

AI-Driven Air Traffic Flow Management for Weather and Capacity-Constrained Operations

An Integrative Approach for Next-Generation ATFCM tools

Andrada I. Gheorghe
EUROCONTROL Brussels,
Belgium
andrada-ioana.gheorghe.ext@
eurocontrol.int

Grigori Gustin Gustin
Software Consulting
LLC Iași, Romania
gr.gustin@gmail.com

Benjamin Cramet
EUROCONTROL
Brussels, Belgium
benjamin.cramet@eurocontrol.int

Eric Ligonnet
EUROCONTROL
Brussels, Belgium
eric.ligonnet@eurocontrol.int

Abstract—Due to an increase in convective weather events, the EUROCONTROL Network Manager (NM) introduced the Cross-Border Weather Procedure (CBWP) to facilitate coordination with Air Navigation Service Providers (ANSPs). The goal of the CBWP is to mitigate the impact of severe weather by addressing capacity constraints in the European airspace network.

ISOBAR serves as a support tool to optimise NM processes, facilitating decision-making and communication between NM Flow positions and ANSPs. This is achieved by integrating bleeding-edge Integrated Network Manager (iNM) operational tools like iFlow (the future Flow function front-end), and with the official weather forecast source from the European Meteorological Network (EMN).

ISOBAR is an open-loop Machine Learning (ML) tool designed for convective weather management, with two main purposes. First, it predicts the probability of a weather-induced regulation materialising at the Traffic Volume (TFV) level. Second, it predicts the associated regulation rate, also known as capacity reduction. This provides operators with early insights into potential problem areas, improving coordination with ANSPs. By providing predictions at both tactical (TACT) and pre-tactical (pre-TACT) phases of flight planning, ISOBAR helps operators to spot differences between capacity reduction and planned demand, and serves as a valuable input for simulations and implementations of possible scenarios.

Keywords—ATFCM, Convective Weather, Cross-Border Weather Procedure, Machine Learning, Traffic Volume Regulation, NMOC, EUROCONTROL

I. INTRODUCTION

The EUROCONTROL Network Manager (NM) is responsible for Demand-Capacity Balancing (DCB) in the European airspace, while ensuring safe and efficient network operations. In 2024, EUROCONTROL observed an increasing number of delays due to convective weather that totalled 5 million minutes, accounting for 36% of the total delays in the network for that year [1]. Convective weather refers to the formation of Cumulonimbus (Cb) clouds and is the leading cause of delays, especially in the summer months, resulting in sector overloads.

While there are multiple ways to manage Cb clouds, the most effective approach — preventing overloads and controlling capacity — is to establish regulations proactively. Regulations are implemented for affected airspace sectors, specifying a start time, end time, and regulation rate (the

adjusted capacity for the sector in question). There are many reasons to regulate airspace and consequently reduce their capacity, but in this project, the focus is on convective weather (Cb clouds) due to its strong operational impact on sector capacity and overall network performance.

To manage convective weather more effectively, the Cross-Border Weather Procedure (CBWP) was introduced in 2018. The goal of the CBWP is to support communication between Air Navigation Service Providers (ANSPs) and EUROCONTROL, and to increase transparency between stakeholders on how regulations are implemented. In 2025, the CBWP evolved into the Capacity and Weather-Based Operations (CWBO), now including capacity. Since EUROCONTROL does not make decisions about what specific measures to enforce to reduce the impact of convective weather, the idea was to improve transparency and especially the planning of such measures. To facilitate this, it has been decided to use a pan-European (cross-border) weather forecast as reference, aggregated by the European Meteorological Network (EMN) [2] from different local Meteorological (MET) offices.

Although regulations act as a safety net for sector overloads, they are set on the day of operations *after* Cb clouds have already formed. The business need and research goal then shape as follows: predicting regulations in advance by anticipating the impact of convective weather events *before* they materialise. This would help operational staff plan more effectively, reduce unforeseen network disruptions, and improve both efficiency and safety. For optimal benefit, the prediction horizon should ideally cover both tactical (TACT) and pre-tactical (pre-TACT) phases of flight planning, that is on the day of operations (D0) and one day prior to operations (D-1), respectively. Furthermore, considering the nature of regulations, the predictions should be made at the Traffic Volume (TFV) level.

ISOBAR, a weather-caused regulation predictor, has been developed to address this need and is also integrated into the CBWP, and now in the CWBO. ISOBAR is a Machine Learning (ML)-based tool designed for operational use in the Network Manager Operations Centre (NMOC) at EUROCONTROL, leveraging ML's ability to detect complex patterns in



large datasets that traditional statistical methods cannot easily capture.

This paper is structured as follows. Section II explores previous work done on the topic, identifying research gaps covered by this work. Section III then incorporates these gaps with the NMOC's business need, setting up a clear approach for the development of the tool. The methodology, including data sources, feature selection, and algorithm testing, is detailed in Section IV. Next, the results of the tool's performance are provided in Section V, followed by a discussion on its usability and data quality in Section VI. Findings are concluded in Section VII, along with directions for future work.

II. LITERATURE REVIEW

Before defining the project goal, in addition to the business need at EUROCONTROL, it is worth mentioning previous work done in the area. In general, approaches to the problem found in the literature can be categorised either by the location of the TFBs considered (data aspect) or by the modelling methodology applied (algorithmic approach).

EUROCONTROL has already conducted several projects that address similar questions at *airport* level [3]. While, at first sight, the broader problem may appear comparable, specific details introduce significant differences. For instance, ground-level regulations rely on different weather attributes such as precipitation level or wind speed, but on the other hand, they lack the added complexity of en-route airspace topology - a challenge that will be covered later in this paper.

The second group of research typically focused on specific subsets of TFBs, often analysing their configuration - such as common entry and exit points or airspace routes - and examining their interaction with the convective weather events. Some studies, particularly those focused on Terminal Areas, analysed deviations from planned flight routes using radar data [4], though usually for a limited set of areas with high interest. Another study even incorporated trajectory data to estimate regulated traffic entry counts for the construction of a two-dimensional grid, integrating forecasts from local MET services, in this case Météo-France [5]. Compared to the readily available EMN convection weather forecast (accessible within EUROCONTROL), the previously mentioned research made use of now-cast/observational data, which does not tackle the challenge faced in the CBWP, namely TACT and pre-TACT forecasting.

Finally, it should be noted that forecasting the impact of convective weather can often be reframed as a broader problem, namely that of estimating the theoretical capacity of the potentially affected airspace sector [6]. This perspective closely aligns with the DCB problem, which several studies have addressed [7], though without explicitly focusing on convective weather. However, this is beyond the scope of this work, as the primary concern is operational decisions based on the current state of the NM and the airspace topology it provides.

III. PROBLEM SETUP

Following an in-depth review of the state-of-the-art and NMOC's stated business requirements, the problem setup can be established. As discussed in Section II, part of the previous studies considered have explored the prediction of weather regulations either at airport or ACC level, narrowing down the training dataset to a specific region. While this approach yields reasonable results, it also tends to reflect recurring local events rather than capturing broader patterns. [5] addressed network impact through hotspot identification; however, this lies outside the scope of regulation prediction and does not provide the level of detail required for Air Traffic Flow and Capacity Management (ATFCM) in an operational context i.e. the information required for the NMOC staff. Finally, the last point refers to the prediction horizon of the built models. Although some have achieved a horizon of as short as 1-2 hours [5], none of the reviewed studies predict weather-caused regulations one day in advance, in a pre-TACT setting.

The research gap becomes clear: predicting weather regulations at TFB-level, across the European network, and with a pre-TACT prediction horizon - that is, one day prior to operations. Therefore, the goal of this paper is to showcase ISOBAR, a multi-horizon model for predicting solely weather-caused regulations at TFB-level for the NM area. Multi-horizon, in the sense that the model can be used for generating both predictions on D0 as well as D-1.

Above all, it is important to note that this project was initiated in response to an operational need identified within the NMOC. As operational stakeholders are increasingly confronted with weather-related regulations, and the latter has become one of the most frequent causes of ATFCM regulations, a prediction tool like this proves to be highly valuable. Nevertheless, the operational usage of this tool also comes with specific requirements and relies on active stakeholder partnership and feedback. These include adequate selection of model metrics for user interpretability, iterative development, and continuous validation to ensure alignment with evolving operational needs. ISOBAR is integrated in the existing CBWP, which aims to reduce weather-related ATFCM delays. The tool is deployed via the iFLOW platform, EUROCONTROL's new prototype system for the NMOC.

IV. METHODOLOGY

In this section, the methodology related to the development of the ISOBAR tool is presented. While this paper describes the most up-to-date model, note that multiple ISOBAR versions have been deployed yearly since summer 2023. At each iteration, an increase in precision was sought to make ISOBAR more reliable for the NMOC.

In this context, Subsection IV-A provides an overview of the data explored in this study, listing selected model features for the training dataset. Next, Subsection IV-B discusses data distribution, while Subsection IV-C describes the algorithms considered, those selected, and presents an overview of the model architecture.



A. Data Sources & Features

Before exploring ML algorithms, it is crucial to understand what data is available for training, including its distribution, quality, and possible impact on model performance. For ISOBAR, the following data sources were considered:

- EMN weather forecast
- Airspace-related data (needed for Acrux Python library)
- Regulations & capacity
- TFB activations & monitorability

Since EMN is used by the NMOC, specifically its pan-European (cross-border) weather forecast [2], it has also been selected as the main weather forecast source for this project. The data is provided by the German MET office (DWD), host of the EMN dataset. Specifically, forecasted weather events (Cb clouds) are represented as polygons defined by geographic coordinates, each linked to a validity window between 09:00 and 21:00 in 3-hour intervals.

Weather forecasts are issued twice daily, both on the day preceding the day of operations. The forecast issued at 07:00 is valid for D-1 while the one issued at 22:00 is valid for D0. The risk matrix (used by operational staff) associated with the EMN weather forecast is shown in Figure 1, where the colour coding (purple, red, orange, yellow) denotes the risk associated to each polygon.

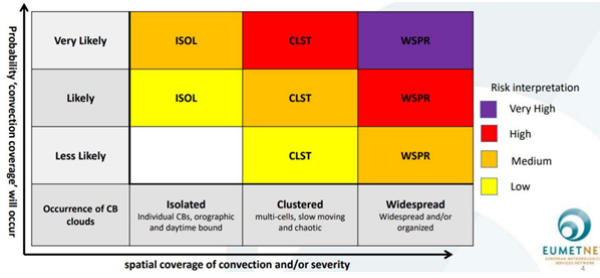


Figure 1. EMN unified risk matrix for forecasters and users.

An internally developed Python library, Acrux, is used to link EMN weather polygons with affected TFBs, a process referred to as clipping. The latter essentially identifies where both TFB and weather polygons overlap. Since Acrux requires knowledge of TFB boundaries to perform this task, airspace-related data (second data source) must also be included in the ISOBAR pipeline. Clipping allows the model to identify areas of potential regulations, shifting the focus to regions that may become problematic to operational staff.

The third data source contains regulations & capacity data and includes the target of the model, namely capacity reduction. Each TFB is associated with an expected hourly capacity value (standard). To align with the research objective, the *regulated* capacity from the regulations dataset should also be mapped to the corresponding hourly timeslots. The capacity reduction - the target variable in ISOBAR - is then defined as the difference between the *regulated* capacity and the *expected* capacity.

The resulting dataset contains both the actual real capacities (regulated or unregulated) and the derived capacity reductions - indicating a regulation when non-zero. Normalising this data into uniform hourly timeslots fixes the time dimension across all TFBs within the airac cycle, simplifying the model and metrics interpretability.

The last data source considered is linked to TFB properties essential to NMOC staff, namely activation and monitorability - features added to ISOBAR in 2025. Their definitions are taken from the ATFCM Operations Manual [8].

A TFB becomes active when its reference sector is activated according to the sector configuration. Once active, the TFB is available for the NMOC and Flow Management Positions (FMPs) to monitor. The difference between activations and monitorability is subtle; a monitorable TFB is conceptually flagged as one that should be observed if conditions are met, while an activated TFB is actually operational - its sector is active at a given time period, meaning monitoring can occur. Therefore, monitoring is only possible if the TFB is both marked as monitorable and currently activated.

Regulations can only be set on activated TFBs - in fact, NMOC staff note that “*a regulated TFB is activated in 90% of the cases*”. In situations where a regulation is imposed on a non-activated TFB, the regulation itself instantaneously triggers its activation. On the other hand, regulations may occasionally occur on non-monitorable TFBs for several operational reasons.

First, collapsed sectors often appear as non-monitorable in the system, yet some TFBs forming these sectors may require regulation. This is particularly relevant for important ACCs such as Maastricht, London, and Reims, which frequently operate with collapsed sector configurations.

Second, a regulation may be applied to a TFB without a subsequent system update of its monitorability status (human error). These cases should be accounted for, consequently, the ISOBAR training dataset includes non-monitorable TFBs. Moreover, as NMOC staff *only* work with activated TFBs, one should not train ISOBAR with irrelevant data (i.e. non-activated TFBs). This would introduce false-positive predictions, namely regulation predictions on non-activated TFBs - which cannot occur in real-life conditions.

As stated previously, these features have been added in 2025, attempting to better reflect operational reality and improve the reliability of ISOBAR’s predictions. While activations are used as a filter for the training dataset, activation *type* and monitorability are taken as features. Solely considering activated TFBs for training the model aims to improve prediction reliability, by reducing false-positive predictions that would occur on non-activated TFBs. The latter are operationally unfeasible, hence the introduction of these features.

Following extensive explanations on data sources, the features selected for training are listed in Table I, along with data type, unit used (when applicable), and the encoding method used. Note that for clipping, more elements than those

TABLE I. DESCRIPTION OF FEATURES USED FOR TRAINING.

Data Source	Feature	Description	Type	Units	Encoding
EMN forecast	cloud_top_height	Altitude of Cb cloud upper boundary	numerical	m	-
	valid_hour	Forecast time of validity (hour)	numerical	h	-
	has_CONVECTIONFRACTION_ISOL	Isolated convection	boolean	-	one-hot
	has_CONVECTIONFRACTION_CLST	Clustered convection	boolean	-	one-hot
	has_CONVECTIONFRACTION_WSPR	Widespread convection	boolean	-	one-hot
	has_LIKELINESS_likely	Convection occurrence likely	boolean	-	one-hot
	has_LIKELINESS_less likely	Convection occurrence less likely	boolean	-	one-hot
	has_LIKELINESS_very likely	Convection occurrence very likely	boolean	-	one-hot
	has_risk_very_high	Very high convective risk	boolean	-	one-hot
	has_risk_high	High convective risk	boolean	-	one-hot
	has_risk_medium	Medium convective risk	boolean	-	one-hot
	has_risk_low	Low convective risk	boolean	-	one-hot
	has_risk_very_low	Very low convective risk	boolean	-	one-hot
	is_amendment	Indicates amended forecast	boolean	-	one-hot
	is_tact_prediction	Indicates tactical prediction horizon	boolean	-	one-hot
Airspace data	tv_impacted_id	Impacted TFV (unique ID)	categorical	-	ordinal
Regulations & capacity	capacity	Actual capacity of TFV	numerical	ac/hr	-
	capacity_reduction	Difference between regulated and expected capacity	numerical	ac/hr	-
TFV activations & monitorability	is_monitorable	Indicates monitorability of TFV	boolean	-	one-hot
	is_activated	Indicates activation of TFV	boolean	-	one-hot
	is_tact_activated	Indicates activation type	boolean	-	one-hot

listed under *Airspace data* have been used, however they have been left out for the purpose of model training. This table only contains features and target (capacity reduction) of the ISOBAR model.

B. Data Exploration

Once the features have been selected, it is important to understand the distribution of the dataset that will be fed to the model, as well as deciding how to sample the data across training-testing. For this purpose, an Exploratory Data Analysis (EDA) has been conducted on the dataset of reference. As EMN forecasts only cover convective weather events, which occur exclusively during the meteorological summer, the dataset is also restricted to this period, ranging from the 2nd of May until the 16th of October. For this paper the focus is on one year of data (for the respective period), specifically 2024, split into training and testing sets with a 2/3-1/3 ratio, respectively. Previous versions of ISOBAR were trained on data from the year prior to deployment; for instance, if the model was intended for operational testing in summer 2023, the training dataset was taken from summer 2022.

Across the different data sources, several distributions are worth discussing. Figure 2 shows the daily count of *weather* regulations throughout summer 2024, along with the average for that period, which amounts to just over 68 regulations per day. There is a clear peak period in the months of June through August, proving how midsummer is heavily impacted by convective weather. On high-intensity days, these spikes reach almost four times the average.

Next, several distributions related to the TFV properties, namely activations and monitorability, are shown below. Figure 3 plots the size of the original dataset, before and after filtering out non-activated TFVs, with only 8.3% of the original dataset kept (overall). Zooming in on those impacted by

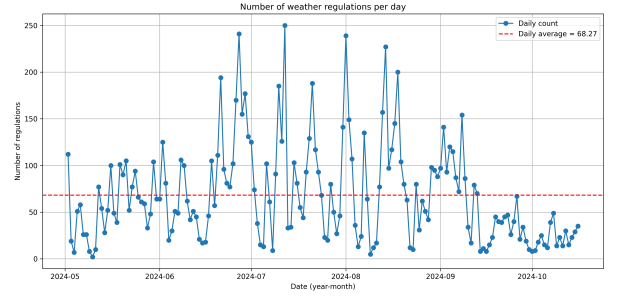


Figure 2. Weather regulations distribution in the dataset of reference.

weather, hence those for which the clipping process identified overlaps between TFV and weather polygons, only 5.2% of datapoints (rows) remain. This highlights the importance of activations filtering for significantly reducing false-positive predictions. One additional observation regarding the *filtered* dataset is the distribution of prediction horizons. Overall, 51.8% of activated TFVs are labelled for D0, while the remaining 48.2% are labelled for D-1, indicating a reasonable balance between the two. This also ensures that the model will not be biased toward better predicting one horizon over the other, as it is exposed to roughly equal amounts of data for both.

Figure 4 shows the number of rows in the dataset with associated regulations (general and weather-related) before and after activations filtering, in blue and red respectively. Two observations can be made. First, around 80% of “general” regulations are consistently impacted by weather, both before and after filtering. Second, the figure highlights the number of rows (with associated regulations) lost after filtering, which are further summarised as percentages in their respective

categories (i.e. general and weather-related). Overall, less than 10% of regulations are lost, confirming the NMOC staff’s approximation that “a regulated TFV is activated in 90% of the cases”.

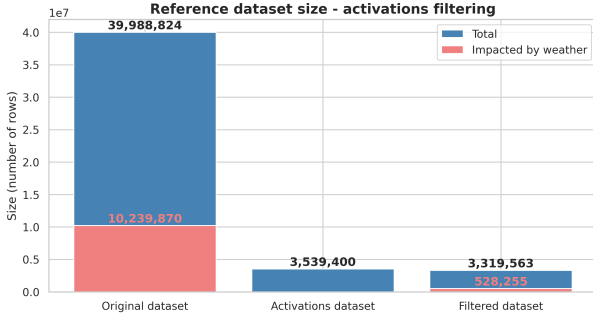


Figure 3. Size of reference dataset after activations filtering.

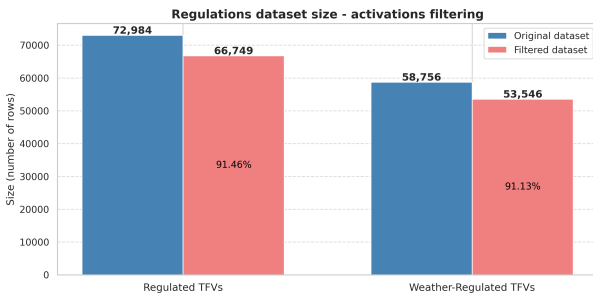


Figure 4. Regulations dataset size after activations filtering.

Finally, Figure 5 shows the number of regulated TFVs from a monitorability perspective. Slightly more than 70% of regulated TFVs are impacted by weather, meaning they have associated weather data resulting from clipping. Note that this percentage is only valid for monitorable TFVs and is calculated using regulated TFV *counts*. This is not the same as the 80% described in Figure 4, where the ratio is calculated based on *hourly rows* in the dataset. For non-monitorable TFVs, the percentages are distributed more evenly, with nearly 55% affected by weather. This highlights the importance of including non-monitorable TFVs for training, to avoid overlooking “non-conventional” regulations - i.e. those applied to TFVs whose monitorability status may not have been updated due to human error.

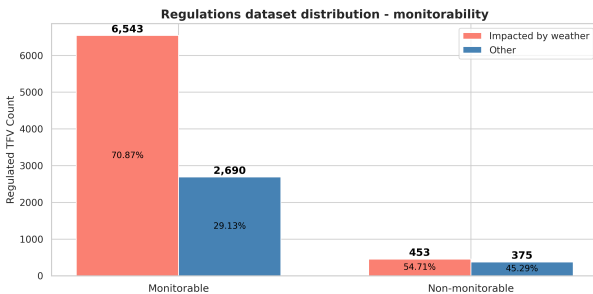


Figure 5. Regulated TFV count and associated monitorability.

C. Modelling

An overview of the model logic can be seen in Figure 6, depicting the different elements constituting the ISOBAR tool. The left side of this figure illustrates the data sources described in Subsection IV-A along with the clipping process. ISOBAR approaches the problem in two stages: first, deciding whether a regulation is necessary for a given TFV, and second, quantifying the capacity reduction in cases where a regulation is applied. This sequential structure was selected based on data characteristics, as most cases involved no regulation, and only a small portion required estimating its impact. Approaching the problem in this way ensures that the model follows the same reasoning process as the operational decision-making it is designed to support.

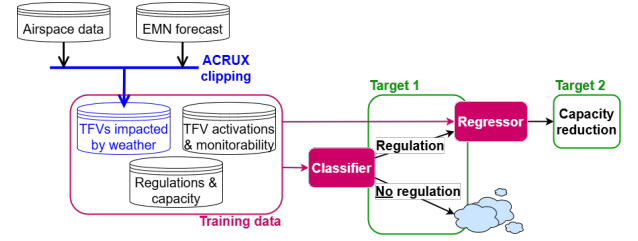


Figure 6. ISOBAR model structure.

Several algorithms have been selected for testing for both sub-models. For the classification task (deciding whether to impose a regulation), Logistic Regression (LR) and Support Vector Machines (SVM) were included as reference baselines, while ensemble methods such as Random Forest (RF), AdaBoost (ADA), and XGBoost (XGB) were selected for their ability to model complex and non-linear interactions and to handle unbalanced datasets, which is relevant given that regulation events are less frequent.

For the regression task (estimating the expected capacity reduction), linear models (LR, Ridge, Lasso) were considered to provide interpretable baselines, whereas RF and XGB were included to account for non-linear dependencies when estimating capacity reductions. This selection enables a balanced evaluation in terms of interpretability, computational cost, and predictive performance considering the operational context of ISOBAR.

The training of these algorithms was conducted as a preliminary evaluation to identify the best-performing models for subsequent hyperparameter tuning and full implementation. Regarding train-test data sampling, a time series cross-validation approach was used to evaluate model performance while preserving the temporal order of observations. The reference dataset was split into multiple sequential train-test folds, where each training set consisted of approximately the first two thirds of the available data up to a given point, and the corresponding test set comprised the subsequent one third - thus respecting the 2/3-1/3 train-test split mentioned in Subsection IV-B. Successive training sets were cumulative, expanding to include previous observations, while test sets always followed immediately after the training period. This

rolling strategy mimics real-world forecasting scenarios by ensuring that only past data informs predictions, preventing temporal leakage and providing a more robust assessment of model generalisation over time. The training results are summarised in Tables II and III for the classifier and regressor, respectively.

TABLE II. PERFORMANCE COMPARISON OF ML CLASSIFIERS.

Model	Accuracy	Precision	Recall	PR-AUC	F1	MCC	TT (s)
XGB	0.976	0.864	0.593	0.532	0.702	0.704	7.442
RF	0.974	0.832	0.566	0.492	0.673	0.673	9.922
ADA	0.972	0.881	0.467	0.437	0.609	0.629	3.635
SVM	0.972	0.912	0.450	0.437	0.602	0.629	0.326
LR	0.971	0.917	0.437	0.428	0.591	0.622	3.914

TABLE III. PERFORMANCE COMPARISON OF ML REGRESSORS.

Model	MAE	MAPE	RMSE	R2-score	TT (s)
XGB	0.471	0.687	1.624	0.518	6.304
RF	0.347	0.697	1.656	0.495	30.233
LR	0.695	0.873	1.938	0.315	0.213
Ridge	0.695	0.872	1.938	0.315	0.156
Lasso	0.854	0.919	2.340	-0.0006	3.660

For model metrics, Precision-Recall Area Under the Curve (PR-AUC) was selected for ranking classification algorithms in Table II, while Mean Absolute Percentage Error (MAPE) was chosen for the regression models in Table III. PR-AUC is suitable for evaluating the classifier with NMOC staff, as it provides a quantifiable measure of model performance. Since ISOBAR is an operational tool, stakeholder feedback must be considered iteratively to maintain constant model improvement, and PR-AUC allows adjustment according to desired precision levels. Similarly, MAPE was chosen for selecting the regression algorithm, mainly for its ease of communication because the error is easily interpretable as a percentage.

Looking at both tables, XGB is the best-performing algorithm, achieving the highest PR-AUC for the classifier and the lowest MAPE and RMSE for the regressor. For the classifier, XGB not only achieves the highest PR-AUC, but also shows superior F1-score, precision, and accuracy compared to the other models. Since PR-AUC is a key metric for communicating ISOBAR performance to NMOC staff, this advantage outweighs the slightly higher training time. For the regressor, although RF achieves a slightly lower MAE, XGB outperforms in terms of MAPE and RMSE, which are more relevant metrics for ISOBAR. Moreover, its training time is only 6 seconds compared to 30 seconds for RF, making it both more accurate and computationally efficient. None of the other algorithms offered a comparable balance between prediction performance and training efficiency. Besides predictive performance, XGB was also considered appropriate given the operational nature of ISOBAR. The algorithm is efficient and widely used, facilitating its integration into an operational environment. In addition, XGB can directly handle missing values, which simplifies pre-processing and ensures robust predictions even

when some input data is not available. Finally, applying the same algorithm for both classification and regression tasks simplifies the overall system. These considerations, together with the observed results, led to XGB being selected for the final ISOBAR implementation.

V. RESULTS

This section presents the most important results following the final model training. These cover model performance, feature importance, and validation with another dataset.

As mentioned previously, PR-AUC was the selected metric; hence to better visualise model behaviour, the precision-recall graphs are shown in Figures 7 and 8 for D0 and D-1 predictions, respectively.

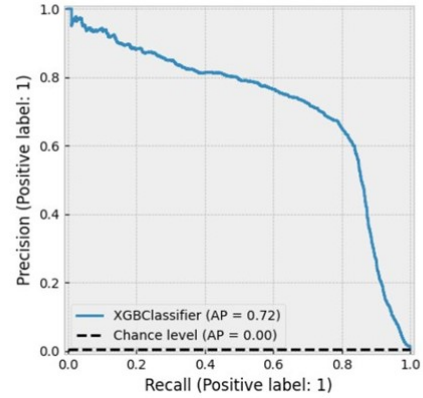


Figure 7. Precision-recall plot for D0 prediction horizon.

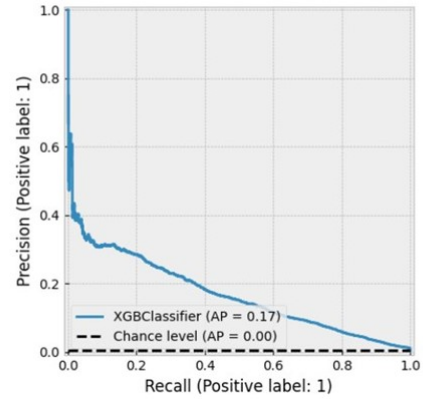


Figure 8. Precision-recall plot for D-1 prediction horizon.

The difference is clear and reflects the importance of input data availability at each prediction horizon, particularly related to activations. As was detailed in Subsection IV-A, activations are set by FMPs on D-1, that is, *during* D-1. Therefore, this data is only usable for data associated with D0 prediction horizons. Data related to D-1 predictions will have neither activations nor monitorability information, hence the resulting

precision-recall graph. When comparing both, it is clear from Figure 7, how much the performance is improved when activations (and monitorability) data are available for model input.

Overall, PR-AUC reaches 0.498, though performance varies across prediction horizons, declining from 0.733 on D0 to 0.167 on D-1. In TACT settings, one can reach a recall of up to 0.8 while maintaining a precision of circa 0.7. In pre-TACT settings, precision - already less than half of that on D0 - already starts declining at recall values above 0.2. This is attributed mainly to the lack of activation information on D-1 and to the higher uncertainty surrounding the pre-TACT phase, including that of weather forecasts. Because activation filtering cannot be applied at D-1, ISOBAR produces a substantially higher number of false-positive predictions, reflected by low precision values.

The raw confusion matrix is given in Figure 9, listing total counts of predictions per category. These total counts give more context to the overall PR-AUC of 0.498, proving that the model correctly classifies the majority of negative cases, yet a considerable portion of positive cases are *incorrectly* classified - leading to a substantial amount of false-positives. This pattern aligns with the precision-recall trade-offs and is influenced by the lower precision observed for D-1 predictions.

Overall, the model achieves relatively high recall for positive cases, but precision is limited by the number of false-positives - reflecting the fluctuating predictive performance observed across both prediction horizons. Nevertheless, moderate predictive performance is acceptable in the pre-TACT phase, which is primarily dedicated to planning, whereas high precision and PR-AUC for D0 predictions are particularly valuable in the TACT phase, where operational decisions require more accurate predictions. Despite the higher rate of false-positives in D-1 predictions, operational stakeholders indicated that they would rather have more potential regulations displayed - and thus, better visibility of areas that may become problematic - than risk omitting one. Since preparation is more important than precision at this stage, a lower precision is acceptable during D-1 operational planning.

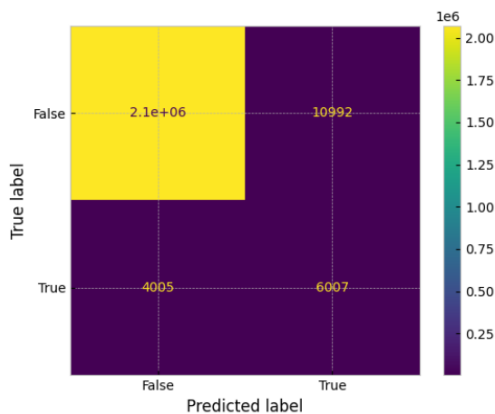


Figure 9. Raw confusion matrix (total counts).

Feature importance is summarised in Table IV, ranking top five features most used by the model for the classifier and regressor components. These are ranked using F-score, which is the number of times a feature was used to split data across all trees in the XGB algorithms. A higher F-score indicates higher feature usage, and thus greater influence on the model.

The analysis shows that `tv_impacted_id` and `capacity` are dominant features for both sub-models, with substantially higher F-scores for the classifier. MET-sourced features such as `cloud_top_height` and `valid_hour` also contribute strongly, capturing both spatial and temporal aspects of forecasts. In contrast, the regressor highlights the importance of the `is_activated` Boolean feature, though with a weaker influence compared with the top four predictors - which are the same for both submodels.

TABLE IV. FEATURE IMPORTANCE ANALYSIS.

Sub-model	Feature	Weight (F-score)
Classifier	<code>tv_impacted_id</code>	7825.0
	<code>capacity</code>	5930.0
	<code>cloud_top_height</code>	3305.0
	<code>valid_hour</code>	1488.0
	<code>has_CONVECTIONFRACTION_CLST</code>	563.0
Regressor	<code>tv_impacted_id</code>	1528.0
	<code>capacity</code>	877.0
	<code>cloud_top_height</code>	869.0
	<code>valid_hour</code>	489.0
	<code>is_activated</code>	215.0

To complement feature importance, SHapley Additive ex-Planations (SHAP) are provided below for the classifier only, since the latter governs the first and most critical decision point i.e. whether regulations occur. This focus ensures interpretability of the most critical model component while avoiding redundancy with the feature importance.

A sample of 100 regulated slots were extracted randomly to reduce computational effort. SHAP are summarised in Figure 10 with features ordered by their overall contribution to the sub-model; the top five predictors appear at the top of the plot.

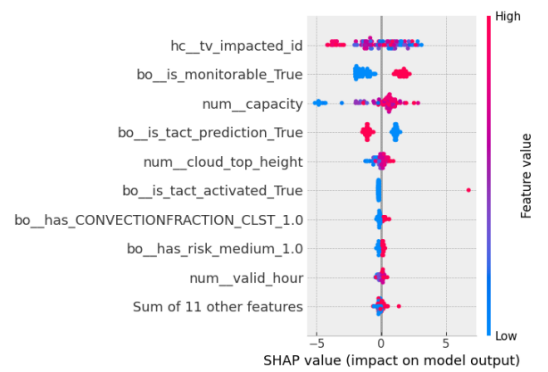


Figure 10. SHAP Analysis for classifier.

Three of them, namely `tv_impacted_id`, `capacity`, and `cloud_top_height`, are consistent with the F-score ranking from Table IV. However, SHAP further reveals the

strong contribution of operational features (*is_monitorable* and *is_tact_prediction*), which were not among the most frequently used splits by the model. In contrast, other features like *valid_hour* and *has_CONVECTIONFRACTION_CLST* are ranked highly by F-score but appear less influential in the SHAP analysis. This reflects differences between the two methods; F-score captures the frequency of feature use in tree splits, while SHAP measures the direct contribution of features to prediction outcomes.

Finally, the model was validated on 2023 (summer) data to ensure its applicability across a differently-distributed dataset, with different convective weather events. While features remain the same, there has been a slight change in EMN data format between 2023 and 2024, particularly concerning amended forecasts. Instead of making use of the amendment flag, an additional dataset is generated on D0, called “updated forecast” (D0U) which contains all previous D0 forecasts, and if needed, respective changes. For this reason, the amount of predictions on D0 is double in 2023 compared to 2024. Nevertheless, duplicates were removed and the analysis was conducted successfully. Results for key metrics are displayed in Table V.

TABLE V. PR-AUC RESULTS FOR THE ORIGINAL MODEL (TRAINED ON SUMMER 2024) AND VALIDATION (TESTED ON SUMMER 2023).

Metric	Original	Validation
PR-AUC (D0)	0.733	0.744
PR-AUC (D-1)	0.167	0.122

Validation results show a slightly higher PR-AUC value for D0 predictions (0.744) compared to the original model (0.733). This likely reflects that D0 regulations are not only predicted more accurately (due to the availability of activations) but they also have a tendency to occur on the same TFBVs, making them more consistent across years and easier for the model to generalise.

In contrast, D-1 predictions have a lower PR-AUC on the validation dataset (0.122 versus 0.167). This is attributed to the inherent uncertain conditions surrounding the pre-TACT phase and potential differences in operational procedures, leading to discrepancies in data distributions between summer 2023 and 2024. Overall, these results validate strong predictive performance for D0 while highlighting the expected limitations for D-1.

VI. DISCUSSION & FUTURE WORK

Alongside the features incorporated into the final version of ISOBAR deployed in iFlow, another research track was pursued. This investigation revealed further opportunities for future work and highlighted areas worth exploring.

A first important finding is related to the demand data, namely the entry counts of aircraft in each TFBV. Although intuitively one would expect demand to play a significant role in regulation prediction, its integration yielded only limited improvements. This outcome was primarily attributed to data

quality issues present in demand and capacity datasets (declared and nominal capacities). The observed improvement in PR-AUC was 0.13, suggesting that data context and quality are more influential than the simple inclusion of demand-related features.

A second point stems from the fact that, often, in such dynamic systems where multiple actors optimise for both shared and own goals (ATFCM delay reduction, safety, route optimisation, penalties, etc.), the data is constantly and dynamically updated - in this case, particularly the activation status and monitorability of TFBVs. It was observed that some TFBVs are activated only after a regulation is applied, and in certain cases, regulations were enforced on non-monitorable TFBVs, highlighting inconsistencies in operational data. The main operational consequence is the omission of regulations due to the short-reported change in the activation status of certain TFBVs. Nevertheless, this happens in exceptional situations, generally in the case of unforeseen convective weather events, and hence it is also not captured by the weather forecast features. A key mitigation measure is to train ISOBAR users to keep in mind that the tool can sometimes be wrong and may omit certain regulations, although rarely. The final decision for regulation setting should always lie with the operational staff; ISOBAR is only meant as a decision-support tool.

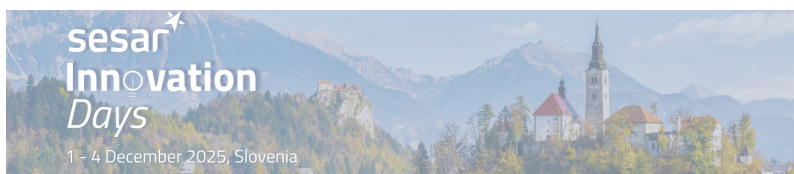
Lastly, issues with regulations data quality were identified. Both anecdotal evidence and EDA show that the *filed* regulation reason does not always correspond to the true reason. For example, regulations attributed to weather were, in some cases, actually the result of staffing constraints or other operational factors.

VII. CONCLUSIONS

This study addressed the subject of weather-caused ATFCM regulations and their impact on the European Network, exploring ML solutions tailored to NMOC staff at EUROCONTROL. To this end, ISOBAR, a tree-based ML model, was developed to predict exclusively weather-caused regulations at TFBV-level, together with their associated capacity reductions. As convective weather (Cb clouds) occurs mainly in summer, the tool was trained on *NMOC-standard* summer data (2nd of May until 16th of October) provided by EMN for the respective years. The latest model version integrates TFBV activation filtering and an extended prediction horizon, enabling predictions for both D0 and D-1, with an overall PR-AUC of 0.498.

While this marks a significant improvement over previous versions, ISOBAR still has limitations. Considerable variations in PR-AUC are observed, with 0.733 for TACT predictions and 0.167 for pre-TACT. This highlights the critical role of input data availability, specifically TFBV activations, which are only available at the end of D-1 and thus usable only for D0 - explaining the decline in PR-AUC across prediction horizons. Nevertheless, the model was validated on 2023 data and demonstrates consistent performance.

Future work will focus on improving the pre-TACT predictions. Within this framework, possible research areas include incorporating an additional layer to predict TFBV activations



or clustering TFVs to enable broader predictions at a coarser level. The latter would be acceptable in the pre-TACT phase, given its planning-oriented purpose.

ACKNOWLEDGMENT

The authors would like to thank NMOC staff for their valuable operational insights and feedback, the iFLOW team for their collaboration on ISOBAR deployment, and Eric Ligonnet for the Acrux Python library enabling clipping. Finally, we acknowledge the teamwork that made this project possible and thank all contributors.

REFERENCES

- [1] Performance Review Commission of EUROCONTROL, “Performance review report of the european air traffic management system in 2024,” EUROCONTROL, Tech. Rep., 2025.
- [2] C. Dubois, “Eumetnet supports eurocontrol’s network manager with onsite meteorologists, aiding the summer cross border convection advisory,” <https://www.eumetnet.eu/eumetnet-supports-eurocontrols-network-manager-with-onsite-meteorologists-aiding-the-summer-cross-border-convection-advisory/> Available online, 2025.
- [3] O. Lattrez, R. B. Montes, and M. Michalski, “Predicting airport atfm regulations using deep convolutional neural networks,” unpublished.
- [4] S. Wang, B. Yang, R. Duan, and J. Li, “Predicting the airspace capacity of terminal area under convective weather using machine learning,” *Aerospace*, vol. 10, p. 288, 2023.
- [5] I. Martínez, J. García-Heras, A. Jardines, A. Cervantes, and M. Soler, “Predicting air traffic flow management hotspots due to weather using convolutional neural networks,” *Engineering Applications of Artificial Intelligence*, vol. 133, Jul 2024.
- [6] K. et al., “Capacity estimation for airspaces with convective weather constraints,” unpublished or unspecified source.
- [7] S. Mas-Pujol, E. Salami, and E. Pastor, “Rnn-cnn hybrid model to predict c-atc capacity regulations for en-route traffic,” *Aerospace*, vol. 9, no. 2, p. 93, Feb 2022.
- [8] *ATFCM Operations Manual*, Maint-2 ed., EUROCONTROL, Network Management Directorate, Jun. 2024, edition approved on 17 May 2024.