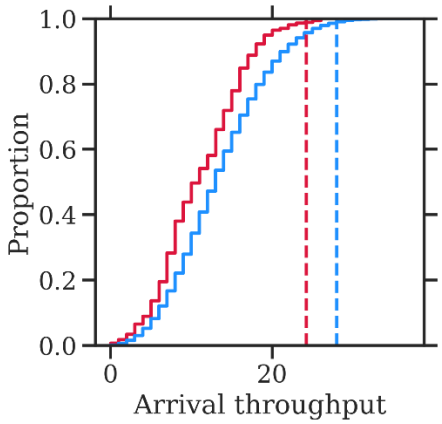


Figure 1. Cumulative proportion of arrival throughput when the visibility was above (blue) and below (red) 800 meters at Zurich airport from June 15th, 2021 to May 31st, 2023. The vertical dashed line indicates the 99th percentile (i.e., the peak service rate).

According to Fig. 1, the arrival throughput is lower than or equal to 28 movements per hour in 99% of the observations when visibility is higher than 800 m. As expected, this peak service rate decreases by approximately four movements per hour (15%) under low-visibility conditions. Accordingly, airport capacity, and consequently its peak service rate, is undeniably influenced by weather conditions, and modelling this influence was precisely the focus of (Dalmau, 2023a). Readers interested in further details about the multiclass classification strategy, which determines

[1] https://www.eurocontrol.int/tool/airport-corner

the actual official runway configuration from the observed runways in use, as well as the concept of peak service rate, are encouraged to refer to the original publication.

2.2 Probability of regulation prediction:

As mentioned in the introduction, ATFM regulations due to adverse weather are implemented with a certain look-ahead time, relying on a weather forecast to determine the expected capacity drop. In practice, however, inaccuracies in weather forecasts (Patriarca, 2023) may lead to an actual period of capacity reduction differing from that predicted several hours ahead. To illustrate the impact of this problem when training machine-learning models, we consider a dataset comprising numerous observations. In this dataset, each observation corresponds to a specific time period, such as 1 h, and includes a wide range of variables representing weather conditions, traffic demand (or a proxy for traffic demand such as calendar information), and the binary label indicating the presence (positive) or absence (negative) of ATFM regulation due to weather. At first glance, this dataset can be used to train a machine learning model to predict the likelihood of ATFM regulation due to weather, conditioned on weather conditions, and traffic demand.

Because ATFM regulations are activated by humans based on the perceived severity of a weather forecast, some positive observations may be linked to regulations that were put into place owing to pessimistic weather forecasts made several hours in advance. However, the actual weather conditions at the time of observation might not have been as severe, rendering the regulation too restrictive or less effective than initially planned. On the other hand, some negative observations could be associated with periods in which a regulation would have been beneficial in preventing airborne holdings but was not implemented due to an overly optimistic forecast. By the time precise weather conditions became certain, it was too late to put the regulation into effect. In essence, dataset labels may contain noise, with some positive observations that should actually be negative, and vice versa. If this noise in the labels is not addressed, it could negatively impact the training of the model, leading to incorrect or non-intuitive relationships among weather, traffic demand, and the likelihood of ATFM regulations.

A similar problem was encountered in (Dalmau, 2023c) where the authors attempted to determine the probability of flight diversion due to weather. In this case, some (but unknown) of the observed flight diversions were attributed to other unpredictable reasons (e.g., medical emergencies, unruly passengers, or night curfew infringements) and should be considered as belonging to the negative class when the objective is to learn the mapping between adverse weather conditions and the probability of diversion due to weather. To address the noise in the labels, confident learning (CL) (Northcutt, 2021) a method to automatically filter out likely mislabelled observations from the dataset, was adopted.

In the prequel of this study (Dalmau, 2023b), CL was also applied to filter out positive observations from the dataset that were regulated but likely no longer effective, as well as negative observations where ATFM regulation was not implemented but could have been beneficial. Subsequently, a machine learning model was trained on the clean dataset to learn, with confidence, the relationship between adverse weather conditions, traffic demand, and the probability of ATFM regulation owing to weather.

Furthermore, the relationship between certain predictors and targets was known in advance. For instance, it is well known that, all else being equal, the lower the visibility, the higher is the probability of regulation. Ideally, the model should autonomously learn these relationships. In practice, the noise in a dataset may lead the model to learn incorrect relationships. For instance, if regulations are (by coincidence) never active in a given airport when thunderstorms are present, the model may interpret that thunderstorms decrease the probability of regulation. A human can easily identify this type of misinterpretation, but a model requires guidance.

Monotonic constraints ensure that certain predictors exhibit monotonic relationships with the target. Two types of monotonic constraint are possible.

$$f(x_1, x_2, \ldots, x, \ldots, x_d) \leq f(x_1, x_2, \ldots, x', \ldots, x_d) \quad (1)$$

whenever $x \leq x'$ is a positive constraint; or

$$f(x_1, x_2, \ldots, x, \ldots, x_d) \geq f(x_1, x_2, \ldots, x', \ldots, x_d) \quad (2)$$

where $x \geq x'$ is a negative constraint. In Eqs. (1) and (2), $x_1, x_2, \ldots, x_d$ are the d input features (i.e., predictors) of Model f.

The results reported by (Dalmau, 2023b) revealed that the combination of confident learning and monotonic constraints results in a robust and reliable machine learning model capable of accurately estimating the probability of ATFM regulation due to weather.

Figure 2 shows, for illustrative purposes, the predictions of this model at Zurich Airport on February 14th, 2023. On this specific day, the morning and evening were affected by severe obscuration caused by freezing fog. This visual example is supported by meteorological reports presented in Table 1.

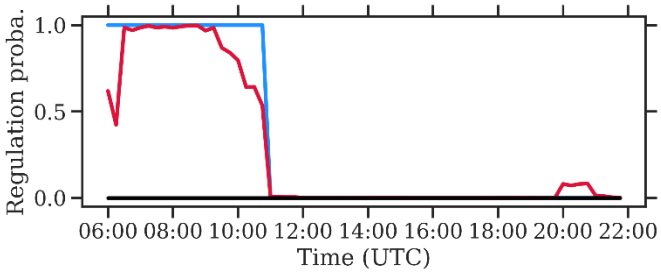


Figure 2. Predicted probability of ATFM regulation due to weather (red) and actual ATFM regulation activation status (blue) at Zurich airport on 14th February, 2023.

Table 1. Meteorological reports at Zurich airport on 14th February, 2023.

METAR LSZH 142150Z VRB02KT **0400** FZFG VV002 M01/M01 Q1032
METAR LSZH 142120Z VRB02KT **0250** FZFG VV001 M01/M01 Q1032
METAR LSZH 142050Z VRB02KT **0200** FZFG VV001 M01/M01 Q1032
METAR LSZH 142020Z 31003KT **0300** BCFG VV002 M02/M02 Q1033
METAR LSZH 141950Z 31003KT 4500 BR FEW002 M02/M02 Q1033
METAR LSZH 141920Z 29002KT 5000 BR NSC M02/M03 Q1033
METAR LSZH 141850Z VRB01KT 6000 NSC M00/M01 Q1033
...
METAR LSZH 141220Z VRB03KT 6000 SCT005 SCT008 02/00 Q1034
METAR LSZH 141150Z VRB02KT 5000 BR FEW004 BKN006 02/M00 Q1035
METAR LSZH 141120Z VRB02KT 4000 BR FEW003 BKN005 01/M01 Q1035
METAR LSZH 141050Z VRB03KT 2500 BR BKN003 OVC005 01/M00 Q1036
METAR LSZH 141020Z VRB02KT 2500 BR BKN003 OVC004 00/M01 Q1036
METAR LSZH 140950Z VRB01KT 1800 PRFG OVC003 M00/M01 Q1036
METAR LSZH 140920Z VRB03KT 1200 PRFG VV003 M00/M01 Q1036
METAR LSZH 140850Z VRB01KT 0900 FZFG VV002 M00/M01 Q1036
METAR LSZH 140820Z VRB01KT **0600** FZFG VV002 M01/M01 Q1037
METAR LSZH 140750Z VRB02KT **0600** FZFG VV002 M01/M01 Q1037
METAR LSZH 140720Z VRB03KT **0600** FZFG VV002 M01/M01 Q1037
METAR LSZH 140650Z VRB02KT **0600** FZFG VV002 M01/M01 Q1037
METAR LSZH 140620Z VRB02KT **0700** FZFG VV003 M01/M01 Q1036

According to Fig. 2, the model effectively captured the ATFM regulation due to weather active from 6AM to 11AM. It is interesting to observe how the predicted probability decreases from 9AM to 11AM as the visibility conditions improve slowly. The model also predicted a small probability of ATFM regulation due to the weather from 8PM to 9PM, approximately, driven by a significant

drop in visibility. Even though the weather conditions were similar, if not worse, than those during the morning period, the probability of regulation due to weather was much lower because of the low traffic demand. This fact reveals that the model learned from historical data that when demand is low, no ATFM regulation is required, even in bad weather.

Readers seeking a deeper understanding of the confident learning method and the regulation prediction model are advised to consult the original publication.

3 Literature review

Airport capacity modelling and prediction have been extensively addressed in the literature by employing diverse models and data sources. For instance, over a decade ago, (Wang, 2011; Wang, 2012) presented real case studies that assessed the impact of weather on airport capacity using quadratic response surface linear regression and random forests. These models were trained using the weather data from a rapidly updated cycle forecast.

Almost concurrently, (Dhal, 2013) attempted to classify airport capacity into *low*, *medium*, or *high* categories up to 24 h in advance. The classifier was built using multinominal logistic regression techniques, with the following variables extracted from terminal area forecasts (TAFs) as predictors: the presence of thunderstorms, cloud ceiling, visibility, temperature, and wind direction and speed. The models that comprise the system presented herein use similar information. However, it is important to note that our models were trained using actual weather observations, rather than forecasts. The rationale behind this decision, along with its proper justification, is discussed in detail later in this paper.

Right after, (Kicinger, 2014) presented a stochastic analytical model for generating probabilistic airport capacity predictions, specifically for strategic traffic flow planning. This study was further extended to a subsequent study (Kicinger, 2016). The proposed model incorporates various types of weather forecast inputs such as deterministic forecasts, deterministic forecasts with forecast error models, and ensemble forecasts. Still on the subject of probabilistic predictions, (Tien, 2018) explored the use of ensemble weather products to quantify uncertainty in airport capacity predictions. Interestingly, the authors demonstrated the viability of using a generic model when airport-specific models were not feasible owing to the lack of data. In a more recent study, (Choi, 2021) proposed the use of artificial neural networks (ANNs) for airport capacity prediction.

Similarly, the authors of (Schultz, 2021) employed machine learning to classify airport performance. As in the study by (Choi, 2021), the authors used ANNs. However, it is essential to note that the primary objective of the authors was not to predict the precise value of airport capacity in movements per hour (i.e., to solve a regression problem) but rather to classify airport performance into various categories, similar to the work of (Dhal, 2013). Weather data were derived from local meteorological reports, whereas airport performance was determined from both flight plan data and reported delays.

Predicting the likelihood of ATFM regulations due to weather in the airspace sector has also been addressed in the past. For instance, (Jardines, 2021) proposed regression and classification models to predict airspace performance characteristics, including entry count, number of flights impacted by weather regulations, and activation of regulations due to weather. Similarly, (Mas, 2021) presented a machine-learning model to capture the relationship between traffic demand, weather, and the presence of ATFM regulations. To the best of our knowledge, the first attempt to predict ATFM regulations at airports was proposed by (Lattrez, 2022). Their model, also built on ANNs, learned the presence or absence of ATFM regulations at airports from historical observations.

4 Architecture of the system

This section outlines the architecture of the system, providing precise details about the input data, constituent models, and information delivered to users through a dedicated human-machine interface. Figure 3 serves as a graphical guide for this section.

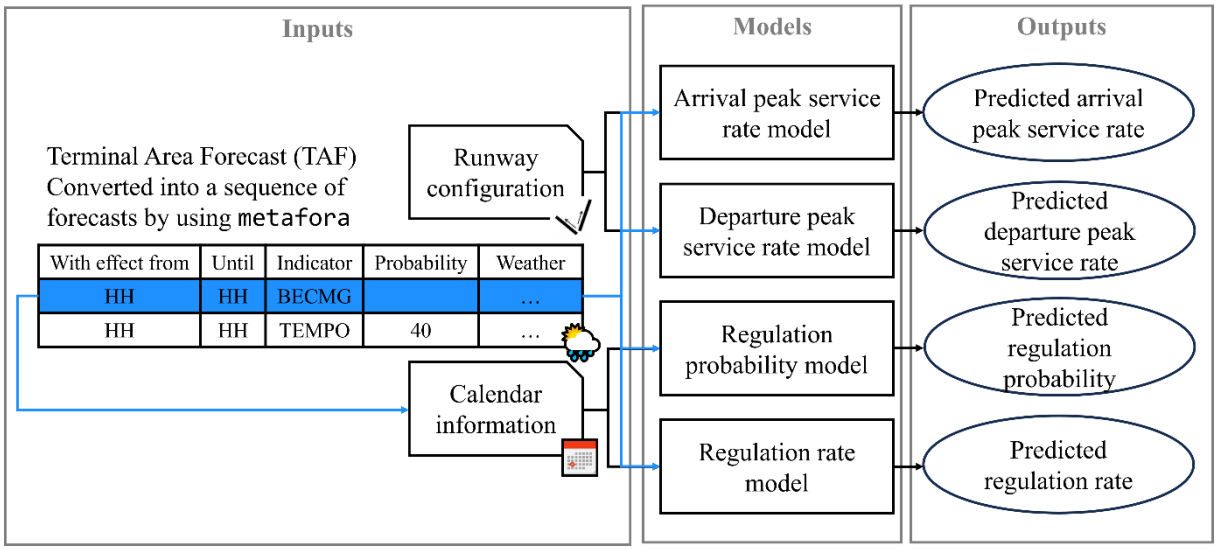


Figure 3. Architecture of the system.

As shown Figure 3, the inputs of the system include:

- Terminal area forecasts: TAFs are specialized weather forecasts created in the vicinity of an airport. They provide detailed meteorological predictions within a 5-mile radius of an airport's runway complex, making them an indispensable tool for aviation planning and operations. TAFs are typically issued four times per day: 0000Z, 0600Z, 1200Z, and 1800Z. These forecasts typically cover a 24-hour period, but in some cases, they can go up to 30 h. TAFs often include multiple time periods (referred to as TAF periods below), each with different weather indicators (e.g., BECMG and TEMPO) and associated probabilities. For each TAF period, TAFs not only provide standard weather elements, such as wind speed and direction, ceiling, and visibility, but they also provide information on significant weather phenomena that may have an impact on aviation safety, such as thunderstorms, fog, or snow. Table 2 shows an illustrative example of TAF for the Barcelona-El Prat (LEBL) airport.

Table 2. Illustrative example of TAF for Barcelona-El Prat airport.

LEBL 070500Z 0706/0806 33006KT 9999 FEW020 TX15/0713Z TN08/0706Z
BECMG 0710/0713 13010KT
TEMPO 0801/0806 3000 SHRA BKN015 FEW020TCU
PROB 40 TEMPO 0804/0806 18015G25KT

This TAF was issued on the 7th day of the month at 0500Z and remains valid for the subsequent 24 h. It comprises four distinct TAF periods, each of which is delineated on a separate line. The initial period commences one hour after issuance and extends until the following day at 0600Z, detailing prevailing weather conditions. The second period forecasts a transition (BECMG) occurring at 1000Z, fully manifesting by 1300Z. Notably, the prevailing wind, initially reported as 6 kt from the NNW, shifted to 10 kt from the SE, while visibility (9999 m) and sky cover (few clouds at 2000 ft) remained constant. A temporary (TEMPO) change is outlined in the third line, effective from the 8th day of the month at 0100Z until 0600Z. This interval anticipates reduced visibility

(3000 m), along with showers, rain, broken clouds at 1500 ft, and a few clouds at 2000 ft with a tower cumulus. Post 0600Z, the conditions are forecast to revert to those predating the temporary alteration. A subsequent temporary change is forecasted from the 8th day of the month from 0400Z to 0600Z, with a 40% probability. This indicates the possibility of variable weather conditions (wind of 15 kt from S and gusts of 25 kt) during this period.

Converting raw (textual) TAFs into a format suitable for feeding machine-learning models poses a significant challenge. Metafora , an open-source Python tool, was specifically developed for this task. Using metafora, the information included in the TAFs and necessary for feeding the models could be easily extracted.

The various TAF periods are then merged with runway configuration data or calendar information to feed the peak service rate and ATFM regulation models, respectively, and generate predictions. The data are described as follows:

- Runway configuration: Let the term *runways in use* refer to the set of runways where movements were observed during a specific time window, and *runway configuration* refer to an official combination of runways that can be utilized according to the airport. For instance, when the runway configuration is 14 / 16 & 28, runway 14 can be utilized for arrivals and runways 16 and 28 for departures. However, there might be situations when, during a specific time window with runway configuration 14 / 16 & 28, no movement is observed in runway 28. That is, the runways in use were 14 / 16, while the runway configuration were 14 / 16 & 28.

  As stated previously, the peak service rate models were trained with the most likely runway configuration that matched the Airport Corner. This implies that these models require future runway configurations to make predictions. Instead of predicting the future runway configuration and then feeding that information into the models, all possible runway configurations were combined with weather data from various TAF periods to generate predictions. As a result, peak service rate models produce as many predictions as combinations of TAF periods and airport runway configurations. In the previous example with four TAF periods, each of the peak service rate models produced 4×3 predicted capacities given that LEBL has three possible runway configurations. This was done to inform airport operators of the airport's capacity for various runway configurations.
- Calendar information: Traffic demand is strongly correlated with calendar features such as the hour of the day, day of the week, and month of the year. Such information must be provided to models that predict the probability and rate of ATFM regulation. Obviously, if demand is very low (e.g., at night), the likelihood of ATFM regulation is very low, even if weather conditions are bad. Each TAF period is divided into 1-hour sub-periods, after which the aforementioned calendar features are generated and fed into the models that predict ATFM regulation information. The preceding LEBL example yielded two predictions for the TEMPO period between 0400Z and 0600Z (one for 0400Z–0500Z and one for 0500Z–0600Z), five predictions for the TEMPO period between 0100Z and 0600Z, etc..

The four models that comprise the system use these data to forecast different quantities related to an airport's state. These are as follows:

- Arrival peak service rate model: This model predicts the 99th percentile of the arrival throughput, that is, the arrival peak service rate, measured in arrivals per hour. This value is a proxy for the actual arrival capacity, which varies depending on the combination of weather conditions and runway configuration. Peak service rate prediction allows airport

operators to identify periods when the predicted arrival capacity may fall below the expected demand and cause operational issues.

- Departure peak service rate model: The same as the previous model, but for departures.
- Regulation probability model: This model predicts the likelihood (in probabilistic terms) that a regulation will be implemented based on the expected weather conditions and calendar information that acts as a proxy for traffic demand. From the airline's perspective, this output indicates the potential impact on flights (e.g., ground or airborne delay), whereas flow managers perceive it as a signal to activate a regulation.
- Regulation rate model: This model predicts the potential entry rate (in arrivals per hour) of hypothetical weather-related regulations. Put simply, this model answers the following question: 'What is the expected rate that a (human) flow manager would use in these weather conditions if a regulation were in place?'. Notably, the predictions of this model must be considered in conjunction with those of the regulation probability model. That is, if the predicted probability of regulation is very low, the predicted regulation rate is irrelevant, because no regulation is expected. It should also be noted that the value predicted by this model is not the optimal rate but rather what has previously been observed (regardless of whether it was a good or bad decision). In other words, if flow managers consistently use overly optimistic or pessimistic entry rates, the model captures and mimics this behavior. Therefore, the predictions of this model must be used with caution. This differs from the peak service rate model, which learns directly from the observed throughput, although the actual throughput may be indirectly influenced by the entry rate during the regulated periods. Interestingly, the peak service rate and regulation (entry) rate model predictions can be combined to identify specific weather conditions and periods of time when rates that are too low or too high are frequently implemented in comparison to the airport's actual throughput. This type of analysis, outside the scope of this study, may help airport operators to fine-tune entry rates and select values that are more aligned with empirical capacities.

5 Machine learning specifics

This section covers the machine learning aspects of the system presented in the previous section. Section 5.1 introduces the dataset used to train the models, while Section 5.2 outlines the *base* model utilized to correlate the input features with the targets, leading to the development of the four independent models.

Table 3 provides a summary of the input features (predictors), target variables (outputs), and loss functions used by the various models discussed. This table serves as a key reference in this section.

Table 3. Input features, target and loss function of the various models.

	Model			
Features	Arrival peak service rate	Departure peak service rate	Regulation probability	Regulation rate
Runway configuration	✓	✓	✗	✗
Weather conditions	✓	✓	✓	✓
Calendar information	✗	✗	✓	✓
Airport identifier	✓	✓	✓	✓
Target	Arrival throughput	Departure throughput	Presence of regulation	Regulation rate

	Model			
Loss function	Mean pinball error	Mean pinball error	Binary cross-entropy	Mean squared error

5.1 Dataset

The dataset used to fit the models comprised a collection of n observations, $\boldsymbol{X} = (\boldsymbol{x}, y)^n$, where each observation contained a vector of input features, $\boldsymbol{x}$, along with the corresponding vector of targets, y. It should be noted that each model was trained with a subset of the input features (e.g., the peak service rate considers weather-related features and runway configuration, whereas the other models were trained on weather-related features alongside calendar information) to predict the corresponding target.

Each observation in $\boldsymbol{X}$ is associated with a 1-hour time window, with each window starting 15 min after the previous one. The dataset focuses on time windows spanning from 4AM to 10PM local time (i.e., [4:00, 5:00)AM, [4:15, 5:15)AM, [4:30, 5:30)AM, etc.). Thus, the dataset excludes night operations because demand is typically low during these hours, rendering the peak service rate an unreliable proxy for capacity, and the probability of ATFM regulation and expected entry rate irrelevant.

The dataset spans from June 15th, 2021 (right after the lifting of COVID-19 restrictions) to January 31st, 2024, and covers the top-50 busiest airports in Europe during 2022 according to Wikipedia[2]. For the train-test split, the observations were randomly assigned, with 80% allocated to the training set and 20% to the test set. However, to avoid information leakage from the training set into the test set, a constraint was implemented to guarantee that all observations pertaining to the same airport and date (e.g., Zurich airport on June 15th, 2021) were exclusively assigned to either the training or test set.

It should be noted that the same four models were used for 50 airports. Thus, a generic universal model was developed for each target. This approach was selected because the amount of data for training airport-specific models may be insufficient to achieve optimal performance. Moreover, the pragmatic feasibility of maintaining distinct models for each airport is questionable in the context of machine-learning operations. Because the departure and arrival peak service rates, as well as the probability and rate of ATFM regulations, are heavily influenced by the airport, one of the input features of the models is the airport identifier. The following sections discuss the target and input features of the dataset in more detail.

5.1.1 Targets

The vector of targets is composed of four elements: (observed) departure and arrival throughput, presence/absence of regulations, and regulation rate.

Delivered throughput

Note that the peak service rate models are tailored to forecast the 99th percentile of throughput based on the input features. Consequently, the initial two targets in the dataset are the observed arrival and departure throughput, representing the number of arrivals and departures observed within each 1-hour interval. To capture the 99th percentile of these targets, a dedicated loss function for quantile regression tasks was employed. Further elaboration is provided in Section 6.

Historical arrival and departure data were extracted from the airport operator data flow (APDF) system, which facilitates seamless information exchange among various entities involved in airport

[2] https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_Europe. Last accessed on March 7th, 2024.

operations. In simpler terms and without delving extensively, APDF serves as a central hub for sharing crucial details such as flight schedules, aircraft movements, passenger lists, baggage tracking, ground handling requests, weather updates, security alerts, and operational performance metrics.

Among the wealth of information provided by the APDF are the planned and actual times for each movement (departure or arrival), as well as the specific runway utilized. For instance, flight VY8002 planned to depart Barcelona at 8:00AM took-off at 8:05AM from runway 25R". This information is essential for dataset construction. More specifically, only the actual movement times are required to compute the historical arrival and departure throughput. However, runway information is also important for calculating runway configurations, as will be elaborated upon later.

In cases where APDF is not readily available, the arrival and departure throughput can be easily extracted from the automatic dependent surveillance broadcast (ADS-B) data.

Presence and rate of regulations

The presence of ATFM regulations owing to weather is indicated by a binary variable that reflects whether a regulation was active for each observation in the dataset. Recall that each observation covers a 1-hour period at a particular airport.

For positive observations, the corresponding rate is derived from the attributes of the regulation. This rate, represented as a non-negative integer, indicates the maximum number of hourly arrivals that the airport can accommodate, as perceived by the human operator that activates the regulation. However, for negative observations, the rate was absent, rendering these observations unsuitable for training the regulation rate model. Put simply, the number of observations available to effectively train the regulation rate model is significantly fewer than that of the other three models.

It is worth noting that historical ATFM regulation data were sourced from the Network Manager (NM), but are not publicly accessible. However, because real-time ATFM regulation data are available to several stakeholders, there may be the potential to construct a similar dataset.

5.1.2 Features

The features in the dataset were divided into four categories based on their nature. The categories are detailed below.

Runway configuration

Unfortunately, no dataset contains historical runway configuration changes for all airports. However, this valuable information for peak service rate models can be estimated based on the observed runways. The method described by (Dalmau, 2023a) was used to determine the runway configuration based on the observed runways in use.

The runways were deduced from individual arrival and departure operations. The specific runway utilized for each operation was extracted from the APDF data. In cases where APDF is not readily available, the landing and take-off runways can be extracted from ADS-B using the `takeoff_from_runway` and `aligned_on_ils` methods, respectively, implemented in open-source tool *traffic* (Olive, 2019). However, a comparison between the APDF and ADS-B revealed that the former was more complete.

Weather conditions

The system described in this study is specifically intended to help operators manage adverse weather conditions. As a result, weather conditions were the primary features of the various models (Table 3). The weather conditions for each observation in the dataset were extracted from

METARs, as the objective of the system was to model the cause-effect relationship between the input features and targets.

It can be agreed upon that the throughput of 28 arrivals per hour during a specific time frame at Zurich airport (for example) was influenced not by the weather forecast made several hours prior but by the actual weather conditions at that precise moment. Therefore, if the objective is to understand the causal relationship between weather and the peak service rate, it would be more beneficial to use METARs for training.

A model trained on TAFs would essentially learn the correlation between the weather forecast and peak service rate, inadvertently learning the inherent inaccuracies of the TAF in the process. While this might seem advantageous at first glance, the authors believe that it is not the best approach. If improvements were made to the TAFs by any meteorologist or system in the future, the model would continue to adjust for the errors observed during its training phase, which is not a desirable property. In contrast, a model trained on METARs simply yields more precise predictions as the quality of the weather forecast improves. This approach ensures that the model's performance improves in tandem with the accuracy of the weather forecast rather than compensating for past inaccuracies that may no longer be applicable.

However, the models are designed to be utilized several hours before operations commence, with the objective of making future predictions. At this point, METARs become irrelevant, as they only offer information about current weather conditions. Consequently, the models must be equipped with TAFs. For this reason, only the information present in both METARs and TAFs (i.e., the intersection of attributes) can be used to train the models. This ensures that the models are trained on the relevant data that will be available when they are put into operation.

METARs, which are used for training, and TAFs, which are employed for inference during operations, are processed using the *metafora* tool. This process is performed to extract a set of shared features, which are as follows:

- Wind speed (m s^{-1}),
- wind compass, including N, E, NE, and NNE (categorical),
- wind gust (m s^{-1}),
- visibility distance (m),
- clouds amount, also known as sky cover (oktas),
- clouds height, also known as ceiling (m),
- cloud vertical visibility, which was only reported when the sky was obscured (m).
- presence/absence of obscuration, such as fog (boolean),
- presence/absence of precipitation, such as rain (boolean),
- presence/absence of thunderstorms (boolean),
- presence/absence of snow (boolean),
- presence/absence of freezing conditions (boolean),
- presence/absence of significant clouds, such as cumulonimbus (boolean), and
- presence/absence of other phenomena such as volcanic hashes or tornadoes (Boolean).

Calendar information

Peak service rate models should be independent of traffic demand. This is because airport capacity, given a specific runway configuration, is influenced by weather conditions rather than by demand. Conversely, the likelihood of flow managers activating regulations and the entry rate are both dependent on traffic demand.

When demand is low, regulations are unnecessary, even under extremely adverse weather conditions, because there are no flights requiring regulation. To maintain system simplicity and allow users to operate it without direct access to traffic demand data, calendar information was used as a surrogate. The logic behind this decision is that, if both the probability of regulation and

its rate correlate with traffic demand, and traffic demand correlates with calendar information, then by extension, the probability of regulation and its rate also correlate, albeit indirectly, with calendar information. The following categorical features are used to characterize the calendar information:

- Hour of the day,
- day of the week, and
- month of the year.

It is worth noting that the notion that circular features, such as hour of the day, day of the week, or month of the year, always require transformation using sine and cosine functions is often misunderstood. While this transformation is commonly used in neural networks to capture periodicity, decision-tree-based algorithms can effectively handle circular features without the need for explicit transformation.

Airport identifier

Finally, it is important to note that even under identical weather conditions, the peak service rates can vary significantly between the two airports. Similarly, given the same weather conditions and traffic demand, the likelihood of ATFM regulation and its rate can change substantially across different airports. Therefore, it is crucial for the model to be supplied with a specific airport identifier for each observation during training as well as for which airport predictions are made during inference. In the dataset, the airport identifier was represented as a categorical feature.

The authors recognize that the determinants of the peak service rate, as well as the probability and rate of a regulation, are not the airport's identifier (e.g., LEBL for Barcelona) but its characteristics (e.g., the number and orientation of runways, geographical location, airport equipment, etc.). Therefore, instead of supplying the model with an airport identifier (i.e., a surrogate), one can convert this single, specific feature into multiple generic features that best represent the airport's characteristics for a given task. This approach is anticipated to enhance the model's generalization and potentially boost performance, despite increasing system complexity. The implications of employing this universal approach will be explored in future studies. As previously discussed, if the targets correlate with an airport's characteristics, theoretically, knowing the identifier should be sufficient to indirectly learn this correlation.

6 Models

A generic base model serves as the foundation for training four distinct models that compose the system. Each model underwent independent training from scratch using the corresponding input features to predict the respective target. The sole distinction between these individual training sessions is the loss function employed.

For the peak service rate model, which addresses a quantile regression task, the mean pinball error (MPE) function was utilized with parameter $\alpha = 0.99$:

$$\text{MPE} = \frac{1}{n'}\sum_{i=1}^{n'}(\alpha \max(y_i - \hat{y}_i, 0) + (1-\alpha)\max(\hat{y}_i - y_i, 0)), \quad (3)$$

where y_i is the actual throughput of the i^{th} observation, $\hat{y}_i$ is the predicted 99$^{\text{th}}$ quantile, and n' is the number of observations used to compute the loss.

The model predicting the probability of regulation, which tackles a binary classification task, employs binary cross-entropy (BCE) as the loss function.

$$\text{BCE} = \frac{1}{n'}\sum_{i=1}^{n'}(y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)), \quad (4)$$

where y_i represents the presence (1) or absence (0) of regulation in the i^{th} observation and $\widehat{p_i}$ represents the predicted probability that the i^{th} observation belongs to class 1.

Finally, for the regulation rate model, which solves a standard regression task, the mean squared error (MSE) was adopted as the loss function.

$$\text{MSE} = \sum_{i=1}^{n'} \frac{1}{n'} (y_i - \widehat{y_i})^2, \quad (5)$$

where y_i and $\widehat{y_i}$ are the actual and predicted rates, respectively, of the i^{th} observation.

Many machine-learning models can be configured to handle quantile regression, binary classification, and standard regression tasks. The base model proposed in this study is based on ensemble methods that produce a strong learner from a group of weak learners. Boosting is a well-known ensemble method that involves training a series of weak learners (e.g., rudimentary decision trees). The training observations for the next learner in traditional adaptive boosting (AdaBoost) (Schapire, 2013) are weighted based on how well the previous learners performed; that is, observations that correspond to incorrect predictions are assigned more weight to concentrate the model's attention on correcting them. Gradient boosting differs from AdaBoost in that instead of assigning weights to observations based on performance, a new learner is trained at each iteration to fit the residual errors of the preceding learners. This ensemble is known as the GBDTs model when decision trees are used as weak learners.

In comparison to simpler models, such as linear regression or conventional decision trees, GBDTs have emerged as the preferred choice because of their capacity to capture complex and nonlinear interactions within the data. GBDTs also possess highly desirable attributes, including the ability to handle missing data effectively and categorical features with high cardinality, as exemplified by the feature representing the airport identifier. Their robustness in the presence of outliers was another pivotal factor that ensured that data anomalies would not unduly influence the results. Furthermore, GBDTs have consistently demonstrated outstanding performance across various practical applications, particularly on tabular datasets, where each row represents an individual observation and each column represents a distinct feature (Lundberg, 2020). It is noteworthy that, while GBDTs may not provide the same level of interpretability as simpler models, techniques such as the Shapley method can still yield valuable insights. Among all GBDTs implementations, Microsoft's *LightGBM* was selected. The reader is referred to (Guolin, 2017) for more information about the distinctive features of *LightGBM* when compared to other popular implementations, such as *XGBoost* (Chen, 2016) or *CatBoost* (Prokhorenkova, 2018).

Notably, *LightGBM* simplifies the encoding of categorical features by advocating straightforward integer encoding. Subsequently, *LightGBM* employs the methodology introduced by (Fisher, 1958) to identify the optimal splits among integer-encoded categories. In accordance with *LightGBM* best practices, integer encoding was applied to non-Boolean categorical features, that is, airport identifiers, wind compasses, and runway configurations. Integers were assigned to airports and wind compasses in alphabetical order; for example, EBBR was mapped to 0, EDDB to 1, etc.. In contrast, integers for runway configurations were assigned based on frequency, which means that runway configuration 0 at EBBR, for example, is the most frequently used configuration at that airport, and 1 is the second most common configuration.

7 Results

The following subsections summarize the performance of the models and their feature attributes. Several illustrative examples are presented and discussed.

7.1 Performance

Table 4 summarizes the performance of the four models in the test set, highlighting the representative metrics and airports.

The metrics include the arrival and departure peak service rates, regulation probabilities for both primary and secondary services, and the mean absolute error (MAE) for the regulation rate. These measures provide a comprehensive evaluation of the predictive accuracy and robustness of the model across different airports. The table below presents these performance indicators in a concise format.

Table 4. Assessing model performance on the test set: for peak service rate models, values in parentheses denote hypothetical performance based on Airport Corner values. Regulation probability and rate models show results solely for airports with over 50 regulated hours due to weather in the test set, ensuring statistically significant evidence.

Airport	Arrival peak service rate	Departure peak service rate	Regulation probability (PS)	Regulation probability (RS)	Regulation rate (MAE)
EBBR	0.18 (0.37)	0.17 (0.32)	-	-	-
EDDB	0.13 (0.38)	0.13 (0.38)	-	-	-
EDDF	0.29 (0.32)	0.28 (0.31)	0.70	0.22	5.4
EDDH	0.11 (0.23)	0.12 (0.23)	-	-	-
EDDK	0.14 (0.25)	0.24 (0.26)	-	-	-
EDDL	0.17 (0.23)	0.17 (0.26)	-	-	-
EDDM	0.26 (0.38)	0.28 (0.38)	0.42	0.23	8.2
EFHK	0.21 (0.31)	0.22 (0.31)	0.82	0.32	2.8
EGBB	0.09 (0.15)	0.1 (0.14)	-	-	-
EGGW	0.1 (0.12)	0.12 (0.14)	0.67	0.32	1.5
EGKK	0.14 (0.22)	0.18 (0.23)	0.68	0.38	3.0
EGLL	0.16 (0.19)	0.19 (0.19)	0.73	0.49	2.7
EGPH	0.09 (0.11)	0.13 (0.18)	-	-	-
EHAM	0.31 (0.53)	0.29 (0.68)	0.81	0.50	8.6
EIDW	0.13 (0.17)	0.21 (0.23)	0.85	0.23	4.1
EKCH	0.15 (0.34)	0.17 (0.34)	-	-	-
ENGM	0.17 (0.26)	0.18 (0.25)	0.66	0.19	2.8
EPWA	0.15 (0.16)	0.16 (0.19)	-	-	-
ESSA	0.15 (0.31)	0.16 (0.3)	0.72	0.39	3.1
LEAL	0.12 (0.13)	0.11 (0.12)	-	-	-
LEBL	0.16 (0.19)	0.17 (0.2)	0.88	0.32	4.5
LEMD	0.26 (0.26)	0.33 (0.34)	0.33	0.27	6.0
LEMG	0.13 (0.15)	0.14 (0.14)	-	-	-
LEPA	0.19 (0.19)	0.18 (0.19)	-	-	-
LFML	0.07 (0.08)	0.1 (0.16)	-	-	-
LFMN	0.13 (0.2)	0.13 (0.2)	-	-	-
LFPG	0.3 (1.19)	0.25 (0.95)	0.73	0.48	7.2
LFPO	0.16 (0.24)	0.15 (0.23)	0.71	0.35	3.6
LGAV	0.15 (0.44)	0.16 (0.56)	-	-	-
LHBP	0.09 (0.31)	0.11 (0.44)	-	-	-
LICC	0.07 (0.07)	0.07 (0.1)	-	-	-
LIMC	0.13 (0.28)	0.14 (0.18)	-	-	-
LIPZ	0.09 (0.14)	0.09 (0.16)	-	-	-
LIRF	0.21 (0.36)	0.19 (0.42)	-	-	-
LIRN	0.09 (0.09)	0.09 (0.09)	-	-	-
LKPR	0.11 (0.26)	0.12 (0.26)	-	-	-

LOWW	0.16 (0.33)	0.18 (0.35)	0.46	0.11	4.7
LPPR	0.08 (0.08)	0.1 (0.12)	0.81	0.63	2.3
LPPT	0.09 (0.1)	0.1 (0.1)	0.79	0.48	2.5
LROP	0.07 (0.07)	0.13 (0.15)	-	-	-
LSGG	0.11 (0.15)	0.11 (0.3)	0.77	0.2	3.4
LSZH	0.15 (0.23)	0.16 (0.24)	0.80	0.57	3.2
LTFM	0.24 (0.27)	0.3 (0.33)	0.78	0.48	8.5

In this table, the MPE measures the effectiveness of the peak service rate models, whereas the mean absolute error (MAE) measures the performance of the regulation rate model. When assessing the regulation prediction model, threshold metrics were utilized to measure the proportion of instances where a predicted class (positive or negative) deviates from the actual class. Precision (PS) and recall (RS) were employed for this evaluation. The precision answers the question: *what proportion of positive predictions was actually correct*? whereas recall answers the question: *what proportion of actual positives was predicted correctly?*. In both cases, a 50% threshold was used, meaning that a predicted probability greater than 50% indicated a positive class, and less than 50% indicated a negative class.

In Table 4, the performance of the peak service rate models is benchmarked against a baseline model derived from the Airport Corner tables. The advantage of using the Airport Corner tables as a baseline lies in the fact that the runway configurations considered by the machine learning model precisely align with those defined in Airport Corner, owing to the matching process described in (Dalmau, 2023a). For each observation in the test set, the capacity (for arrivals or departures) reported in the Airport Corner for the inferred runway configuration can be contrasted with the peak service rate prediction generated by the machine learning model, which considers both the runway configuration as well as the weather conditions.

Table 4 demonstrates that the machine learning models provide more reasonable estimations of the peak service rate compared to the capacities as defined in the Airport Corner in terms of mean pinball error. The extent of the performance improvement varies depending on the airport. For instance, at Paris Charles De Gaulle airport (LFPG), machine learning models exhibit a significantly lower mean pinball error in the test set compared with the baseline model, whereas at Porto airport (LPPR), the machine learning models are almost identical to the Airport Corner tables.

Regarding the performance of the regulation probability and rate models, only airports that have experienced over 50 h of regulation due to weather throughout the clean test set (i.e., after applying the CL methodology and filtering out noisy observations) are displayed in the table to ensure the statistical relevance of the results.

As shown in Table 4, the regulation probability model demonstrates modest performance across the majority of airports when trained on the clean train set and tested on the clean test set. This performance is largely influenced by the inherent complexity of the problem and limited availability of positive observations for learning. The precision of the model ranged from 42% to 88%, with recall varying between 19% and 63%. Thus, it is important to highlight that the performance of the model may not satisfy the acceptability criteria for all airports. For instance, at Madrid-Barajas (LEMD), the model achieves a precision of 33% and a recall of 27%, indicating a clear room for improvement. The acceptability criteria for precision and recall metrics are currently unknown and require validation exercises across multiple airports to establish minimum acceptable thresholds. These criteria are likely to vary significantly, depending on the operational context of each airport. Factors such as the sensitivity of air traffic flow management decisions to prediction errors, specific weather patterns affecting each region, and regulatory standards imposed by aviation authorities could all influence these thresholds. Therefore, a comprehensive assessment involving collaboration between airports and flow managers is essential for effectively defining these criteria.

It should be noted that because human operators are those who activate ATFM regulations based on weather forecasts with some lead time, the model was trained to learn the human perception of weather severity. However, instead of using weather forecasts (i.e., TAFs), the model was trained on observed weather (i.e., METARs). Based on the most frequent patterns learned by the model, CL was used to filter out observations in which the actual weather conditions were not sufficiently adverse when a regulation was in place. This strategy was put into effect under the debatable presumption that when humans choose to implement regulations, weather forecasts are typically accurate. This discussion suggests that it is unrealistic to expect the model to achieve a perfect prediction of regulations with extremely high recall and precision.

Furthermore, the precision and recall metrics in Table 4 were computed under the assumption of a 50% threshold, which means that 51% is considered a positive prediction (active regulation) and 49% a negative prediction (no regulation). Adjusting this threshold allows users to tailor the recall versus precision tradeoff according to their specific requirements. The model produces a probability rather than a positive or negative outcome. Thus, airports with a relatively high predicted probability may require special attention, but decision making is always dependent on the human interpretation of that probability based on experience. Figure 4 illustrates this idea, showing the probability of regulation with a color code as a function of the hour of the day (horizontal axis) for several airports (vertical axis). In summary, while precision and recall are useful metrics for assessing prediction quality, the model is primarily used as an advisory tool, recognizing its limitations owing to the complexity of the task.

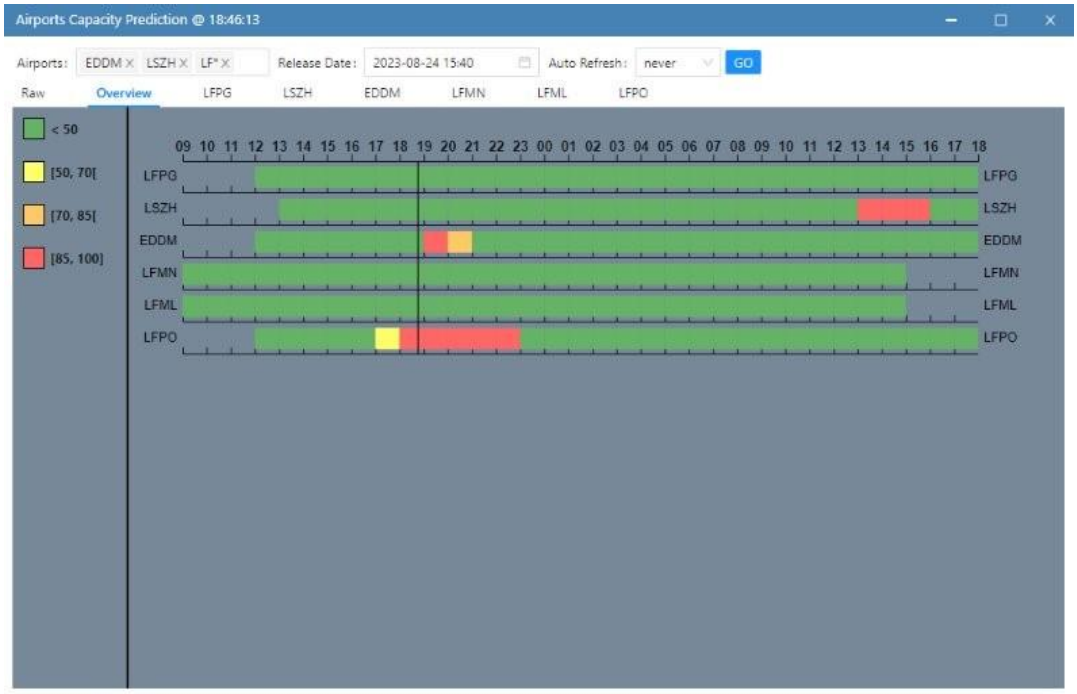


Figure 4. The current human-machine interface for the ATFM regulation probability model. The colour code corresponds to the likelihood of ATFM regulation.

Finally, the performance of the regulation probability model is significantly contingent on the train-test split. As indicated by (Dalmau, 2023b), when a smaller dataset was employed, the results varied considerably. Some airports demonstrated improved performance, while others showed a decline. It is generally postulated that an increase in the size of the dataset reduces the model's sensitivity to the train-test split, thereby ensuring more consistent performance across different splits.

7.2 Feature attribution

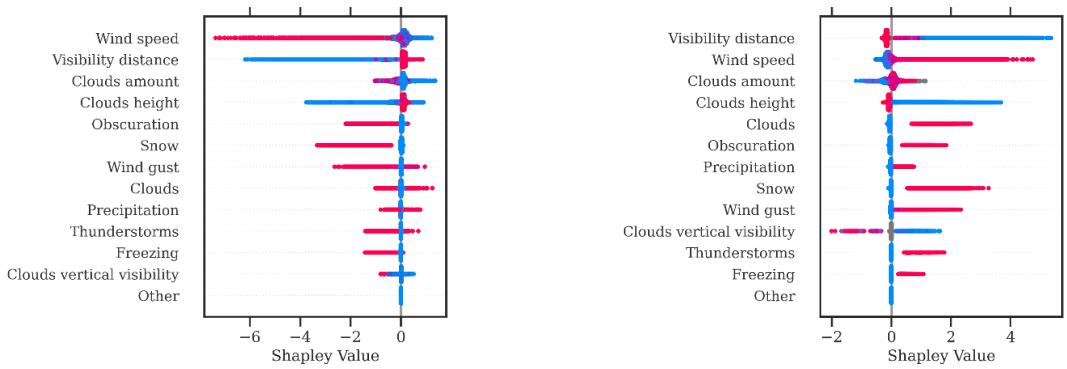
The principles of game theory can be used to interpret the prediction of any machine model for a given observation, assuming that each feature is a player in a game and the output of the model is the payout. Let us consider the following scenario: all players participate in the game and join the game in random order. The attribution of a player is the average change in the payout received by players in the game when they join them. More formally, the Shapley value $\phi_i(\boldsymbol{x}, \boldsymbol{\theta})$ of feature i for a given input vector $\boldsymbol{x}$ and model parameters $\boldsymbol{\theta}$ is defined as the expected marginal contribution of i to the prediction across all possible feature permutations (Lundberg, 2020). In other words, the Shapley value quantifies how much each feature contributes to the model output, on average, when considering all possible subsets of features that include i:

$$\phi_i(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{S \subseteq 1,2,\ldots,d \setminus i} \frac{(d-|S|-1)!|S|!}{d!} \left(f\left(\boldsymbol{x}^{(S \cup i)}, \theta\right) - f\left(\boldsymbol{x}^{(S)}, \boldsymbol{\theta}\right) \right), \quad (6)$$

where S is a subset of features not including i and $\boldsymbol{x}^{(S)}$ denotes the input vector with only the features in S.

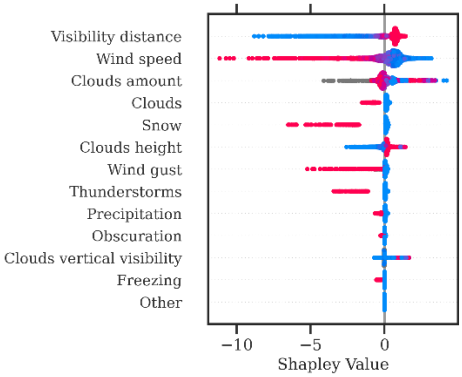
In practical applications, precisely computing the Shapley values is a computationally intensive task. To address this issue, a new explanation method called *TreeExplainer* has been developed for tree-based models such as GBDTs. *TreeExplainer* can approximate Shapley values in polynomial time and is used in the referenced paper. Further details regarding *TreeExplainer* can be found in (Lundberg, 2020). It is worth noting that the Shapley values are given in the same units as the output of the model: movements per hour for the peak service rate models; the logit[3] for the probability of regulation model; and maximum arrivals per hour for the regulation rate model.

Figure 5 depicts the distribution of Shapley values for both the numerical and Boolean features associated with weather conditions, encompassing all observations within the test set. The y-axis indicates the name of the features in order of the mean absolute Shapley value from the top to the bottom. Each dot on the x-axis shows the Shapley value of the associated feature on the prediction for one observation, and the color indicates the magnitude of that feature: red indicates high, while blue indicates low. By definition, positive values (resp. negative). Shapley values increased (resp. decrease) prediction with respect to the expected value of the target in the training set. To streamline the interpretation of the results, only the arrival peak service rate results were presented. However, the findings for departure peak service rates are similar and comparable.



(a) Arrival peak service rate *(b) Regulation probability*

[3] In a machine learning model, logits represent the raw scores outputted by the model before being transformed into probabilities. Logits are converted into probabilities using the softmax function, which ensures that the values range between 0 and 1 and sum up to 1 across all classes.



(c) Regulation rate

Figure 5. The current human-machine interface for the ATFM regulation probability model. The colour code corresponds to the likelihood of ATFM regulation.

As depicted in Fig. 5(a), the wind speed significantly influences the arrival peak service rate, with stronger winds correlating with lower arrival peak service rates. This is followed by the impact of visibility, sky cover (cloud amount), and ceiling (cloud height). Lower ceilings and reduced visibility typically result in a decrease in the peak arrival service rate. Figure 5(a) further illustrates that the presence of snow, thunderstorms, obscuration, and freezing noticeably impaired the arrival peak service rate. However, the effects of precipitation and significant clouds remain unclear. Interestingly, certain weather phenomena, such as tornadoes, appear to have a minimal impact on the arrival peak service rate. It is important to bear in mind that these rare events are seldom encountered during training. Therefore, any conclusions drawn regarding the influence of these factors should be interpreted with caution.

Figure 5(c) presents patterns that, although similar, are not identical to those in Fig. 5(a). Visibility distance appears to be the most influential factor in determining the entry rate of a regulation, followed by wind speed and sky cover. The presence of cumulonimbus clouds and snow also significantly affected the predictions of the model, typically resulting in a decrease in the predicted entry rates. In general terms, and, as anticipated, the patterns observed in Figs. 5(a) and 5(c) are comparable. This is because one would expect the entry rate of regulations implemented by human operators to align with the peak service rate (i.e., capacity) of the airport.

Finally, Fig. 5(b) shows the inverse feature attributions compared to the other two models. The Shapley values in Fig. 5(b) are expressed in logit terms and not in probability terms. In any case, a positive Shapley value indicates that the feature contributes to the prediction of the observation by increasing the probability relative to the expected value in the training set, whereas a negative value indicates the opposite.

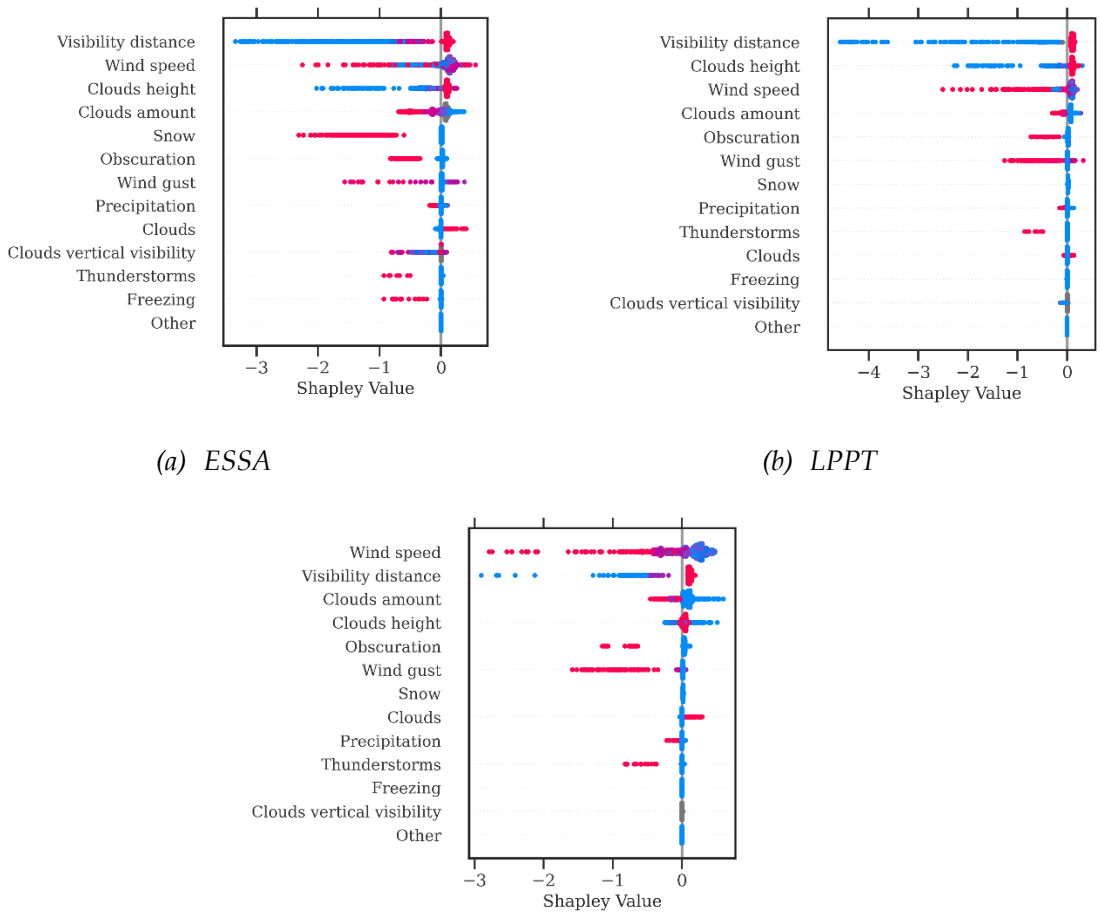
Figure 5(b) shows that the model that predicts the probability of regulation due to weather, which was trained on a clean training set with CL and subject to monotone constraints, has learned patterns that are intuitive and obvious from a human point of view. For example, when the wind speed is low, this feature has a negative impact on the predicted logit (i.e., it tends to lower the predicted probability of regulation). The impact of this feature becomes positive when its value is high. The opposite reasoning applies to visibility and ceiling. As expected, high visibility and ceiling have a negative, if not nonexistent, effect on the model's prediction. In contrast, when visibility or the ceiling value is low, the likelihood of regulation increases dramatically.

Other Boolean features, such as the presence of cumulonimbus or snow, tend to increase the probability of regulation owing to weather when their status is true, albeit with very different magnitudes. For instance, the presence of precipitation has a marginal impact on the model's

output, whereas snow can increase the logit to four for some specific predictions. Notably, the Shapley value of a feature in an observation also depends on the values of other features.

Figure 6(b) also demonstrates the effect of monotone constraints, which guarantee a consistent attribution of features and result in a model that is robust to noise in the training set. It is important to note that *LightGBM* does not currently support the implementation of monotonous constraints in quantile regressors. As a result, the models that predict peak service rates are not subject to monotone constraints, and the patterns they learn may not be as intuitive to human understanding as those learned by the weather-induced regulation model. Indeed, in Fig. 5(a), there are few observations in which the presence of thunderstorms is shown to have a positive effect on the predicted peak service rate. This behavior is counterintuitive because thunderstorms typically disrupt flight operations and reduce peak service rates.

As the Shapley values are computed per observation, it is possible to extract the Shapley values of the observations corresponding to an airport to understand the specific impact of weather. The remainder of this section compares, for illustrative purposes, the Shapley values of three airports with very different characteristics: Stockholm Arlanda Airport (ESSA), Humberto Delgado Airport (LPPT) (informally known as Lisbon Airport), and Alicante–Elche Miguel Hernández Airport (LEAL). To ensure a concise interpretation and considering the similar feature attributions observed in the peak service rate and regulation rate models, the analysis focuses solely on the arrival peak service rate and regulation probability models.



(a) ESSA *(b) LPPT*

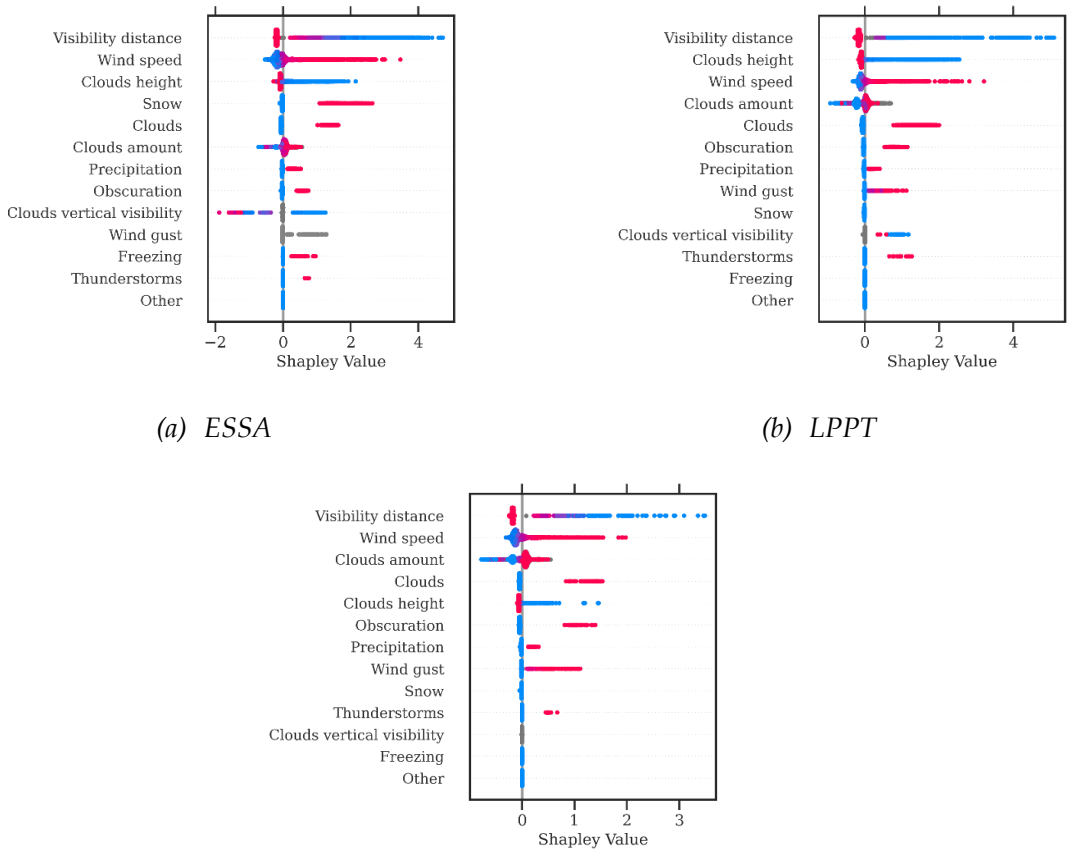
(c) LEAL

Figure 6. Shapley values distribution of the arrival peak service rate model in the observations of the test set for various airports.

As illustrated in Fig. 6, visibility emerged as the most influential factor affecting the arrival peak service rate at the ESSA, followed by wind speed and ceiling. A similar trend was observed for the LPPT, albeit with variations in the order and magnitude of these factors. The primary distinction between these two airports pertains to the effects of the snow and freezing conditions. Snow is a frequent occurrence at ESSA, often impacting operations, whereas it rarely affects operations at LPPT owing to its geographical location. Interestingly, the model successfully discerned this difference from the data, despite not being explicitly informed about airport locations.

In contrast, at LEAL, the frequency and impact of the low-visibility conditions were less significant. Instead, the wind speed and gusts seem to exert a substantial influence on the arrival peak service rate. For instance, the stormy weather on November 5th, 2023, had a significant impact on LEAL. These adverse conditions negatively affect the airport's capacity, leading to several flight diversions. Similar to the LPPT, the presence and impact of snow and freezing conditions at the LEAL are negligible. The results of this illustrative comparison highlight the importance of considering local weather patterns and geographical factors when interpreting the model predictions.

Upon comparing the Shapley values for the regulation probability model across these three airports, distinct differences became apparent. Figure 7 corroborates the findings from Fig. 5(b), highlighting visibility distance as the most influential feature affecting the probability of ATFM regulation. This consistency across the models underscores the critical role of visibility in airport operations. As anticipated, the most substantial impact was observed in the LPPT, where low visibility was prevalent.



(a) ESSA *(b) LPPT*

(c) LEAL

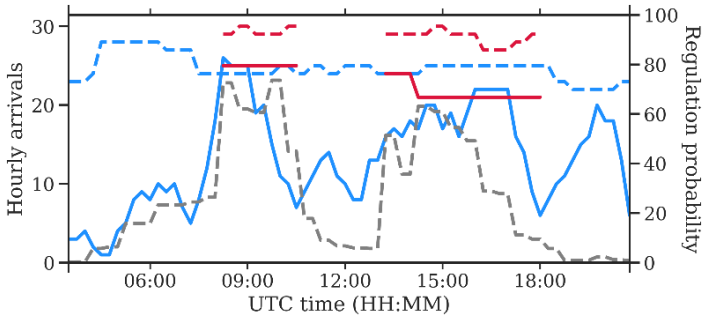
Figure 7. Shapley values distribution of the regulation prediction model in the observations of the test set for various airports.

In alignment with the previous discussion, Fig. 7(a) indicates that snow emerges as the fourth most influential feature at ESSA. The impact of this feature is negligible at both the LPTT and LEAL, primarily because of the infrequency of snowfall at these locations.

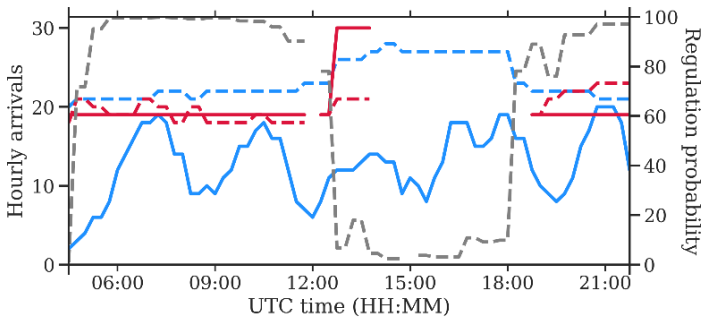
Strong winds and a low ceiling also displayed a strong correlation with the probability of ATFM regulation, ranking as some of the most influential features across all three airports in this illustrative comparison. Intriguingly, at the LPPT, the ceiling was the second most significant factor, while at the LEAL, it only ranked fifth. This variation emphasizes the distinct weather patterns and operational characteristics unique to each airport.

7.3 Illustrative examples

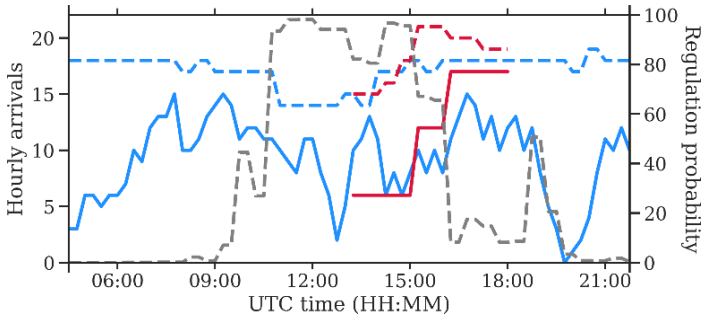
Figure 8 shows a visual representation of the system's application over multiple days for the test set at various European airports. Each instance provides a comparative view of the model's predictions and the actual observed values. This comparison focused on the accuracy and reliability of the models in various operational contexts.



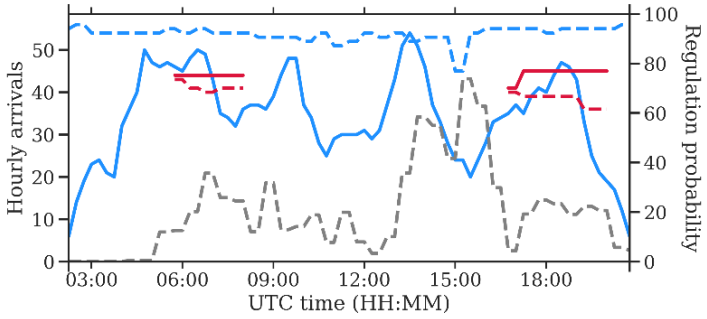
(a) LSZH on October 8th, 2021



(b) LFPO on November 11th, 2021



(c) LSGG on December 4th, 2023



(d) EDDF on August 2nd, 2023

Figure 8. Examples from the test set. Red: regulation rate. Blue: hourly arrivals (predicted peak service rate and observed throughput). Grey: regulation probability. Solid: observed. Dashed: predicted.

Figure 8(a) presents both the predictions and observations at Zurich Airport (LSZH) on October 8th, 2021. This particular day was marked by low ceiling conditions, which led to the implementation of two ATFM regulations due to weather, one in the morning and the other in the afternoon. The figure shows that the arrival peak service model was successful in detecting a decrease in capacity during an adverse weather event. Moreover, when the regulations were in effect, the probability of regulation was relatively high. This case is particularly intriguing, as the predicted rate is more optimistic than the actual rate, which corresponds more closely with the predicted peak service rate. This suggests that human operators tended to set slightly higher rates under similar conditions. However, in this instance, the chosen rate was in excellent agreement with the anticipated capacity.

Figure 8(b) shows both the predictions and actual observations at the (LFPO) on November 11th, 2021. This day was characterized by conditions of low visibility, prompting the implementation of two ATFM regulations due to weather—one in the morning and the other in the evening. In this instance, the predicted peak service rate exhibited a more significant decrease in the morning, dropping by up to eight movements per hour compared to the afternoon. Moreover, when the ATFM regulations were in effect, the predicted probability of regulation was extremely high. This underscores the confidence of the model in the severity of low-visibility events. The predicted regulation rate was also notably accurate, being slightly lower, but closely aligned with the predicted peak service rate.

Figure 8(c) provides a comparison of both the predictions and actual observations at Geneva Airport (LSGG) on December 4th, 2023. According to Table 5, this day was marked by snowfall, which significantly reduced visibility and ceiling, and had a substantial impact on operations. Notably, an ATFM regulation was implemented at approximately 1PM, yet the regulation prediction model had been forecasting a high probability of regulation from approximately 10AM. A detailed examination of the observations revealed that the weather conditions were indeed adverse at that moment, characterized by light snowfall and a low ceiling. This highlights the ability of the model to anticipate measures to respond to deteriorating weather conditions. In this instance, the observed regulation rate was lower than both the predicted rate and the peak service rates, which were notably similar. This discrepancy could be due to various factors, including operational decisions and variations in weather conditions.

Table 5. Some meteorological reports at Geneva airport on December 4th, 2023.

METAR LSGG 041350Z 23003KT 0900 R04/1800N R22/1700U SN VV005 M00/M01 Q1011
METAR LSGG 041320Z 23004KT 1000 R04/1600D R22/1600D SN VV006 M00/M01 Q1012
METAR LSGG 041250Z VRB02KT 1100 R04/1800N R22/1700N SN VV006 M00/M01 Q1012
METAR LSGG 041220Z VRB01KT 0900 R04/1200N R22/1200N SN VV005 M00/M01 Q1012
METAR COR LSGG 041150Z VRB01KT 1000 R04/1500N R22/1600D SN VV006 00/M01 Q1013
METAR LSGG 041120Z VRB01KT 1000 R04/1400D R22/P2000N SN VV011 00/M01 Q1013
METAR LSGG 041050Z VRB02KT 3500 -SN FEW008 OVC022 00/M02 Q1013
METAR LSGG 041020Z 02003KT 6000 -SN FEW012 SCT030 OVC055 00/M02 Q1014
METAR LSGG 040950Z VRB01KT 7000 -SN FEW035 OVC055 00/M02 Q1014
METAR LSGG 040920Z VRB02KT 9999 -SN FEW035 OVC055 00/M02 Q1014

Finally, Fig. 8(d) presents a compelling case for Frankfurt Airport (EDDF) on August 2nd, 2023. At first glance, the peak service rate model did not capture the decrease in capacity, and the regulation model failed to accurately predict the exact timeframe of the regulation. However, a closer examination of the METARs for that day (see Table 6) revealed that the morning began with light rain, which did not significantly affect visibility, and a ceiling ranging from 700 feet to 4800 feet. Showers and convective cloud activity started at approximately 1PM, peaking with wind gusts between 3PM and 4PM. Subsequently, the weather conditions calmed, with ceilings and visibility improving to CAVOK by around 5PM. The predicted probability of regulation and predicted capacity drop shown in Fig. 8(d) are well aligned with these observations, with both the maximum predicted probability and capacity drop coinciding with the highest observed wind gusts of the day at 3PM.

Table 6. Some meteorological reports at Frankfurt airport on August 2nd, 2023.

METAR EDDF 020520Z 17009KT 9999 -RA FEW029 BKN047 16/13 Q1002
METAR EDDF 020620Z 17009KT 140V200 9999 -RA FEW032 BKN048 15/14 Q1001
METAR EDDF 020720Z 17010KT 9999 -RA BKN008 BKN025 OVC031 16/15 Q1000
METAR EDDF 020820Z 17012KT 9999 -DZ BKN007 16/15 Q0999
METAR EDDF 021220Z 19012KT 9999 SCT006 BKN021 OVC028 18/17 Q0996
METAR EDDF 021320Z 18012KT 9999 -SHRA FEW008 BKN042 FEW///TCU 19/18 Q0995
METAR EDDF 021520Z 23022G34KT 9000 +SHRA FEW029 SCT046 FEW///CB 17/14 Q0995
METAR EDDF 021550Z 22014G26KT 9999 SHRA FEW///TCU 17/15 Q0995
METAR EDDF 021650Z 20010KT CAVOK 20/16 Q0994
METAR EDDF 20011KT CAVOK 20/17 Q0995

The first regulation of the day was activated at 4:39AM, owing to the expected convective clouds. However, it was cancelled at 6AM as the weather turned out to be better than anticipated, with convective clouds not appearing until after 1 PM. Similarly, the second regulation was activated at 3PM in response to convective cloud activity, but was cancelled before its scheduled end at 5:45PM due to an unexpected improvement in weather conditions. This example illustrates both the proactive measures taken by flow managers to prevent potential issues and their effective reactivity to avoid over-regulation and unnecessary delays when weather conditions improve unexpectedly.

8 Conclusions

This study introduces a decision-support system designed to enhance airport operations in the face of challenging weather conditions. The system employs machine learning methodologies to model airport capacity and assess the effectiveness of implementing air traffic flow management regulations and their likely entry rates.

The essence of the system lies in its ability to forecast the peak service rate (used as a proxy for capacity) depending on weather conditions, facilitating a uniform examination of operational capacity across all airports by utilizing historical data. However, it is important to note that for airports experiencing rare congestion or underuse, the peak service rate might deviate considerably from the actual operational capacity. Hence, the peak service rate is only applicable when the time period under consideration for training includes sufficient instances where the demand surpasses the actual capacity. In situations where airports are consistently underused, the peak service rate is more indicative of the peak demand than the actual operational capacity.

The system also includes a model to predict the probability of air traffic management regulations due to adverse weather, providing stakeholders with a rough indication of the severity of a weather event. To address the inherent noise in the dataset labels arising from decisions made in advance by operators using uncertain information, confident learning techniques and monotone constraints are proposed to train this model. The experiments demonstrate a modest model performance for major European airports that frequently encounter adverse weather conditions.

The final element of the system enhances the predicted regulation probability by incorporating the most likely entry rate, which is determined by the flow managers' previous implementation under comparable circumstances.

By enhancing the estimation of capacity, regulation likelihood, and potential entry rate, the proposed system offers a valuable tool for operators to make informed decisions and optimize airport operations during adverse weather conditions.

It should be noted that the system predictions were based on terminal area forecasts (TAFs). Consequently, if the system presented in this paper were utilized to predict airport capacities and support the implementation of flow measure measures (such as defining more precise entry rates), the quality of the model's predictions would be contingent on the accuracy of the TAFs. This emphasizes the need for more accurate forecasts, as operational decisions are made with some lead times.

Data Access Statement

The data that support the findings of this study are confidential and not publicly available due to privacy restrictions. Therefore, they cannot be shared for reproducibility purposes.

Contributor Statement

Ramon Dalmau: Conceptualization, Methodology, Formal Analysis, Data Curation, Software, Writing –Original Draft, and Visualization. Jonathan Attia: Conceptualization, Supervision, Resources, Project Administration, and Writing –Review & Editing.

Conflict Of Interest (COI)

There is no conflict of interest.

References

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Choi, S., & Kim, Y. J. (2021). Artificial neural network models for airport capacity prediction. *Journal of Air Transport Management, 97*, 102146. https://doi.org/10.1016/j.jairtraman.2021.102146

Dalmau, R., & Gawinowski, G. (2023c). Learning with confidence the likelihood of flight diversion due to adverse weather at destination. *IEEE Transactions on Intelligent Transportation Systems, 24*(5), 5615–5624. https://doi.org/10.1109/TITS.2023.3235741

Dalmau, R., Attia, J., & Gawinowski, G. (2023a). Modelling the impact of adverse weather on airport peak service rate with machine learning. *Atmosphere, 14*(10), Article 1476. https://doi.org/10.3390/atmos14101476

Dalmau, R., Attia, J., & Gawinowski, G. (2023b). Modelling the likelihood of air traffic management regulations due to weather at airports. In *Proceedings of the 13th SESAR Innovation Days (SID)*, Seville, Spain.

De Vivo, C., Ellena, M., Capozzi, V., Budillon, G., & Mercogliano, P. (2022). Risk assessment framework for Mediterranean airports: A focus on extreme temperatures and precipitations and sea level rise. *Natural Hazards, 111*, Article 50. https://doi.org/10.1007/s11069-021-05066-0

Dhal, R., Roy, S., Taylor, C. P., & Wanke, C. R. (2013). Forecasting weather-impacted airport capacities for flow contingency management: Advanced methods and integration. https://doi.org/10.2514/6.2013-4356

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association, 53*(284), 789–798. https://doi.org/10.1080/01621459.1958.10501479

Furtak, K., & Wolińska, A. (2023). The impact of extreme weather events as a consequence of climate change on the soil moisture and on the quality of the soil environment and agriculture – A review. *CATENA, 231*, 107378. https://doi.org/10.1016/j.catena.2023.107378

Jardines, A., Soler, M., & García-Heras, J. (2021). Estimating entry counts and ATFM regulations during adverse weather conditions using machine learning. *Journal of Air Transport Management, 95*, 102109. https://doi.org/10.1016/j.jairtraman.2021.102109

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)* (pp. 3149–3157). https://doi.org/10.5555/3294996.3295074

Kicinger, R., Chen, J.-T., Steiner, M., & Pinto, J. (2014). Probabilistic airport capacity prediction incorporating weather forecast uncertainty. https://doi.org/10.2514/6.2014-1465

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

Lattrez, O., Barragán-Montes, R., & Michalski, M. (2022). Predicting airport ATFM regulations using deep convolutional networks. In *Proceedings of the 12th SESAR Innovation Days (SID)*, Budapest, Hungary.

Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mas-Pujol, S., Salamí, E., & Pastor, E. (2021). Predict ATFCM weather regulations using a time-distributed recurrent neural network. In *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)* (pp. 1–8). https://doi.org/10.1109/DASC52595.2021.9594303

Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research, 70*, 1373–1411. https://doi.org/10.1613/jair.1.12125

Olive, X. (2019). traffic: A toolbox for processing and analysing air traffic data. *Journal of Open Source Software, 4*(44), 1518. https://doi.org/10.21105/joss.01518

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

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Proceedings of the 31st Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada.

Schapire, R. E. (2013). Explaining AdaBoost. In B. Schölkopf, Z. Luo, & V. Vovk (Eds.), *Empirical inference: Festschrift in honor of Vladimir N. Vapnik* (pp. 37–52). Springer. https://doi.org/10.1007/978-3-642-41136-6_5

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*, 103119. https://doi.org/10.1016/j.trc.2021.103119

Tien, S.-L. A., Taylor, C., Vargo, E., & Wanke, C. (2018). Using ensemble weather forecasts for predicting airport arrival capacity. *Journal of Air Transportation, 26*(3), 123–132. https://doi.org/10.2514/1.D0105

Voskaki, A., Budd, T., & Mason, K. (2023). The impact of climate hazards to airport systems: A synthesis of the implications and risk mitigation trends. *Transport Reviews, 43*(4), 652–675. https://doi.org/10.1080/01441647.2022.2163319

Wang, Y. (2011). Prediction of weather impacted airport capacity using ensemble learning. In *IEEE/AIAA 30th Digital Avionics Systems Conference* (pp. 2–612611). https://doi.org/10.1109/DASC.2011.6096002

Wang, Y. (2012). Prediction of weather impacted airport capacity using RUC-2 forecast. In *IEEE/AIAA 31st Digital Avionics Systems Conference (DASC)* (pp. 3–313312). https://doi.org/10.1109/DASC.2012.6382312