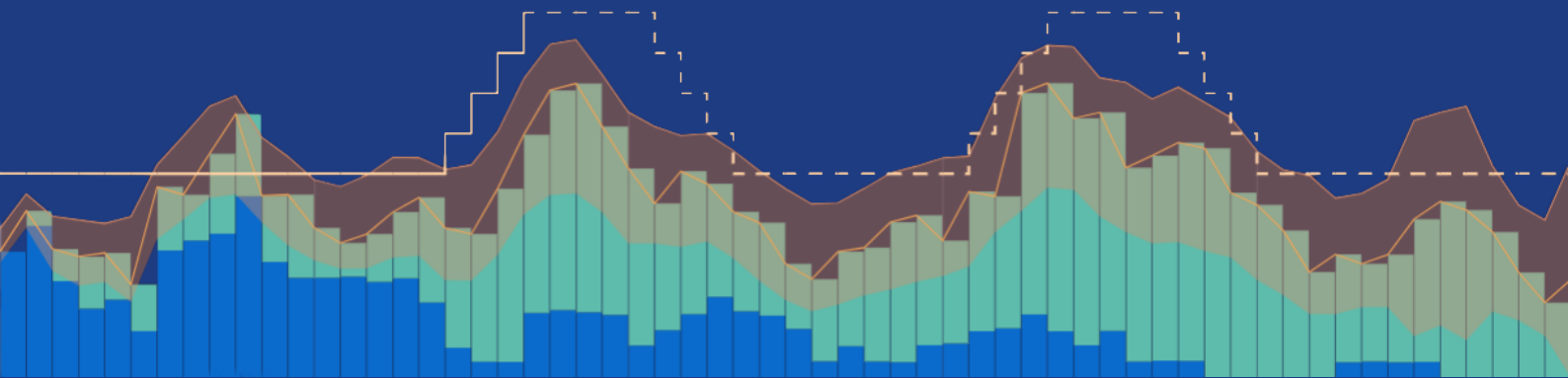


Probabilistic Forecasting of Inbound Demand Using Conformal Prediction

A Case Study at LVNL

J.P.H.W. Simons



Probabilistic Forecasting of Inbound Demand Using Conformal Prediction

A Case Study at LVNL

Thesis report

by

J.P.H.W. Simons

to obtain the degree of Master of Science
at Delft University of Technology
to be defended publicly on 8 July 2026 at 14:00

Thesis committee:

Chair:	Ir. O. Stroosma
Supervisors:	Dr.ir. J. Ellerbroek Dhr. F. Dijkstra Ir. B. van Dillen
External examiner:	Dr. A. Amiri-Simkooei
Place:	Faculty of Aerospace Engineering, Delft
Project Duration:	November 2025 - July 2026
Student number:	5056462

Cover image by Hessel Meulenbeld (LVNL) and Isis Prummel
An electronic version of this thesis is available at <https://repository.tudelft.nl/>.

Preface

This thesis represents the final milestone in fulfilling the requirements for my Master of Science degree in Aerospace Engineering at Delft University of Technology. It marks the end of seven memorable years as a student in Delft. The research presented in this document was conducted in collaboration with Luchtverkeersleiding Nederland (LVNL) and supported by the Knowledge & Development Centre Mainport Schiphol (KDC) through the Centre of Excellence.

I would like to express my gratitude to Joost Ellerbroek, Ferdinand Dijkstra, and Bob van Dillen for their supervision. Their guidance and encouragement were crucial to this thesis, and the freedom they gave me to explore my own ideas made it an engaging endeavour. Furthermore, this research would not have been possible without the support of several LVNL employees, who contributed through their expertise and access to operational data. Finally, I would like to thank my peers at the KDC and iLabs, as well as iLabs CEO Matthijs, for their practical advice and insightful coffee-break chats throughout the process.

Contents

Nomenclature	iv
List of Figures	v
List of Tables	vi
1 Introduction	1
I Scientific Article	2
II Appendices to the Scientific Article	20
A PFI Results	21
B Calibration Stratification	25
B.1 Effect of Daytime vs. Nighttime Operations on Calibration	25
B.2 Effect of Prediction Horizon on Calibration	26
C ACI Framework	28
C.1 Effect of Step Size on ACI Performance	28
C.2 Effect of Information Lag on ACI Performance	32
D Synthetic Inbound Demand	35
D.1 Difference Between Offline and Online Synthetic Inbound Demand	35
D.2 Bunch Forming Due to LVNL's ATFM Regulations	36
E Effect of Target Error Rate on Prediction Interval Width	38
F Performance of DST vs. Median-Quantile ML Model	39
III Literature Review and Research Proposal	40
2 Literature Review	41
2.1 Background Information	41
2.2 Previous Work on Aggregate Inbound Demand Predictions	43
2.3 Uncertainty Quantification with a Wrapper Model	44
2.4 Time Series Analysis and Forecasting	45
2.5 Conclusion	49
3 Research Objective	50
3.1 Research Objective	50
3.2 Research Question	50
3.3 Scope	50
4 Project Plan	52
4.1 Methodology	52
4.2 Expected Results	53
4.3 Project Schedule	53
5 Conclusion	56
References	60

Nomenclature

ACC	Area Control Centre
ACI	Adaptive Conformal Inference
ANSP	Air Navigation Service Provider
ATFM	Air Traffic Flow Management
CBA	Cross-Border Area
CQR	Conformalized Quantile Regression
CTOT	Calculated Take-Off Time
DCB	demand and capacity balancing
DST	Decision Support Tool
EFD	ETFMS Flight Data
EHAA	Amsterdam FIR
EHAM	Amsterdam Airport Schiphol
ETA	estimated time of arrival
ETFMS	Enhanced Tactical Flow Management System
FIR	Flight Information Region
FMPC	Flow Management Position Controller
FNR	false negative rate
FPR	false positive rate
GBDT	Gradient-Boosted Decision Tree
ICAO	International Civil Aviation Organization
KDC	Knowledge & Development Centre Mainport Schiphol
KNMI	Koninklijk Nederlands Meteorologisch Instituut
LVNL	Luchtverkeersleiding Nederland
LVP	Low Visibility Procedure
MAD	mean absolute deviation
ML	machine learning
MWIS	Mean Winkler Interval Score
NM	Network Manager
PFI	Permutation Feature Importance
PI	prediction interval
RVR	Runway Visual Range
SKV	Schiphol Kansverwachting

List of Figures

B.1	Overlapping conformity score distributions for both daytime (0400 - 2159 UTC) and nighttime (2200 - 0359 UTC) operations for the 150-minute and 300-minute prediction horizons. . . .	26
B.2	Conformity score distributions for daytime and nighttime operations across 60-, 180-, and 300-minute prediction horizons.	26
B.3	Comparison of the expansion factor ($Q_{1-\alpha}$) by prediction horizon for daytime (0400 - 2159 UTC) and nighttime (2200 - 0359 UTC) operations resulting from the Conformalized Quantile Regression (CQR) framework.	27
C.1	2016-step (1 week) rolling mean coverage level of various models for a 60-minute prediction horizon.	28
C.2	2016-step (1 week) rolling mean coverage level of various models for 180-minute prediction horizon.	29
C.3	Coverage level for various step sizes γ in the Adaptive Conformal Inference (ACI) model. .	29
C.4	mean absolute deviation (MAD) in percentage points (p.p.) from the target coverage level for various step sizes γ in the Adaptive Conformal Inference (ACI) model.	29
C.5	Step size optimised for mean absolute deviation (MAD) and Mean Winkler Interval Score (MWIS) (2016-step rolling mean), evaluated on the validation dataset using grid search. . .	30
C.6	2016-step (1 week) rolling mean coverage level of CQR and ACI model and a Bernoulli sequence for a 150-minute prediction horizon evaluated on the test dataset.	31
C.7	Step size optimised for mean absolute deviation (MAD) for two different rolling window sizes, evaluated on the validation dataset using grid search.	32
C.8	Coverage level for the CQR, operational ACI ($\gamma = 0.0008$), and ideal ($\gamma = 0.0008$) ACI models evaluated on the test dataset.	33
C.9	MAD in percentage points (p.p.) for the CQR, operational ACI, and ideal ACI models evaluated on the test dataset.	33
C.10	Coverage level of various models ($\gamma = 0.0008$) for 300-minute prediction horizon with a 2016-step rolling average (1 week) evaluated on the test dataset.	33
D.1	Distribution of differences between synthetic inbound demand used for training (offline) and operational use (online).	36
D.2	Distribution of absolute differences between prediction interval widths after applying the ACI framework using the online and offline synthetic inbound demand definitions.	36
D.3	Example of bunch forming due to ATFM regulations imposed by LVNL on 2025-10-05 with a 300-minute prediction horizon.	37
E.1	Effect of target error rate on calibrated prediction interval width evaluated on the test dataset.	38
F.1	Mean and median absolute inbound demand prediction error of the DST and the median (50 th percentile) quantile regression model, plotted as a function of prediction horizon. . . .	39
2.1	Graphical explanation of demand and capacity balancing (DCB) (EUROCONTROL, 2019).	42
2.2	Example of EHAM inbound demand prediction by the DST (Vegter et al., 2023).	42
2.3	Illustration showing the effect of intervals on inbound demand calculation (Dekker, 2023). .	43
2.4	ACI applied to the 2020 US presidential election for county-level predictions (Gibbs & Candès, 2021).	48
2.5	Electric power load demand for the Indian state of Tamil Nadu (Suresh et al., 2025).	49
4.1	Gantt chart, last updated on 2026-06-19.	55

List of Tables

A.1	Explanation of the features examined in the probabilistic model, including the PFI results evaluated on the validation dataset.	21
A.2	Low Visibility Procedure (LVP) categories (Wijngaard et al., 2008). VIS is the abbreviation for visibility, and RVR for Runway Visual Range.	24
C.1	Performance of various models using a 2016-step rolling mean evaluated on the test dataset. MAD in percentage points (p.p.).	30

Introduction

In recent years, demand for air travel has risen steadily, despite periodic downturns caused by global crises, and is expected to continue growing in the coming decades. European Union Aviation Safety Agency et al. (2025) predicts an annual growth of 1.1% in flight volume between 2025 and 2050 in its baseline scenario. However, the “European air traffic network is saturated” (EUROCONTROL, 2025a), making it necessary to continuously balance demand against the available airspace capacity within EUROCONTROL’s member states.

As the designated Air Navigation Service Provider (ANSP), Luchtverkeersleiding Nederland (LVNL) is responsible for the safe and efficient handling of all civil air traffic in the Amsterdam Flight Information Region (FIR) EHAA up to flight level 245. Amsterdam Airport Schiphol is the primary hub in this FIR. With over 470,000 flights and 66.8 million passengers in 2024, it is the largest airport in the Netherlands and one of the busiest in Europe (Royal Schiphol Group, 2025). Given these volumes, LVNL bears significant responsibility for ensuring the safety and efficiency of aircraft movements within Dutch airspace. Occasionally, it is necessary to hold inbound aircraft at their departure airport to prevent airspace capacity from being exceeded. This causes passenger delays and increases costs for airlines, and should therefore be kept to a minimum. The challenge is to allow as much air traffic as possible to enter the Amsterdam FIR without exceeding capacity limits. This process is known as demand and capacity balancing.

To support this balancing process, LVNL uses the Decision Support Tool (DST). This software tool calculates and visualises inbound demand, allowing LVNL to apply precautionary regulations when necessary to lower demand peaks. However, the current DST generates deterministic forecasts and therefore does not quantify prediction uncertainty. A shift to a probabilistic framework would better capture the inherent uncertainty of the problem and enable an Air Navigation Service Provider (ANSP) such as LVNL to make more informed decisions. This research aims to develop such a probabilistic forecasting method for aircraft inbound demand to enhance the efficiency of Air Traffic Flow Management (ATFM), specifically targeting reductions in unnecessary ATFM delays and unforeseen capacity breaches.

The research was conducted in cooperation with the Knowledge & Development Centre Mainport Schiphol (KDC), a partnership between LVNL, KLM, Amsterdam Airport Schiphol, and the Dutch Ministry of Infrastructure and Water Management. Through the Centre of Excellence, the KDC aims to foster collaboration with academia. This research was carried out at LVNL, which provided access to operational data and the expertise of its employees.

This thesis is structured as follows. First, Part I presents the scientific paper, while Part II contains its appendices. The literature review and research proposal, which were fundamental to the development of the scientific paper, are presented in Part III. More specifically, Chapter 2 presents a literature study, evaluating state-of-the-art methodologies and identifying gaps for further research. Subsequently, Chapter 3 outlines the research objective and corresponding research question that this thesis aims to answer. Chapter 4 details the methodology, expected results, and planned project schedule, followed by a brief conclusion to the research proposal in Chapter 5.

Part I

Scientific Article

Probabilistic Forecasting of Inbound Demand Using Conformal Prediction

J.P.H.W. Simons

Under the supervision of Dr.ir. J. Ellerbroek¹, Dhr. F. Dijkstra², and Ir. B. van Dillen¹

*Section Control & Simulation, Faculty of Aerospace Engineering
Delft University of Technology, Delft, The Netherlands*

¹*Faculty of Aerospace Engineering, Delft University of Technology*

²*Knowledge and Development Centre Mainport Schiphol*

Abstract—Air navigation service providers currently rely on deterministic demand forecasts for Air Traffic Flow Management (ATFM), which inherently fail to quantify forecast uncertainty. This paper presents a top-down probabilistic forecasting framework for inbound air traffic demand using Air Traffic Control the Netherlands (LVNL) as a case study. The framework models the uncertainty of an existing deterministic forecasting system using quantile regression, thereby capturing heteroscedasticity directly at the aggregate demand level. To improve empirical coverage under non-exchangeable conditions, Conformalized Quantile Regression is combined with Adaptive Conformal Inference. The framework was applied using approximately two years of operational ATFM data. Compared to a statistical error-margin baseline, the proposed machine learning approach produced substantially narrower prediction intervals and achieved empirical coverage closer to the target, leading to a 25.1% lower Mean Winkler Interval Score (MWIS). Furthermore, Adaptive Conformal Inference more closely tracked the nominal 90% target coverage level over time than static conformal calibration, while also slightly reducing average interval width and MWIS. A retrospective operational analysis indicates that the probabilistic framework enables more flexible risk-based decision-making compared to deterministic forecasting alone. For a representative 150-minute prediction horizon, specific probabilistic decision thresholds simultaneously reduced both the false positive and false negative rates of capacity breaches relative to the deterministic baseline across the evaluated test dataset. These results demonstrate the potential of adaptive conformal prediction techniques to support probabilistic inbound demand forecasting within ATFM operations.

1. Introduction

The European airspace is facing critical capacity challenges, with the network effectively “saturated” according to EUROCONTROL [1]. To ensure that air traffic demand does not exceed available capacity, local Air Navigation Service Providers (ANSPs) linked to EUROCONTROL can request Air Traffic Flow Management (ATFM) regulations from EUROCONTROL’s Network Manager. These regulations manage network congestion by assigning ground delays to specific aircraft before they depart. By proactively holding these

flights, the anticipated demand peak is flattened, preventing airspace overloads.

To accurately determine when to request these ATFM regulations, ANSPs currently rely on deterministic demand forecasts. For example, Luchtverkeersleiding Nederland (LVNL) uses a software tool called the Decision Support Tool (DST), which provides a single expected value. While these deterministic forecasts support timely decision-making, they do not quantify the prediction uncertainty. As quantified error margins are not available, ANSPs must rely on operational judgement regarding forecast uncertainty during the process of demand and capacity balancing (DCB). If ANSPs are risk-averse, pre-emptive regulations could lead to unnecessary ground delays when the actual demand naturally falls below the forecast. Conversely, accepting excessive risk could lead to in-flight delays if the actual demand turns out to exceed the maximum capacity. This paper presents a case study on LVNL’s inbound demand predictions, where the existing deterministic demand forecast is complemented with probabilistic information. Quantifying the uncertainty inherent in the demand predictions enables ANSPs to transition toward risk-based DCB. A risk-based approach could enhance the demand and capacity balancing process, as ANSPs can take into account the probability of an airspace overload, ultimately minimising unnecessary ATFM delays while simultaneously reducing the number of unexpected sector overloads.

In pursuit of improving inbound demand estimates and quantifying this uncertainty, significant research has already addressed aircraft arrival time estimates, trajectory predictions, and their variance [2–6]. Furthermore, predicting the uncertainty of inbound demand through a bottom-up approach has been explored [7–10]. However, top-down probabilistic frameworks have received comparatively limited attention in the literature [11, 12]. To the author’s knowledge, conformal prediction has not yet been applied to top-down probabilistic inbound demand forecasting in ATFM. Therefore, the main scientific contribution of this paper lies in the evaluation of a top-down probabilistic forecasting framework for aircraft inbound demand using adaptive conformal prediction. The proposed framework aims to obtain strong prediction interval coverage behaviour while minimising interval width under changing operational conditions. Although demonstrated using LVNL’s operational environment, the proposed frame-

work is designed to be generalisable, potentially making it transferable to inbound demand forecasting at other ANSPs.

The remainder of this paper is structured as follows. Section II discusses the necessary background information, while Section III details the proposed methodology. The experimental setup is described in Section IV. Section V presents the results, which are subsequently discussed in Section VI. Finally, concluding remarks are provided in Section VII.

II. Background Information

This section introduces the operational and methodological concepts required for probabilistic inbound demand forecasting. First, it describes the ATFM context in which LVNL uses deterministic inbound demand forecasts. It subsequently defines inbound demand as used in this study and compares the proposed top-down probabilistic approach to bottom-up uncertainty aggregation methods. Finally, it introduces residual error modelling, quantile regression, and conformal prediction as the methodological basis for the proposed approach.

A. Operational Context

LVNL is responsible for the Amsterdam Flight Information Region (FIR) EHAA below flight level 245. The traffic volume within this airspace is predominantly composed of aircraft heading to and from Amsterdam Airport Schiphol (EHAM), with traffic to and from regional airports constituting only a minor fraction. This research concentrates on inbound demand. Consequently, it focuses exclusively on air traffic inbound to EHAM, consistent with the approach used by LVNL's DST when predicting inbound demand for the Area Control Centre. The inbound demand capacity of EHAA is constrained by the runway capacity at EHAM. To ensure that the capacity is not exceeded, ground delays are preferred over in-flight delays due to the higher financial and environmental impact of the latter. The goal of the predictions made by LVNL's DST is to provide the necessary insight to balance airspace demand and capacity, through a process called demand and capacity balancing (DCB). This entails lowering demand peaks that exceed the capacity limit by filling demand troughs when surplus capacity exists, as illustrated by Fig. 1.

If inbound demand is projected to exceed capacity, Flow Management Position Controllers (FMPCs) from ANSPs connected to EUROCONTROL, such as LVNL, can request ATFM regulations from EUROCONTROL's Network Manager. These regulations enforce capacity constraints by imposing a Calculated Take-Off Time (CTOT) on departing inbound flights, requiring aircraft to depart within a tolerance window ranging from 5 minutes before to 10 minutes after the assigned time. By assigning these ground delays, ANSPs can reduce peaks of inbound traffic demand.

Probabilistic demand information can potentially enhance the demand and capacity balancing process by quantifying the uncertainty surrounding deterministic predictions, as illustrated in Fig. 2. In this scenario, the deterministic prediction exceeds the hypothetical capacity limit around 13:15 UTC. Relying solely on this point estimate, the FMPC could decide to impose ATFM regulations. However, the uncertainty mar-

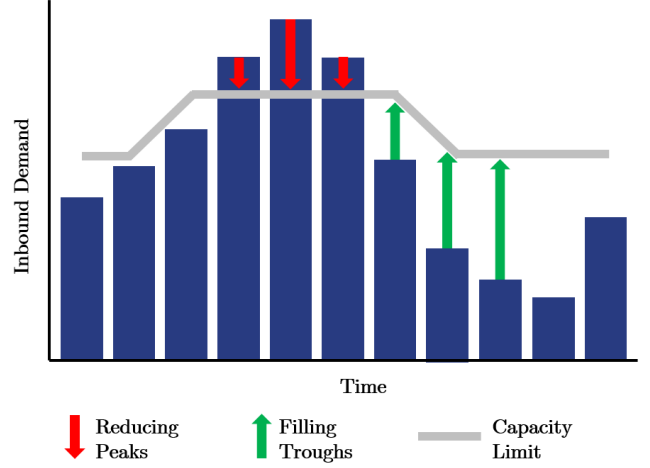


Fig. 1 Visualisation of demand and capacity balancing (DCB).

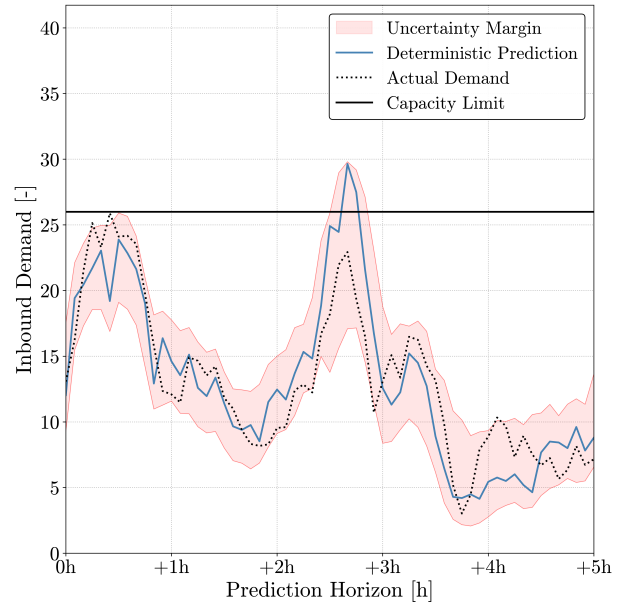


Fig. 2 Probabilistic forecasting envelope for 13-09-2025 10:30 UTC. The capacity limit is hypothetical. The actual demand is only available in hindsight.

gin reveals that the deterministic estimate lies near the upper bound of the predicted uncertainty interval, and that there is a high probability that the actual demand will remain safely below the capacity limit. This outcome is confirmed by the actual demand curve. Thus, with the probabilistic insight provided by the uncertainty margin, the FMPC could choose not to implement the ATFM regulations, preventing unnecessary ground delays. Conversely, probabilistic information can also prevent missed capacity breaches in a different scenario (not visible in Fig. 2). If a deterministic prediction falls slightly below the capacity limit, the FMPC might normally withhold regulations. However, if the uncertainty interval indicates a heavily upward-skewed uncertainty, the risk of a breach becomes visible, potentially triggering requests for ATFM regulations that would otherwise have been overlooked. These are two concrete examples of how providing probabilistic metrics around the deterministic prediction can enhance the

demand and capacity balancing process, potentially reducing unnecessary ground delays and unforeseen capacity breaches.

B. Inbound Demand Definition

LVNL defines inbound demand as the number of intervals between inbound aircraft passing the EHAA Cross-Border Area (CBA) within a 20-minute time window. A visualisation of the EHAA CBA is shown in Fig. 3 [13]. In contrast to a simple count of aircraft crossing the boundary, this interval definition is based on the amount of intervals between consecutive inbound aircraft within a time window. Intervals that span the start or end of a time window are fractionally split, contributing to each window in proportion to the time spent within it (illustrated in Fig. 4). This interval-based definition of demand more realistically represents the pressure of inbound traffic on an airspace sector than a simple count of aircraft arrivals.

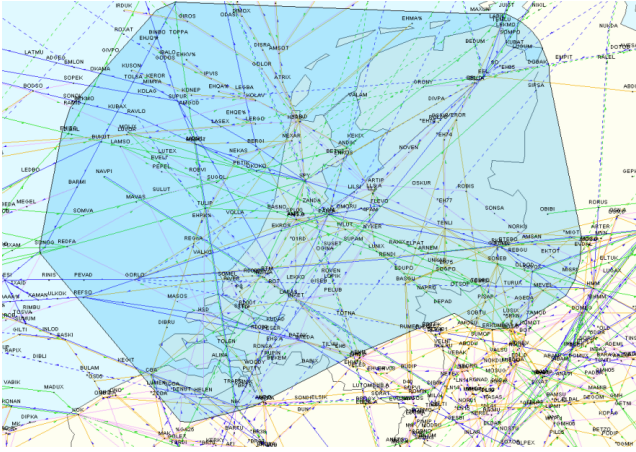


Fig. 3 Visualisation of the EHAA CBA [13].

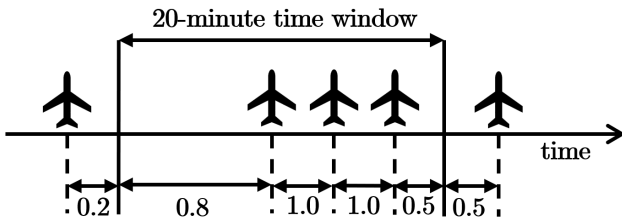


Fig. 4 Illustration of the inbound demand calculation with fractional splits for intervals overlapping multiple time windows. Each aircraft symbol on the timeline represents the estimated EHAA CBA entry time of an inbound flight. For the indicated 20-minute time window, the demand equals 3.3.

C. Bottom-up vs. Top-Down Approach

Having detailed the operational context and inbound demand definition, the next step is to compare the proposed forecasting framework with existing probabilistic demand forecasting methods.

The predictability of air traffic demand is fundamentally driven by the uncertainty inherent in individual flight arrival times. Ruskamp [14] notes that “traditionally, demand forecasting in Air Traffic Management (ATM) has relied on trajectory-based methods”. It is therefore unsurprising that

much of the literature has adopted bottom-up approaches to probabilistic inbound demand forecasting, where aggregate demand uncertainty is calculated by summing the uncertainty of individual flight arrival times. Extensive research has focused on refining individual trajectory predictions, arrival time estimates, and their uncertainty [2–6], as well as on quantifying demand uncertainty by aggregating these individual flight uncertainties [7–10]. However, using this bottom-up approach to inbound demand uncertainty quantification at major airports like EHAM presents critical limitations. Not only does modelling the aggregate uncertainty of individual flights require the simulation of dozens or even hundreds of flights simultaneously, but the methodology is also highly vulnerable to error propagation. Any systematic bias or modelling inaccuracy at the individual flight level will compound when aggregated, distorting the probabilistic forecast of inbound demand.

To circumvent the compounding complexities of individual flight uncertainty aggregation, a top-down probabilistic approach views aggregate demand as a singular entity. Top-down probabilistic forecasting remains largely unexplored for air traffic demand. Wanke et al. [11, 12] took a top-down approach to demand forecasting, but omitted time windows with high demand to circumvent the self-defeating prophecy of these predictions. In this phenomenon, high demand predictions exceeding capacity limits trigger ATFM regulations that alter actual traffic, leaving the original unconstrained demand unobservable (further elaborated in Section III.A). Wanke et al. [11] identified three explanatory variables that correlate with prediction uncertainty: prediction horizon, demand, and the share of proposed and active flights.

While disregarding individual arrival time uncertainties for a metric fundamentally driven by them may appear counter-intuitive, modelling demand at the aggregate level may benefit from partial error cancellation. Micro-level deviations, such as individual aircraft shifting into adjacent time windows, may partially offset one another at the aggregate level. Thus, although individual flight arrival times might deviate significantly, the overall impact on the demand count can be marginal. Furthermore, treating the demand as a collective entity can capture large-scale phenomena driving inbound demand uncertainty, such as weather patterns, airspace capacities, and temporal patterns in the demand prediction when appropriate explanatory variables are included. Lastly, modelling the demand at the aggregate level significantly reduces computational cost compared to a bottom-up approach, which is an important factor for an operational model that continuously needs to perform probabilistic forecasting.

D. Uncertainty Quantification

LVNL already employs the DST, which generates deterministic forecasts of inbound demand. The objective is not to replace these predictions, but to quantify the uncertainty through secondary residual error modelling. The underlying notion is that the residual errors of the deterministic model are a measure of the forecast uncertainty. This technique of secondary residual error modelling, as described by Brando et al. [15], involves creating a “wrapper model” that predicts

the uncertainty of a black-box algorithm. O’Sullivan et al. [16] took a similar approach to estimate the uncertainty of bus arrival times, for example. This work was inspired by the paper by Pereira et al. [17], who note that such a methodology is transferable to other regression tasks provided there is access to the raw predictions, corresponding observations, and variables correlating with the prediction error. Because the framework models these empirical residuals rather than the physical arrival process, it treats the primary forecaster as a black box, making it model-agnostic. Thus, while the internal logic of LVNL’s DST is known and will be used to retroactively simulate missing historical prediction data as described in Section IV.B, the proposed framework does not require such structural insight to function.

Gradient-Boosted Decision Tree (GBDT) models are well suited for this probabilistic approach because they have potential benefits over traditional statistical methods and deep learning architectures. Recent studies, such as the results from the M5 forecasting competition, indicate that decision-tree-based machine learning techniques outperform statistical and hybrid methods for highly correlated time series presented in tabular data [18–20]. Additionally, some GBDTs can natively capture heteroscedasticity through quantile regression. Heteroscedasticity, a phenomenon of fluctuating variance in a time series, is a critical feature for inbound demand forecasting, as prediction errors naturally scale with variables such as traffic volume and prediction horizon [11].

E. Conformal Prediction

To construct prediction intervals, quantile regression is commonly used, as it allows for the direct estimation of conditional quantiles. This capability is essential for capturing the heteroscedastic nature of inbound demand uncertainty. However, standard quantile regression does not provide statistical guarantees for finite-sample coverage. To achieve this, conformal prediction can be applied. Pioneered by Vovk et al. [21], this framework provides marginal coverage guarantees under the assumption of data exchangeability. Exchangeability assumes that the order of a dataset is irrelevant, and that it can be shuffled without affecting the joint probability distribution. A critical limitation is that the time series of inbound demand prediction errors is expected to violate the assumption of exchangeability due to serial correlation. Physically, a delayed aircraft missing its scheduled arrival window (leading to an over-prediction) must inevitably land in a subsequent one (leading to an under-prediction), creating strong temporal dependence in the prediction errors. Thus, the data cannot be randomly shuffled, as that would disrupt the temporal dependency of the time series. This violates the condition of exchangeability, causing standard conformal prediction techniques to lose their marginal coverage guarantees. Furthermore, changes in prediction uncertainty that were unobserved during model training (i.e., distribution shifts) also violate the exchangeability assumption.

To address this exchangeability issue, adaptive conformal prediction techniques can be implemented as an additional algorithmic layer. Operating in the online phase, an adaptive approach dynamically updates the target error rate based on

the observed errors of the previous predictions. This feedback loop allows the model to dynamically adjust the prediction intervals to distribution shifts. Consequently, the adaptive approach can improve empirical coverage in the long run [22]. Over sufficiently long time horizons, the empirical coverage converges to the nominal target coverage level.

Adaptive conformal prediction techniques have been applied to a wide variety of time series. For example, in the seminal paper presenting Adaptive Conformal Inference (ACI), Gibbs and Candès [22] applied it in combination with Conformalized Quantile Regression (CQR) to predict stock market volatility and vote totals during the 2020 US presidential election, showing that the adaptive approach is less prone to drift and has smaller deviations around the nominal target coverage level compared to static conformal prediction methods.

Numerous types of (adaptive) conformal prediction techniques exist. To the best of the author’s knowledge, they have not yet been applied to inbound demand forecasting in the field of ATFM. Therefore, the primary contribution of this research lies in the evaluation of a probabilistic framework for inbound demand predictions using a conformal prediction approach within the field of ATFM, rather than in the development of novel (adaptive) conformal prediction algorithms or the optimisation of the underlying architectures. This rationale also applies to the selection of the GBDT algorithm used.

III. Methodology

To quantify the uncertainty inherent in deterministic inbound demand forecasts, this research employs a top-down probabilistic framework that models the distribution of the residual errors of the DST using exogenous variables. The primary goal is to generate prediction intervals that reliably achieve the target coverage level while remaining as narrow as possible.

In theory, a perfect probabilistic model would fully capture the heteroscedastic nature of demand prediction uncertainty, making dynamic adjustment unnecessary. In practice, however, air traffic operations are subject to changing conditions that can cause the uncertainty patterns observed during training to change (violating the exchangeability assumption). These distribution shifts cause the prediction intervals to become too narrow or too wide for the actual prediction uncertainty, leading to under- or overcoverage, and reducing the statistical validity of the static prediction intervals. Therefore, the proposed framework includes an online adjustment mechanism that continuously evaluates its predictive performance. When the observed demand repeatedly falls outside the predicted intervals, indicating that uncertainty has increased beyond what the model captured, future intervals are widened. Conversely, when uncertainty is lower than expected, the intervals are contracted, reflecting the reduced prediction uncertainty. In this way, the framework adaptively adjusts interval width over time, allowing the probabilistic forecasts to remain reliable and informative under changing operational conditions.

To implement this adaptive behaviour, the methodology follows four sequential stages, as illustrated in Fig. 5:

- 1) **Create synthetic inbound demand:** Construction of

a synthetic unconstrained demand target to circumvent the self-defeating nature of high inbound demand caused by ATFM regulations (Section III.A).

- 2) **Train machine learning (ML) model:** Training a ML model via quantile regression to capture heteroscedastic prediction uncertainty (Section III.B).
- 3) **Create and calibrate prediction intervals (PIs):** Creating inbound demand prediction intervals and calibrating these initial prediction intervals using Conformalized Quantile Regression (CQR) (Section III.C).
- 4) **Adaptively adjust PIs in online phase:** Applying Adaptive Conformal Inference (ACI) to continuously adjust the intervals in the online phase, ensuring robustness against distribution shifts (Section III.D).

Finally, this section concludes by detailing how these dynamically adjusted intervals are used to calculate the probability of exceeding capacity limits (Section III.E).

The detailed explanation of these steps is provided below.

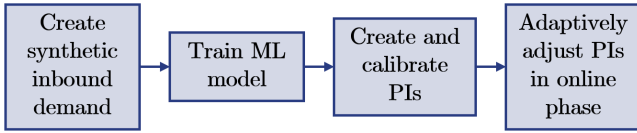


Fig. 5 Overview of the complete process performed to obtain adaptive PIs.

A. Construction of a Synthetic Unconstrained Demand Target

For this methodology, only the time series of predictions, observed values, and variables correlating with the prediction uncertainty are required, as explained in Section II.D. The first two are compared with each other to calculate the prediction error, which serves as the dependent variable for the quantile regression model. A fundamental challenge in obtaining observed target values when training a secondary residual error model for ATFM is the self-defeating prophecy of the inbound demand prediction. If the DST predicts that inbound demand will exceed capacity, LVNL's FMPC can request ATFM regulations through EUROCONTROL's Network Manager. Because these interventions alter the arrival times of inbound aircraft and thus inbound demand, the original unconstrained demand is never observed, creating a scenario where the predicted demand affects the observed demand [11, 12].

Comparing predicted demand directly against regulated observed demand would thus train the model incorrectly. To overcome this, a synthetic true inbound demand is created using synthetic arrival times for each flight. This synthetic baseline acts as a retrospective target approximating what the DST should have predicted in unconstrained conditions (i.e., if LVNL does not apply any future ATFM regulations from the prediction time onwards). It is calculated by assuming that, for flights receiving restrictive ATFM regulations imposed by LVNL that have not yet departed, the synthetic arrival time is the estimated take-off time of the most recent flight message plus the true flight time. Once a flight has departed, ATFM regulations from LVNL can no longer impact the arrival time through ground delays, and therefore the synthetic arrival time used equals the true arrival time. For flights

subject to overlapping regulations from different sources, this adjustment is applied when the restrictive delay from the most recent message originates from LVNL. If the restrictive delay originates from another source, the synthetic arrival time is the estimated take-off time of the last flight update without LVNL restrictive ATFM delay plus the true flight time. Flights that never receive LVNL restrictive ATFM delay use the true arrival time for the synthetic demand calculation. Cancelled flights are never included in the synthetic true inbound demand, as they constitute natural uncertainty in the inbound demand that the probabilistic model should capture. By creating a synthetic true inbound demand, it is possible to calculate the prediction error based on what the DST should have predicted at each moment in time. The flight time is defined as the time from take-off to EHAA CBA entry.

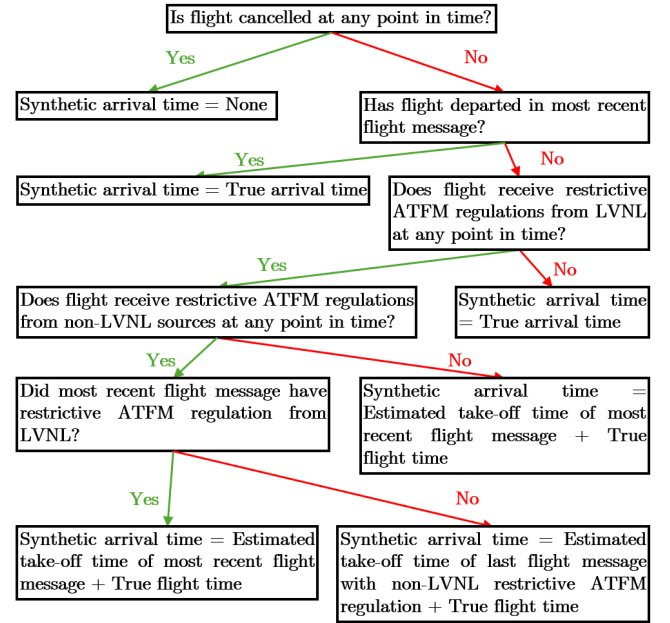


Fig. 6 Decision tree outlining the logic used to derive the synthetic arrival times for individual flights. These are aggregated to calculate the synthetic demand.

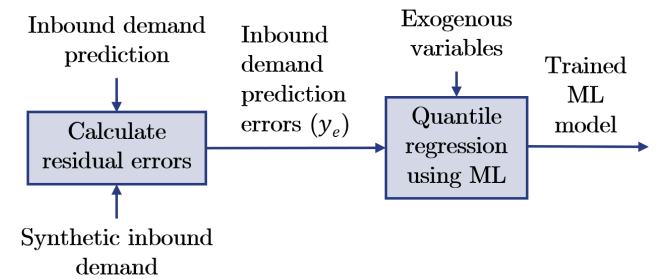


Fig. 7 Flowchart of the training process of an ML model using quantile regression.

B. Generation of Prediction Intervals

After the synthetic demand has been constructed, the residuals can be computed by comparing the synthetic demand with the deterministic predictions. These inbound demand prediction errors, combined with the explanatory variables, are then used to train a machine learning (ML) model that

predicts the desired quantiles of the inbound demand prediction errors, as illustrated in Fig. 7. The ML algorithm models the relationship between inbound demand prediction errors (dependent variable) and relevant explanatory features (independent variables) such as traffic volume and prediction horizon. By learning how these features relate to the errors, the model captures the heteroscedasticity of the DST prediction uncertainty.

To estimate the quantiles of the error distribution, the model employs quantile regression. By definition, a quantile specifies the threshold below which a given proportion of observations is expected to fall (e.g., 95% of the observations should lie below the 95th percentile). To achieve these specific estimates, the ML model is trained to minimise the pinball loss function $\rho_\tau(y_e, \hat{y}_e)$, as defined by Eq. (1) [23].

$$\rho_\tau(y_e, \hat{y}_e) := \begin{cases} \tau(y_e - \hat{y}_e) & \text{if } y_e - \hat{y}_e > 0, \\ (1 - \tau)(\hat{y}_e - y_e) & \text{otherwise} \end{cases} \quad (1)$$

Here, $\tau \in (0, 1)$ indicates the desired quantile level, y_e represents the DST prediction error, and $\hat{y}_e$ denotes the predicted quantile of the inbound demand prediction error. Minimising this asymmetric loss trains the ML model to provide conditional quantiles for the demand prediction error given the feature vector X_t , at time step t .

The specific ML algorithm employed in this study is LightGBM, an efficient Gradient-Boosted Decision Tree (GBDT) algorithm. Recent literature demonstrates that LightGBM achieves strong forecasting accuracy, particularly when processing correlated time series with numerous explanatory features [18]. Besides its computational efficiency relative to other gradient boosted algorithms [18], LightGBM natively supports quantile regression through pinball loss optimisation. Additionally, it has an inherent ability to process categorical features, which circumvents the need for one-hot encoding and thus prevents a large increase in dimensionality when categorical features are applied.

C. Calibration of Prediction Intervals

The calibration step is depicted in Fig. 8. First, the trained prediction error model from Section III.B is applied to a held-out calibration dataset. Using the same set of exogenous variables as input, the trained model can predict the lower and upper quantiles of the inbound demand prediction error.

These bounds are then superimposed onto the deterministic inbound demand prediction to construct an uncalibrated prediction interval (PI) for the demand instead of the demand

error. Mathematically, the inbound demand prediction interval is defined by Eq. (2) [16], where $\hat{q}_{\frac{\alpha}{2}}(X_t)$ and $\hat{q}_{1-\frac{\alpha}{2}}(X_t)$ denote the lower and upper bounds of the prediction interval, respectively, corresponding to a miscoverage rate α .

$$PI = [\hat{q}_{\frac{\alpha}{2}}(X_t), \hat{q}_{1-\frac{\alpha}{2}}(X_t)] \quad (2)$$

The coverage definition is presented by Eq. (3) [16].

$$P\left(\hat{q}_{\frac{\alpha}{2}}(X_t) \leq y_t \leq \hat{q}_{1-\frac{\alpha}{2}}(X_t)\right) = 1 - \alpha \quad (3)$$

For example, constructing a 90% prediction interval (i.e., $\alpha = 0.1$) requires training two quantile models, one for $\tau = 0.05$ and one for $\tau = 0.95$. The resulting interval represents the range within which an inbound demand observation y_t should fall with a probability $1 - \alpha$. To differentiate the inbound demand error (used in Section III.B) from the inbound demand at time step t , the former is denoted by y_e and the latter by y_t .

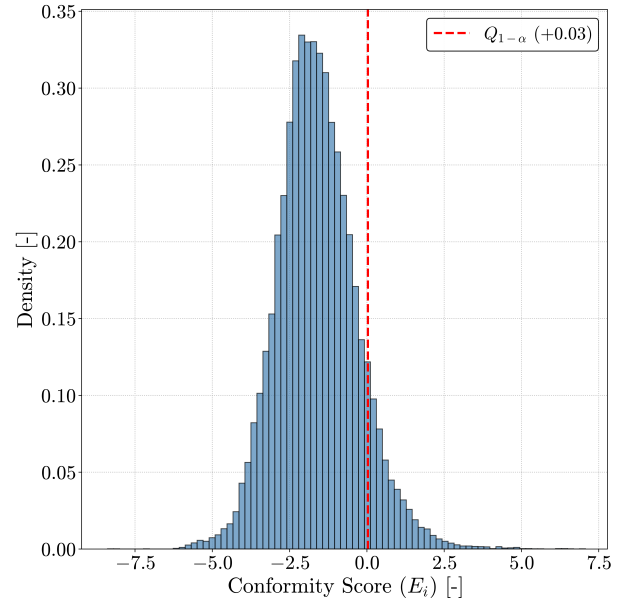


Fig. 9 Conformity score distribution for the 150-minute prediction horizon for daytime operations.

Once the uncalibrated inbound demand prediction interval has been generated, the intervals are refined using Conformalized Quantile Regression (CQR). This is a conformal prediction technique that provides finite-sample marginal coverage guarantees, assuming the time series are exchangeable. For each prediction in the calibration dataset, a conformity score E_t is computed by comparing the synthetic true demand with the upper and lower bounds, according to Eq. (4) [24].

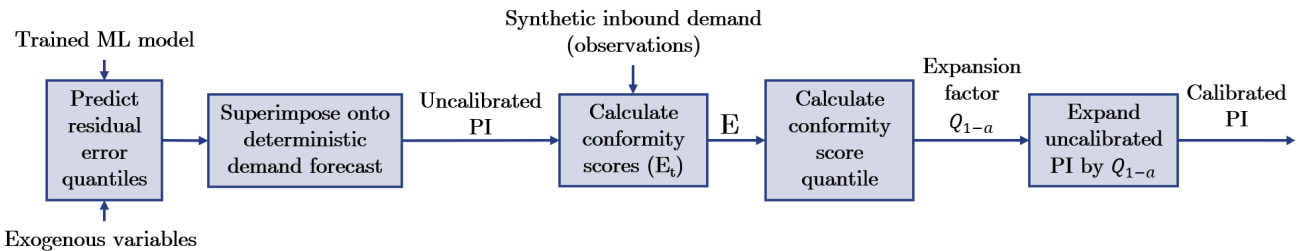


Fig. 8 Flowchart of the calibration process of the prediction interval (PI) using CQR.

A positive score indicates that the observation lies outside the initial, uncalibrated prediction interval, whereas a negative score implies it falls within the bounds. The set of all conformity scores across the calibration dataset is denoted by E .

$$E_t := \max \left\{ \hat{q}_{\frac{\alpha}{2}}(X_t) - y_t, y_t - \hat{q}_{1-\frac{\alpha}{2}}(X_t) \right\} \quad (4)$$

The conformity scores are then ordered, and the empirical quantile corresponding to $\frac{n+1}{n}(1-\alpha)$ (denoted by $Q_{1-\alpha}(E)$) is identified. The fraction $\frac{n+1}{n}$ is used as a finite-sample correction, where n stands for the number of scores in the set E . The empirical quantile acts as a calibration offset, used to expand or contract the uncalibrated intervals. This final expansion/contraction step is shown mathematically by Eq. (5) [24], with $C(X_{t+1})$ denoting the calibrated prediction interval.

$$C(X_{t+1}) = \left[\hat{q}_{\frac{\alpha}{2}}(X_{t+1}) - Q_{1-\alpha}(E), \right. \\ \left. \hat{q}_{1-\frac{\alpha}{2}}(X_{t+1}) + Q_{1-\alpha}(E) \right] \quad (5)$$

To improve calibration, a separate $Q_{1-\alpha}$ value is computed for each prediction horizon and time of day category of the prediction window. This distinction is made because prediction errors (i.e., prediction uncertainty) typically increase with prediction horizon and are significantly higher during the day than during the night. Daytime operations are defined as time windows starting between 0400 and 2159 UTC, and nighttime operations from 2200 to 0359 UTC. This stratification allows the calibration offset to incorporate differences in prediction uncertainty across operating conditions.

Fig. 9 illustrates the distribution of conformity scores and the corresponding $Q_{1-\alpha}$ value for the 150-minute prediction horizon, daytime operations scenario.

In essence, the CQR algorithm evaluates the trained model on unseen calibration data to determine the precise adjustment required to achieve the target coverage level. This adjustment is subsequently applied symmetrically to the lower and upper bounds of the prediction intervals made in the online phase.

D. Dynamic Adjustment of Prediction Intervals

CQR relies on the assumption of exchangeability to provide marginal coverage guarantees. However, inbound demand is expected to violate the exchangeability condition, as discussed in Section II.E. To improve empirical coverage under non-exchangeable conditions, Adaptive Conformal Inference (ACI) is employed. As a starting point, ACI uses the trained model and the expansion factor obtained from the CQR calibration

step ($Q_{1-\alpha}$), and then dynamically adjusts the intervals in an online fashion, as summarised in Fig. 10. Importantly, the adaptive calibration mechanism does not alter the underlying predictive model itself, but dynamically adjusts the calibration procedure in response to changing empirical coverage behaviour.

In the online phase, ACI repeatedly updates a time-varying target error rate α_t based on recent predictive performance, with the aim of converging to the nominal miscoverage level α over time. By continuously adjusting α_t , the algorithm shifts the empirical quantile extracted from the conformity scores ($\frac{n+1}{n}(1-\alpha_t)$), making the expansion factor $Q_{1-\alpha_t}$ a time-varying value and thus effectively shifting the red line in Fig. 9. The target error rate α_{t+1} is clipped to (0, 1) after each update step.

The update rule for the target error rate is given by Eq. (6) [22].

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \text{err}_t) \quad (6)$$

In this equation, α is the nominal target error rate, γ is the step size, and α_t and α_{t+1} are the target error rates for the current and subsequent steps, respectively. The discrete error indicator err_t is defined by Eq. (7) [22].

$$\text{err}_t := \begin{cases} 1 & \text{if } y_t \notin C(X_t), \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Tuning γ involves balancing reactivity against stability. A large γ makes the adaptive calibration highly responsive but risks becoming erratic, whereas a low γ leads to more stable prediction intervals but reacts more slowly to distribution shifts.

Each prediction horizon has its own α_t sequence. Additionally, when calculating the dynamic expansion factor $Q_{1-\alpha_t}$, the corresponding prediction horizon and time of day are taken into account, similar to the CQR setup. This ensures that both the static calibration and the subsequent adaptive adjustments respect the prediction error characteristics.

Furthermore, to realistically reflect operational conditions, the temporal delay in the ACI feedback loop is taken into account. For instance, when a prediction is made 3 hours in advance, the corresponding true inbound demand is only observed 3 hours and 20 minutes later (considering the 3-hour prediction horizon plus the 20-minute time-window duration). As a consequence, the longer the forecasting horizon, the greater the temporal lag before the algorithm can observe the realised error and update the subsequent target coverage

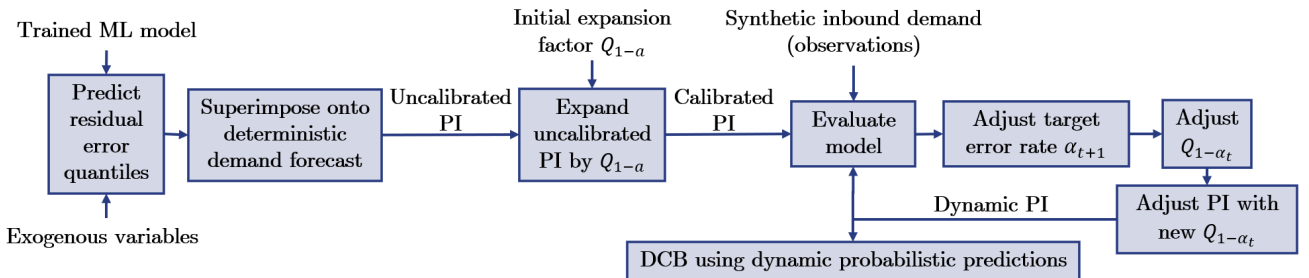


Fig. 10 Flowchart schematically presenting the dynamic adjustment of the prediction interval (PI) through ACI applied in the online phase. The initial calibration factor used to expand the uncalibrated PI comes from the CQR step.

levels, which in turn delays the adaptation of the corresponding prediction intervals.

E. Exceedance Probability

By interpolating between a set of quantiles for the inbound demand, an approximation of the cumulative distribution function of the residual errors can be obtained. The continuous approximation is based on the quantiles 0.01, 0.05, 0.10 to 0.90 in steps of 0.10, 0.95, and 0.99. This representation of the conditional distribution enables the computation of the probability that inbound demand exceeds a given capacity limit. For example, if the maximum allowable capacity coincides with the 75th percentile of the predicted demand distribution at a particular time step, the probability of an airspace overload at that time is 25%. Since the absolute theoretical minimum (0%) and maximum (100%) quantiles cannot be explicitly computed, the exceedance probability is set to 100% or 0% if the capacity threshold is below the lowest or above the highest modelled quantiles (i.e., 0.01 and 0.99). After training, the predicted quantiles are calibrated and subsequently sorted to prevent quantile crossing. The sorting step preventing quantile crossing is crucial to ensure that the cumulative distribution function is monotonically increasing.

Applying the methodology described in this section augments the deterministic inbound demand forecasts produced by LVNL's DST with a probabilistic uncertainty layer. Furthermore, it enables the estimation of the probability that demand will breach capacity limits.

IV. Experimental Setup

This section presents the experimental setup designed to apply the proposed probabilistic framework and evaluate its statistical and operational performance. It outlines the hypotheses, dataset split, explanatory features, model configuration and optimisation, and validation metrics employed.

A. Hypotheses

The aim of the research is to develop a top-down probabilistic forecasting framework for inbound demand that minimises interval width while maintaining reliable empirical coverage. To evaluate the mathematical and operational performance of this framework, three hypotheses are formulated.

The following hypothesis is proposed to test the framework's ability to model varying uncertainty and minimise interval width:

Hypothesis 1: The proposed top-down quantile regression framework captures heteroscedastic residual uncertainty more effectively than a statistical error-margin baseline, as measured by the MWIS.

The statistical error-margin baseline is constructed by calculating empirical residual quantiles from the training data and applying these as margins around the deterministic DST forecast. Unlike the ML framework, this baseline does not condition the interval width on explanatory variables, except for the prediction horizon.

The necessity of an adaptive algorithm applied to statically calibrated prediction intervals within the operational context of ATFM is tested using the following hypothesis:

Hypothesis 2: The ACI framework will maintain rolling empirical coverage closer to the nominal target than the static CQR approach, while achieving a comparable or smaller average interval width.

Finally, the following hypothesis evaluates the potential operational contribution of the model within LVNL's decision-making process concerning ATFM regulations:

Hypothesis 3: Transitioning from deterministic point forecasts to a probabilistic framework enables a measurable reduction in pre-emptive, unnecessary ground delay interventions while maintaining or reducing the rate of missed capacity exceedances.

B. Dataset

The primary dataset evaluated in this study comprises ETFMS Flight Data (EFD) of flights inbound to Amsterdam Airport Schiphol, consisting of 753 days of data. ETFMS stands for Enhanced Tactical Flow Management System. The data originate from EUROCONTROL's Network Manager and are provided by LVNL. To ensure data integrity, flights lacking explicit termination or cancellation messages are omitted from the dataset. This filtering process results in the removal of 2,172 flights from a total of 518,446, corresponding to 0.42% of the initial dataset. Of these removed flights, 1,097 can be attributed to having an arrival time falling outside the temporal scope of the dataset. The remaining exclusions, which account for just 0.21% of the initial dataset, are due to flight diversions and/or data logging errors. Because these two causes cannot be reliably distinguished, all of these flagged flights are removed from the dataset.

As established in Section II.E, the inbound demand time series exhibits strong serial correlation. Furthermore, the ACI algorithm requires a strictly chronological, continuous online phase to continuously update the target error rate α_t . Therefore, the dataset is split into the following uninterrupted, chronological parts:

- Training dataset (372 days, 49%): Spans from 2022-12-25 to 2023-12-31, used to train the base LightGBM model.
- Calibration dataset (156 days, 21%): Spans from 2024-04-24 to 2024-09-26, reserved for computing conformity scores through the CQR method.
- Validation dataset (55 days, 7%): Spans from 2025-01-30 to 2025-03-25, used for optimising the step size γ and running the Permutation Feature Importance (PFI) analysis.
- Test dataset (170 days, 23%): Spans from 2025-08-09 to 2026-01-25, acting as the unseen online phase for model evaluation.

The four subsets are separated by temporal gaps due to data availability.

As mentioned in Section III, only the predictions, observed values, and explanatory variables are required to obtain a probabilistic wrapper model. In the absence of historical DST inbound demand predictions, these baselines are retrospectively reconstructed from ETFMS Flight Data (EFD), using the exact logic employed by the DST (which is explained in Section II.B). This ensures that the reconstruction serves as an accurate representation of the true DST predictions. The EFD dataset used contains flight progress information for ATFM purposes, crucially including the estimated and actual entry times of the EHAA CBA necessary for the inbound demand calculation.

C. Model Features

As established in Section III, the DST prediction error is the dependent variable, modelled using a set of 56 explanatory (independent) variables. These features encompass operational and temporal metrics, including, among others, the prediction horizon, day of the week, hour of the day, the share of airborne flights, boundary-proximity flight shares, and the deterministic DST prediction itself. Seventeen of the 56 features are specifically engineered to quantify time series patterns. These variables track the evolution of a forecast by measuring the magnitude of step changes, rolling standard deviations, and adjacent-window volatility as well as single time-window prediction volatility, while strictly preventing data leakage. For example, high predictive volatility can potentially serve as a strong indicator of large prediction errors, which should lead to wider prediction intervals. Finally, the feature space includes 18 meteorological variables derived from Schiphol Kansverwachting (SKV) weather forecasts provided by LVNL*.

D. Model Configuration

To ensure alignment with LVNL's deterministic DST, predictions are generated for 20-minute sliding time windows with 5-minute intervals.

The probabilistic framework is applied across a 0- to 5-hour prediction horizon. ATFM regulations need to be applied at least 2 to 3 hours in advance to translate into ground delays. By modelling up to 5 hours in advance, the forecast covers this lower limit with some margin.

For the purposes of this study, the nominal target coverage level is set to 90%, corresponding to a 10% nominal target error rate.

This configuration is used to demonstrate the framework's capabilities, with the understanding that the parameters can be tailored to operational preferences.

E. Model Optimisation

Prior to testing the hypotheses, the model's hyperparameters are tuned. This process determines the optimal number of leaves, learning rate, and total iterations of the LightGBM model. To do this, a Random Search algorithm is employed

on the training dataset, using the pinball loss function. The search space is defined as follows: the number of leaves is sampled from integers in the range [20, 200], with a step size of 10; the learning rate is drawn from 20 evenly spaced values on the interval [0.0001, 1]; and the number of estimators is sampled from integers in the range [100, 2000] with a step size of 100. A total of 100 random trials are used, and the lower and upper quantiles are evaluated separately. While potentially more exhaustive hyperparameter tuning methods exist, Random Search has been proven to be a highly efficient and competitive alternative at a fraction of the computational cost [25].

Additionally, the step size γ in the ACI framework requires optimisation. This is achieved by iteratively investigating the model's performance for various step sizes using the validation dataset.

Finally, Permutation Feature Importance (PFI) is employed using the validation dataset to identify and discard non-contributing features. When analysing a specific feature, its values in the dataset are randomly shuffled, creating meaningless noise. The trained model is then used to create prediction intervals. This is repeated five times for each feature, mitigating the effect of random variation. If the model performance (measured by its MWIS) remains unchanged, the feature is deemed uninformative and can be removed in the optimised model. This process retains only the features that meaningfully contribute to the model's predictive performance.

F. Model Evaluation

To evaluate the prediction intervals, two metrics are important: the obtained coverage level and average interval width. These can be combined into a single metric through the Mean Winkler Interval Score (MWIS). The MWIS balances interval width and coverage accuracy, acting as a useful performance metric for the analysis and comparison of prediction intervals. It rewards narrow prediction intervals but heavily penalises violations where the observed value falls outside the bounds. The Winkler Interval Score is mathematically defined by Eq. (8) [26], and the MWIS is the mean Winkler Interval Score across the whole dataset. $u = \hat{q}_{1-\frac{\alpha}{2}}(X_t)$ and $l = \hat{q}_{\frac{\alpha}{2}}(X_t)$ denote the upper and lower bounds of the prediction interval, respectively, and W represents the Winkler Interval Score.

$$W = \begin{cases} (u - l) + \frac{2}{\alpha}(l - y_t), & \text{if } y_t < l \\ (u - l), & \text{if } l \leq y_t \leq u \\ (u - l) + \frac{2}{\alpha}(y_t - u), & \text{if } y_t > u \end{cases} \quad (8)$$

The mean absolute deviation (MAD) from the target coverage level, given a certain rolling window period, is used to analyse the stability and reactivity of the different models. The rolling window period is set to 2016 steps, constituting exactly one week. This allows analysis of model performance under sustained distribution shifts. As a consequence, distribution shifts of a much smaller time scale are not clearly visible in the results.

In order to evaluate the operational impact of the proba-

*LVNL receives this data from the Koninklijk Nederlands Meteorologisch Instituut (KNMI)

bilistic approach, false positive rate (FPR) and false negative rate (FNR) baselines are first established, mathematically described by Eq. (9) and Eq. (10). FN denotes false negatives, TP true positives (i.e., actual capacity breaches which are correctly predicted), TN true negatives (i.e., actual non-capacity breaches which are correctly predicted), and FP false positives. In this context, a false negative is defined as a missed capacity overload, which is a scenario where no ground delay is requested because the predicted demand remains below the threshold, resulting in a capacity breach. Conversely, a false positive constitutes an unnecessary intervention. It is a scenario where an anticipated overload triggers ATFM regulations, but the synthetic true demand naturally remains within capacity limits.

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \quad (9)$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (10)$$

The baseline mimics current operations, assuming ATFM regulations are requested by the Flow Management Position Controller (FMPC) when the deterministic DST prediction exceeds a fixed capacity limit of 65 intervals per hour (equivalent to 21.67 per 20-minute time window). This deterministic baseline is subsequently compared against the performance of the probabilistic framework, where interventions are instead triggered by specific predictive quantiles (e.g., a 20% capacity breach risk level corresponding to the 80th percentile and upper bound of a 60% prediction interval). The result produces a continuous curve of false negative rates (FNRs) and false positive rates (FPRs). If the DST baseline falls to the upper right of the curve, it indicates that the model successfully captures the heteroscedastic uncertainty of inbound demand and shows how operational improvements can potentially be realised.

As mentioned above, ATFM regulations preferably need to be requested 2 to 3 hours in advance. For that reason, when a single prediction horizon is analysed, a 150-minute horizon is used.

By applying the approach outlined in this section with the methodology presented in Section III, it is possible to evaluate the usefulness of a probabilistic top-down conformal prediction framework for ATFM inbound demand forecasting.

V. Results

This section presents the results of the proposed probabilistic forecasting framework. The section is structured as follows: Section V.A covers model optimisation, Section V.B benchmarks the machine learning model against a statistical baseline, Section V.C evaluates the benefit of applying the adaptive ACI algorithm on top of the static CQR framework, and Section V.D explores the potential operational value of the probabilistic approach relative to LVNL's deterministic DST baseline.

Fig. 11 illustrates a representative model output. It shows the result of the adaptive prediction interval, in this case for a 150-minute prediction horizon, constructed around the DST

point forecasts. When deployed in an operational context, the combination of all prediction horizons forms a probabilistic forecast envelope spanning from zero to five hours ahead, shown by Fig. 2. The synthetic inbound demand shown in the figures is unavailable during live operations and can only be reconstructed in hindsight.

A. Model Optimisation Results

Based on the PFI analysis in which 56 features are evaluated using the validation dataset, the 17 features listed in Table 1 are used in the optimised model because they meaningfully contribute to the probabilistic model performance. Table 1 shows the results when PFI is performed on the test dataset, to indicate the importance of the features for the model. SUGOL, ARTIP, and RIVER are the three initial approach fixes of EHAM, and KLM is the largest airline at the airport.

Table 1 Results of a Permutation Feature Importance (PFI) analysis using the 17 features from the optimised model, run on the test dataset to indicate feature importance for model performance.

Metric	$\Delta\%$ MWIS
DST demand prediction	+51.32%
Share of airborne flights	+11.79%
Demand prediction difference between current and leading time window	+8.64%
Share of flights with ATFM delay	+7.51%
Sum of predicted demand differences between trailing 12 time windows (i.e., up to 1h in the past, prediction made at same moment in time)	+6.55%
Inbound demand prediction difference between current and trailing time window	+5.21%
Hour category	+2.90%
Prediction horizon	+2.03%
Share of KLM flights	+1.44%
Demand of flights over SUGOL	+1.37%
Sum of predicted demand differences between leading 12 time windows (i.e., up to 1h in the future, prediction made at the same moment in time)	+1.32%
Volatility of prediction of same time window (up to 1h ago)	+1.29%
Demand of flights over ARTIP	+1.05%
Share of flights scheduled to arrive in first or last three minutes of a time window	+0.85%
Demand of flights over RIVER	+0.83%
Share of short-haul flights	+0.37%
Demand prediction difference between previous and current prediction for same time window	+0.16%

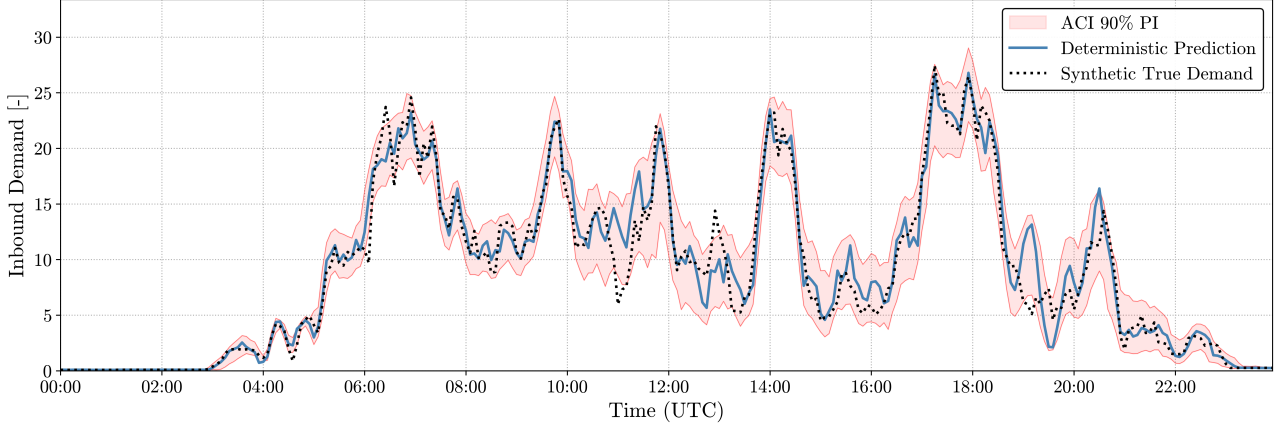


Fig. 11 Adaptive PI for 2025-12-12 with a 150-minute prediction horizon.

Evidently, the DST inbound demand prediction has a very strong impact on the performance of the model. Furthermore, the prediction horizon appears to be of relatively little importance. This is in contrast with Wanke et al. [11], who found the prediction horizon to be one of the three important features in demand uncertainty. With all 56 features, the ACI model achieves a MWIS of 5.68 on the test dataset. Using only the 17 features mentioned in Table 1, the MWIS only increases by 0.2% to 5.69. Thus, with a minor decrease in performance, the size of the model can be reduced substantially, increasing its computational efficiency. Further reducing the feature set to the top six features with $\Delta\%MWIS > 5.0\%$ results in a MWIS of 5.85, a substantial 3.0% increase. None of the seventeen weather-related features has a noticeable impact on model performance.

A 0.101 learning rate, 90 leaves, and 800 iterations are used in the final model to train the quantiles. Additionally, a step size γ of 0.0008 is used in the ACI framework, striking a balance between reactivity and stability. This model configuration is used to obtain the results presented below.

B. Comparison With a Statistical Error-Margin Baseline

The statistical error-margin baseline is compared against the uncalibrated, non-adaptive machine learning model rather than the final ACI framework. This distinction is made because the adaptive conformal calibration layer is designed to adapt empirical calibration behaviour independently of the underlying predictive model. Evaluating the uncalibrated machine learning model therefore isolates the contribution of the heteroscedastic uncertainty modelling itself relative to the statistical baseline.

The numerical results presented in Table 2 show that although the ML model undercovers relative to the nominal 90% target, its deviation from the target is smaller than that of the statistical baseline. Furthermore, it produces 26.1% narrower prediction intervals on average. This combination leads to a 25.1% lower MWIS.

A similar trend in superior performance of the ML approach is observed when considering only daytime operations (0400 - 2159 UTC), which represents the most capacity-constrained, and therefore operationally relevant, period for

LVNL. In this subset, the ML model achieves a MWIS of 7.14, compared to 8.76 for the statistical baseline.

These results indicate that the machine learning approach better captures the heteroscedasticity of the forecast uncertainty than the statistical error-margin PI does, providing evidence in support of the first hypothesis.

Table 2 Empirical coverage, prediction interval (PI) width, and MWIS averaged across all evaluated prediction horizons for the statistical error-margin baseline and the uncalibrated machine learning (ML) model evaluated on the test dataset.

Metric	Statistical Baseline PI	Uncalibrated ML PI
Coverage [-]	93.4%	87.3%
Average PI Width [-]	6.06	4.48
MWIS [-]	7.61	5.70

C. Adaptive vs. Static Calibrated Model Performance

Fig. 12 shows the rolling mean empirical coverage for the 150-minute prediction horizon. It illustrates the improved temporal stability of the ACI framework, as its deviations from the nominal target coverage level are substantially smaller than those of the other two models. The rolling mean of the CQR framework exhibits a noticeable drift from the target coverage level from time step 17,000 onwards. However, the ACI model keeps fluctuating around this target. This improved performance of the ACI framework is observed at all prediction horizons. Numerically, the tighter fluctuations of the ACI framework are confirmed by its MAD of 0.74 percentage points, compared to 1.49 percentage points for the CQR framework, averaged across all evaluated prediction horizons. This behaviour illustrates the ability of the ACI framework to adapt its prediction interval calibration under distribution shifts and maintain empirical coverage closer to the nominal target level over time.

This observation is also confirmed by Fig. 13. Where the uncalibrated model systematically undercovers, and the CQR model consistently overcovers, the ACI model closely matches the nominal target coverage level across all predic-

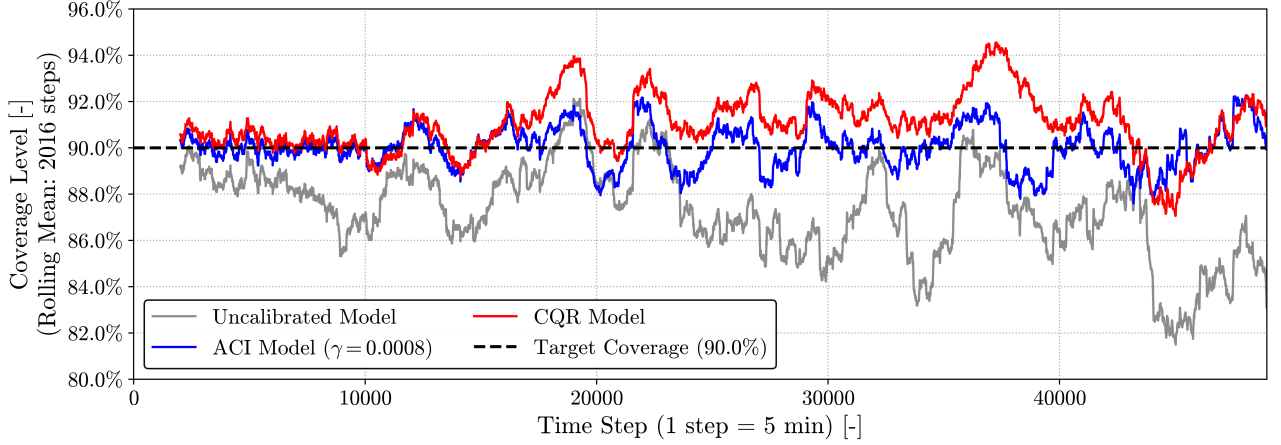


Fig. 12 Rolling 2016-step (1 week) average PI coverage level for a 150-minute prediction horizon across the test dataset for the uncalibrated, static (CQR), and adaptive (ACI) models.

tion horizons. Similarly, the adaptive approach has a lower mean deviation from the nominal target across all evaluated prediction horizons than the CQR framework.

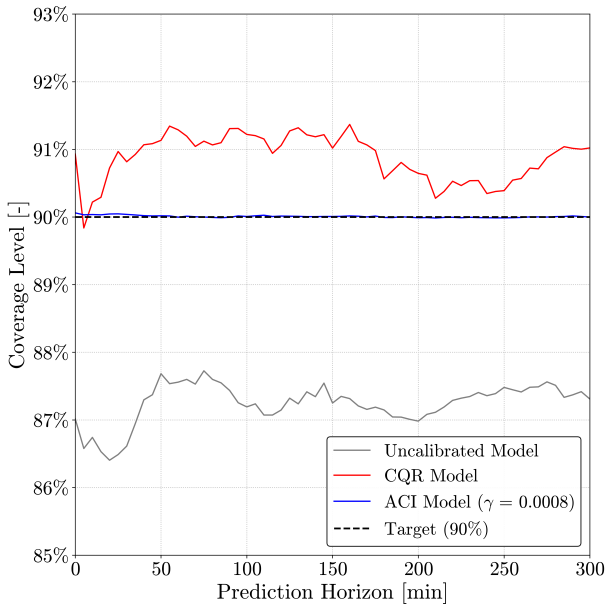


Fig. 13 Average PI empirical coverage per prediction horizon for the uncalibrated, static (CQR), and adaptive (ACI) models evaluated on the test dataset.

The numerical results are summarised in Table 3. Although the absolute MWIS differences are small (5.69 for the ACI approach versus 5.70 for the CQR model), ACI achieves improved coverage stability while also slightly reducing the average interval width, as shown by Fig. 14.

Thus, while the rolling coverage of the static conformal calibration model drifts from the target, the adaptive calibration mechanism enables the ACI framework to more effectively track the nominal target coverage level over time while maintaining comparable interval width. These findings align with the second hypothesis.

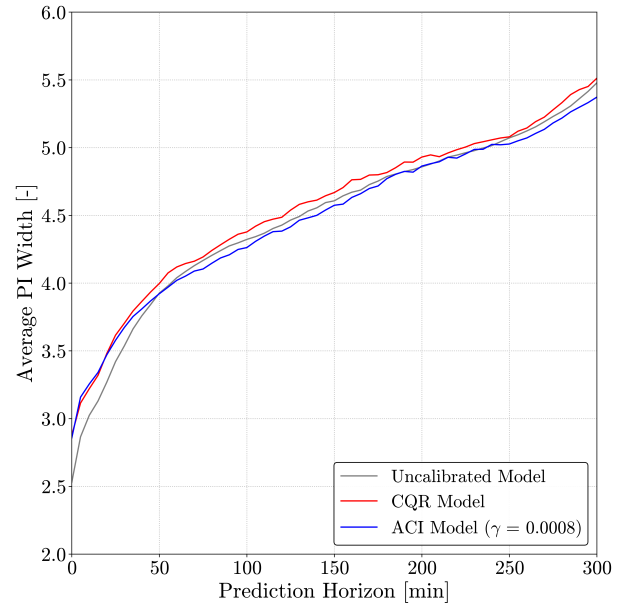


Fig. 14 Average prediction interval (PI) width per prediction horizon for the uncalibrated, static (CQR) and adaptive (ACI) models evaluated on the test dataset.

Table 3 Empirical coverage and MWIS of the uncalibrated, static (CQR), and adaptive (ACI) models evaluated on the test dataset averaged across all evaluated prediction horizons.

Model	Coverage [-]	MWIS [-]
Uncalibrated Model	87.3%	5.70
CQR Framework	90.9%	5.70
ACI Framework	90.0%	5.69

D. Operational Effectiveness Evaluation

Fig. 15 illustrates the operational trade-off between the FNR and FPR for varying probabilistic risk levels of capacity breaches.

Operationally, a highly conservative policy (i.e., low risk tolerance for capacity breaches) results in frequent interventions, yielding a high FPR but a low FNR. Conversely, a less conservative policy reduces unnecessary interventions at the expense of an increased probability of missed capacity exceedances. The continuous probabilistic curve can therefore assist the FMPC in navigating the trade-off between unnecessary regulations and missed capacity breaches.

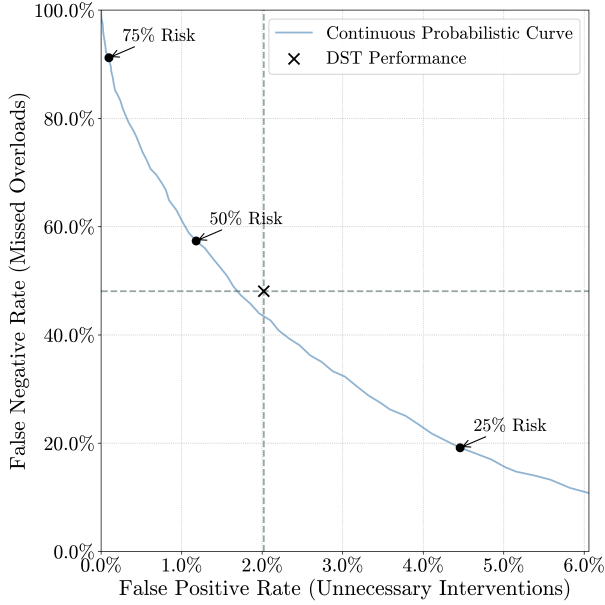


Fig. 15 Trade-off between the false negative rate and false positive rate for varying risk levels at a capacity limit of 65 aircraft per hour, compared to the deterministic DST baseline (prediction horizon = 150 min) evaluated on the test dataset.

Considering predictions generated with a prediction horizon of 150 minutes, the deterministic DST model produces an FNR of 48.1% and an FPR of 2.0%. The relatively low FPR compared to the FNR is explained by the large difference between capacity exceedance events and non-events, as breaches are relatively rare in the dataset. Incorporating predictive uncertainty into the decision-making process makes alternative operating points available. For example, selecting a risk tolerance of 44%, corresponding to the 56th percentile of the probabilistic distribution, leads to a relative reduction of 1.6% in FNR and 14.1% in FPR compared to the DST performance. These results are consistent with the third hypothesis.

Importantly, these probabilistic quantiles should not be interpreted as improved deterministic demand forecasts that replace the original DST prediction. Rather, the deterministic forecast remains the main demand estimate, while the probabilistic framework provides additional information regarding the uncertainty surrounding that estimate. The selected quantile merely defines the operational decision threshold used to determine when an intervention should be initiated.

The improved retrospective performance observed in

Fig. 15 is visible across all prediction horizons. It suggests that the proposed framework captures meaningful correlations in the forecast uncertainty that are not represented by the deterministic point forecast. Consequently, the probabilistic framework may support more informed operational decision-making by enabling the FMPC to dynamically balance the risk of unnecessary ATFM regulations against the risk of missed capacity exceedances according to the desired operational risk tolerance.

VI. Discussion

This section is concerned with a critical discussion of the methodology and findings presented in this paper. First, the assumptions and limitations underlying the conducted research are discussed in Section VI.A. Thereafter, Section VI.B presents recommendations for future research.

A. Limitations

Four principal limitations are identified in this research and will be discussed below.

A fundamental assumption underlying this research is the use of synthetic inbound demand as the ground truth for both model training and prediction interval evaluation. This has two limitations. As the true unconstrained demand is inherently unobservable, an exact ground truth cannot be established, and thus the accuracy of the synthetic demand cannot be evaluated. The synthetic demand methodology nevertheless approaches this quantity in a principled manner. Furthermore, the synthetic demand is constructed using the true flight time of each aircraft, as described in Section III.A. However, in an operational setting, when the performance of the prediction intervals is constantly evaluated through the ACI framework, the true flight time of an aircraft that has not yet arrived is unknown at the time of evaluation. Consequently, a modification to this approach would be required before using the ACI framework in an operational setting. When comparing dynamically adjusted prediction interval widths from the synthetic demand approach used for training with those from an operationally deployable approach, 50% of the absolute differences are below 0.21, and 99% are smaller than 1.17. The synthetic demand calculation methodology is retained for model training because it constitutes a closer representation of the hypothetical unconstrained demand than the operationally applicable method.

Ideally, diverted flights are taken into account in the same way as cancelled flights are, as they both constitute natural demand uncertainty. However, the demand calculation excludes diverted flights because they cannot be reliably distinguished from flights with data recording errors in this dataset. As mentioned in Section IV.B, these flights account for only 0.21% of the dataset, resulting in a negligible impact on the model.

The operational performance metric relies on an exceedance probability derived from calibrated prediction intervals, as described in Section III.E. Adaptive prediction intervals are not used for this purpose because the cumulative distribution function requires a strictly monotonically increas-

ing curve. This property can be violated when adaptively adjusted quantiles cross. To resolve this, quantiles are sorted once following calibration, which precludes further adaptive adjustments.

Determining the optimal step size γ proves to be challenging, as it governs a fundamental trade-off between stability and reactivity. A small step size yields stable prediction intervals with relatively low variance around the target coverage level under normal conditions, but responds slowly to sudden changes in uncertainty arising from distribution shifts. Conversely, a large step size produces highly reactive intervals that rapidly widen and contract in response to these distribution shifts. This improves coverage, but at the cost of more erratic prediction intervals. The effect of a given step size differs between prediction horizons, indicating that the adaptive calibration dynamics are dependent on this feature. This is likely caused by the increasing impact of delayed feedback at longer prediction horizons. The results suggest that horizon-specific step sizes may therefore be preferable, with larger values of γ for shorter prediction horizons and smaller values for longer prediction horizons. This is intuitive, as shorter horizons can benefit from faster adaptation, whereas longer horizons are affected by greater feedback delays, making overly reactive updates less desirable. The choice of γ is also sensitive to the timescale over which the performance is assessed. Smaller rolling windows produce larger deviations from the nominal target coverage level because they are more responsive to short-term fluctuations, whereas larger windows reflect longer-term calibration stability. As a result, the step size that minimises the MAD from the target coverage level over a one-week rolling window differs from the one that performs best over a 24-hour window. Moreover, the step size that minimises the MWIS does not necessarily coincide with the one that minimises the MAD from the target coverage level. Consequently, no universally optimal γ is found, although the analysis does make it possible to identify rough bounds within which to search for a suitable step size.

Lastly, the ideal objective in uncertainty quantification is conditional coverage, meaning that the prediction interval attains the nominal coverage level not only on average, but also conditional on the available operational context, such as the prediction horizon, time of day, traffic situation, and other relevant factors. In practice, conditional coverage is only possible by imposing strong assumptions on the data-generating process. Such assumptions are unlikely to hold in a dynamic operational environment, and therefore the ACI approach provides improved long-run empirical coverage behaviour without providing coverage guarantees. Nevertheless, conditional reliability is partially assessed in practice by stratifying model performance across relevant dimensions, such as prediction horizon and time of day. Although no formal conditional coverage guarantee can be claimed, the empirical results indicate that the framework maintains relatively consistent performance across several operational conditions.

B. Recommendations for Future Work

Building upon the findings and limitations presented in this study, several promising directions for future research are

identified.

First, as discussed in Section VI.A, selecting an appropriate step size γ for the ACI framework remains non-trivial. The optimal value reflects an inherent trade-off between reactivity and stability, which is user-dependent. The choice of timescale introduces an additional degree of freedom that further complicates calibration. To eliminate the need for manual tuning, future work should investigate advanced adaptive frameworks, such as Aggregated ACI [27] and Dynamically-tuned ACI [28]. Evaluating whether these methods can autonomously maintain robust coverage during sudden uncertainty shifts without sacrificing interval stability is a highly relevant next step.

Second, while this study takes a top-down approach to probabilistic inbound demand forecasting because of its computational efficiency and structural advantages, its empirical performance relative to bottom-up uncertainty quantification remains an open question. Future research should therefore compare the predictive performance and computational cost of the top-down and bottom-up approaches to determine the optimal methodology for probabilistic inbound demand forecasting.

Finally, and critically, to bridge the gap between theoretical validity and operational utility, the framework must be deployed in a live environment. Shadow testing alongside current systems will uncover real-world limitations that theoretical evaluation cannot capture. Most importantly, capturing practitioner feedback on how these probabilistic intervals inform human-in-the-loop decision-making will be essential for outlining a detailed roadmap for full operational implementation.

VII. Conclusion

The goal of this research is to investigate the application of a conformal prediction framework to inbound demand forecasting in order to equip decision-makers with uncertainty information about deterministic predictions. The Dutch Air Navigation Service Provider (ANSP), Luchtverkeersleiding Nederland (LVNL), is used as a case study to apply the methodology to real-world operational scenarios. The findings show that the proposed probabilistic approach can effectively model uncertainty around deterministic inbound demand forecasts. By generating calibrated prediction intervals, the framework can help ANSPs take forecast uncertainty into account during the demand and capacity balancing process, rather than relying solely on point predictions.

The study identifies key explanatory features that allow the model to capture the heteroscedastic nature of inbound demand uncertainty. In particular, the magnitude of the inbound demand prediction, the share of airborne flights, changes in predicted demand over time, and the share of flights with Air Traffic Flow Management delays are found to be important predictors of uncertainty. Furthermore, the empirical results show that integrating an Adaptive Conformal Inference framework with Conformalized Quantile Regression (CQR) improves temporal coverage stability while maintaining comparable prediction interval width. Compared with the

uncalibrated model and the static CQR framework, the adaptive approach achieves average empirical coverage closer to the nominal target level across the evaluated prediction horizons.

The operational analysis indicates that the probabilistic framework has the potential to support more flexible decision-making by allowing different operating points to be selected according to the preferred risk tolerance for capacity breaches. For a 150-minute prediction horizon, one specific probabilistic risk threshold reduces the false negative and false positive rates for capacity breach detection by 1.6% and 14.1%, respectively, relative to the deterministic demand prediction baseline in the test dataset evaluation.

Future work should focus on exploring more specialised adaptive conformal prediction techniques to address the challenge of step-size tuning. In addition, the top-down probabilistic approach employed in this study should be benchmarked against a probabilistic bottom-up forecasting framework. Lastly, further research should address practical limitations of the model when deployed in an operational environment.

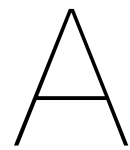
References

- [1] EUROCONTROL, “Air traffic in the European network in summer 2025,” <https://www.eurocontrol.int/summer-2025>, 2025.
- [2] Tielrooij, M., Borst, C., Van Paassen, M. M., and Mulder, M., “Predicting arrival time uncertainty from actual flight information,” *Proceedings of the 11th USA/Europe Air Traffic Management Research and Development Seminar*, FAA/EUROCONTROL, 2015, pp. 577–586.
- [3] Strottmann Kern, C., de Medeiros, I. P., and Yoneyama, T., “Data-driven aircraft estimated time of arrival prediction,” *2015 Annual IEEE Systems Conference (SysCon) Proceedings*, IEEE, 2015, pp. 727–733. <https://doi.org/10.1109/SYSCON.2015.7116837>.
- [4] Dalmau, R., Trzmiel, A., and Kirby, S., “Combining Machine Learning Models to Improve Estimated Time of Arrival Predictions,” *European Journal of Transport and Infrastructure Research*, Vol. 25, No. 1, 2025. <https://doi.org/10.59490/ejtir.2025.25.1.7488>.
- [5] Ayhan, S., Costas, P., and Samet, H., “Predicting estimated time of arrival for commercial flights,” *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 33–42. <https://doi.org/10.1145/3219819.3219874>.
- [6] Vos, R., Sun, J., and Hoekstra, J., “A transformer-based trajectory prediction model to support air traffic demand forecasting,” *Proceedings of the 11th International Conference for Research in Air Transportation (ICRAT)*, 2024.
- [7] Meyn, L., “Probabilistic methods for air traffic demand forecasting,” *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2002, p. 4766. <https://doi.org/10.2514/6.2002-4766>.
- [8] Gilbo, E., and Smith, S., “Probabilistic prediction of aggregate traffic demand using uncertainty in individual flight predictions,” *AIAA Guidance, Navigation, and Control Conference*, 2009. <https://doi.org/10.2514/6.2009-6194>.
- [9] Chatterji, G., Sridhar, B., Sheth, K., Kim, D., and Mulfinger, D., “Methods for Establishing Confidence Bounds on Sector Demand Forecasts,” *AIAA Guidance, Navigation, and Control Conference and Exhibit*, American Institute of Aeronautics and Astronautics, 2004. <https://doi.org/10.2514/6.2004-5232>.
- [10] Gonze, F., Huens, E., Jungers, R. M., Simonetto, A., and Boucquoy, J., “Probabilistic occupancy counts and flight criticality measures in air traffic management,” *Journal of Air Transportation*, Vol. 26, No. 3, 2018, pp. 94–103. <https://doi.org/10.2514/1.D0087>.
- [11] Wanke, C., Callaham, M., Greenbaum, D., and Masalonis, A., “Measuring uncertainty in airspace demand predictions for traffic flow management applications,” *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2003, p. 5708. <https://doi.org/10.2514/6.2003-5708>.
- [12] Wanke, C., Song, L., Zobell, S., Greenbaum, D., and Mulgund, S., “Probabilistic congestion management,” *Proceedings of the 6th Europe-USA ATM Seminar*, Baltimore. US, 2005.
- [13] Verboon, T., Tuinstra, E., van den Berg, R., Voogt, M., de Gelder, N., and Dijkstra, F., “Multi-Airport Concept: Improved Management of Air Traffic Flows in the Netherlands,” Report by To70, NLR, and Ferway, 2020. URL <https://kdc-mainport.nl/wp-content/uploads/2021/04/19.282.05-Final-report-Multi-Airport-Concept.pdf>.
- [14] Ruskamp, T. M., “Prediction of Traffic Take-Off Times at Out-stations,” Master’s thesis, Delft University of Technology, 2025.
- [15] Brando, A., Torres-Latorre, C., Rodriguez-Serrano, J. A., and Vitria, J., “Building uncertainty models on top of black-box predictive APIs,” *IEEE Access*, Vol. 8, 2020, pp. 121344–121356. <https://doi.org/10.1109/ACCESS.2020.3006711>.
- [16] O’Sullivan, A., Pereira, F. C., Zhao, J., and Koutsopoulos, H. N., “Uncertainty in bus arrival time predictions: Treating heteroscedasticity with a metamodel approach,” *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, No. 11, 2016, pp. 3286–3296. <https://doi.org/10.1109/TITS.2016.2547184>.
- [17] Pereira, F. C., Antoniou, C., Fargas, J. A., and Ben-Akiva, M., “A metamodel for estimating error bounds in real-time traffic prediction systems,” *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No. 3, 2014, pp. 1310–1322. <https://doi.org/10.1109/TITS.2014.2300103>.
- [18] Makridakis, S., Spiliotis, E., and Assimakopoulos, V., “M5 accuracy competition: Results, findings, and conclusions,” *International Journal of Forecasting*, Vol. 38, No. 4, 2022, pp. 1346–1364. <https://doi.org/10.1016/j.ijforecast.2021.11.013>.
- [19] Januschowski, T., Wang, Y., Torkkola, K., Erkkilä, T., Hasson, H., and Gasthaus, J., “Forecasting with trees,” *International Journal of Forecasting*, Vol. 38, No. 4, 2022, pp. 1473–1481. <https://doi.org/10.1016/j.ijforecast.2021.10.004>.
- [20] Grinsztajn, L., Oyallon, E., and Varoquaux, G., “Why do tree-based models still outperform deep learning on typical tabular data?” *Advances in Neural Information Processing Systems*, Vol. 35, 2022, pp. 507–520. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html.

- [21] Vovk, V., Gammerman, A., and Shafer, G., *Algorithmic learning in a random world*, Springer, 2005.
- [22] Gibbs, I., and Candès, E., “Adaptive conformal inference under distribution shift,” *Advances in Neural Information Processing Systems*, Vol. 34, 2021, pp. 1660–1672. URL <https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html>.
- [23] Koenker, R., and Bassett Jr., G., “Regression quantiles,” *Econometrica: Journal of the Econometric Society*, 1978, pp. 33–50. <https://doi.org/10.2307/1913643>.
- [24] Romano, Y., Patterson, E., and Candès, E., “Conformalized Quantile Regression,” *Advances in Neural Information Processing Systems*, Vol. 32, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html>.
- [25] Bergstra, J., and Bengio, Y., “Random search for hyperparameter optimization,” *Journal of Machine Learning Research*, Vol. 13, 2012, pp. 281–305. URL <https://jmlr.org/papers/v13/bergstra12a.html>.
- [26] Winkler, R. L., “A decision-theoretic approach to interval estimation,” *Journal of the American Statistical Association*, Vol. 67, No. 337, 1972, pp. 187–191. <https://doi.org/10.1080/01621459.1972.10481224>.
- [27] Zaffran, M., Feron, O., Goude, Y., Josse, J., and Dieuleveut, A., “Adaptive conformal predictions for time series,” *International Conference on Machine Learning*, PMLR, 2022, pp. 25834–25866. URL <https://proceedings.mlr.press/v162/zaffran22a.html>.
- [28] Gibbs, I., and Candès, E., “Conformal inference for online prediction with arbitrary distribution shifts,” *Journal of Machine Learning Research*, Vol. 25, No. 162, 2024, pp. 1–36. URL <http://jmlr.org/papers/v25/22-1218.html>.

Part II

Appendices to the Scientific Article



PFI Results

This chapter presents the results of the Permutation Feature Importance (PFI) analysis. Table A.1 summarises the tested features and their corresponding PFI values. The analysis is conducted on the validation dataset using a model configured with 90 leaves, 800 iterations, a learning rate of 0.101, and a target error rate of 10%.

Table A.1: Explanation of the features examined in the probabilistic model, including the PFI results evaluated on the validation dataset.

Feature Name	Unit	Explanation	PFI MWIS change [-]
Prediction horizon	[min]	Forecasting horizon	+1.74%
Month	[-]	Categorical feature	+0.02%
Day of the week	[-]	Categorical feature	+0.07%
Hour category	[-]	Categorical feature	+0.83%
DST inbound demand prediction	[-]	Deterministic inbound demand prediction made by LVNL	+44.66%
Border flight count	[-]	Number of flights within the first three or last three minutes of the time window	+1.30%
Border flight share	[-]	Share of flights within the first three or last three minutes of the time window	+0.70%
Share airborne flights	[-]	Share of flights within a time window that are already airborne	+12.49%
Share flights with ATFM delay	[-]	Share of flights with ATFM delay (regardless of the cause/source) within a specific time window	+6.73%
Magnitude of ATFM delay	[min]	Magnitude of ATFM delay assigned to flights within a specific time window	+0.12%
Share KLM flights	[-]	Share of flights from KLM and KLC within a time window	+1.04%
Share short haul flights	[-]	Share of flights originating from airports with ICAO codes starting with E, L, or B (i.e. almost only European flights)	+0.30%
Sine of time	[-]	Sine of the time ($= \sin(2\pi \frac{t}{24})$)	+0.32%
Cosine of time	[-]	Cosine of the time ($= \cos(2\pi \frac{t}{24})$)	+1.38%

Continued on next page...

Table A.1 – continued from previous page

Feature Name	Unit	Explanation	PFI MWIS change [-]
Count RIVER flights	[-]	Number of flights scheduled to fly over initial approach fix RIVER within a time window	+0.51%
Share RIVER flights	[-]	Share of flights scheduled to fly over initial approach fix RIVER within a time window	+0.68%
Count SUGOL flights	[-]	Number of flights scheduled to fly over initial approach fix SUGOL within a time window	+0.60%
Share SUGOL flights	[-]	Share of flights scheduled to fly over initial approach fix SUGOL within a time window	+0.02%
Count ARTIP flights	[-]	Number of flights scheduled to fly over initial approach fix ARTIP within a time window	+0.39%
Share ARTIP flights	[-]	Share of flights scheduled to fly over initial approach fix ARTIP within a time window	+0.09%
Backward difference	[-]	Inbound demand prediction difference between preceding and current time window	+5.74%
Absolute backward difference	[-]	Absolute inbound demand prediction difference between preceding and current time window	+0.04%
Backward jaggedness sum	[-]	Sum of predicted inbound demand differences between preceding 12 time windows (i.e. up to 1h ago, prediction made at same moment in time)	+6.50%
Backward jaggedness std.	[-]	Standard deviation of predicted inbound demand differences between preceding 12 time windows (i.e. up to 1h ago, prediction made at same moment in time)	+0.28%
Forward difference	[-]	Inbound demand prediction difference between current and leading time window	+10.04%
Absolute forward difference	[-]	Absolute inbound demand prediction difference between current and leading time window	+1.30%
Forward jaggedness sum	[-]	Sum of predicted inbound demand differences between leading 12 time windows (i.e. up to 1h in the future, prediction made at same moment in time)	+0.75%

Continued on next page...

Table A.1 – continued from previous page

Feature Name	Unit	Explanation	PFI MWIS change [-]
Forward jaggedness std.	[-]	Standard deviation of predicted inbound demand differences between leading 12 time windows (i.e. up to 1h in the future, prediction made at same moment in time)	+0.56%
Step change	[-]	Difference in inbound demand prediction for same time window between preceding prediction (made 5min ago) and current prediction	+0.12%
Absolute step change	[-]	Absolute difference in inbound demand prediction for same time window between preceding prediction (made 5min ago) and current prediction	+0.00%
Volatility same time window sum	[-]	Sum of predicted inbound demand differences between 12 preceding predictions for same time window (i.e. up to 1h ago)	+0.02%
Volatility same time window std.	[-]	Standard deviation of predicted inbound demand differences between 12 preceding predictions for same time window (i.e. up to 1h ago)	+0.67%
Delay minutes past 1h	[min]	Sum of assigned delay minutes of aircraft of past 1h at time of prediction	+0.00%
Delay count past 1h	[-]	Number of aircraft with ATFM regulations supposed to arrive in the past 1h	+0.04%
Delay minutes past 1h to 3h	[min]	Sum of ATFM delay of aircraft supposed to arrive 1h to 3h ago	+0.00%
Delay count past 1h to 3h	[-]	Number of aircraft with ATFM regulations supposed to arrive 1h to 3h ago	+0.02%
Delay minutes past 3h to 8h	[min]	Sum of ATFM delay of aircraft supposed to arrive 3h to 8h ago	-0.01%
Delay count past 3h to 8h	[-]	Number of aircraft with ATFM regulations supposed to arrive 3h to 8h ago	+0.02%
Cumulonimbus chance	[-]	Predicted chance of cumulonimbus clouds at EHAM	+0.00%
Snow chance	[-]	Predicted chance of snow at EHAM	-0.01%
Gusts	[kts]	Predicted wind gust speed at EHAM	-0.01%
RVR <350 m	[-]	Predicted chance that Runway Visual Range <350m	+0.06%
Dew point	[C]	Predicted dew point at EHAM	+0.11%
Continued on next page...			

Table A.1 – continued from previous page

Feature Name	Unit	Explanation	PFI MWIS change [-]
Moderate or heavy snow	[-]	Predicted chance for moderate or heavy snow at EHAM	-0.01%
Wind speed	[kts]	Predicted wind speed at EHAM	+0.05%
Visibility	[-]	Categorical feature. Predicted visibility at EHAM (G, M, A, B, C, D)	+0.01%
Temperature	[C]	Predicted temperature at EHAM	+0.29%
Freezing precipitation chance	[-]	Predicted chance of freezing precipitation at EHAM	+0.00%
Thunderstorm	[-]	Predicted chance of thunderstorms at EHAM	+0.00%
Wind direction	[deg]	Predicted wind direction at EHAM	+0.06%
RVR <550m and/or ceiling <200ft chance	[-]	Predicted chance that Runway Visual Range <550m and/or ceiling <200ft at EHAM	+0.05%
RVR <1500m and/or ceiling <300ft chance	[-]	Predicted chance that Runway Visual Range <1500m and/or ceiling <300ft at EHAM	+0.09%
Visibility <5km and/or ceiling <1000ft chance	[-]	Predicted chance that visibility <5km and/or ceiling <1000ft at EHAM	+0.03%
Visibility <5km and/or ceiling <2000ft chance	[-]	Predicted chance that visibility <5km and/or ceiling <2000ft at EHAM	+0.04%
Wind speed std.	[kts]	Predicted standard deviation of wind speed at Schiphol (EHAM)	+0.00%
Wind direction std.	[deg]	Predicted standard deviation of wind direction at Schiphol (EHAM)	+0.02%

Visibility is treated as a categorical feature, with categories assigned according to the Low Visibility Procedure (LVP) definitions provided by Table A.2 (Wijngaard et al., 2008).

Table A.2: Low Visibility Procedure (LVP) categories (Wijngaard et al., 2008). VIS is the abbreviation for visibility, and RVR for Runway Visual Range.

Visibility classification	Visibility [VIS / RVR]		Ceiling
Good	> 5 km [VIS]	and	> 1000 ft
Marginal	1.5 - 5 km [VIS]	or	300 - 1000 ft
LVP phase A	550 - 1500 m [RVR]	or	200 - 300 ft
LVP phase B	350 - 550 m [RVR]	or	< 200 ft
LVP phase C	200 - 350 m [RVR]		
LVP phase D	< 200 m [RVR]		

B

Calibration Stratification

This chapter details the calibration stratification used in the CQR framework. Specifically, Section B.1 discusses the importance of differentiating between daytime and nighttime operations when calibrating the prediction intervals. Subsequently, Section B.2 shows why the calibration step is performed separately for each prediction horizon.

B.1. Effect of Daytime vs. Nighttime Operations on Calibration

Fig. B.1 illustrates the importance of splitting the calibration step between a daytime (0400 - 2159 UTC) and nighttime (2200 - 0359 UTC) phase when calculating the expansion factor ($Q_{1-\alpha}/Q_{1-\alpha_t}$), using the conformity score distributions for 150-minute (Fig. B.1a) and 300-minute (Fig. B.1b) prediction horizons as examples. In both cases, the nighttime conformity score distributions are narrower than the corresponding daytime distributions. This stratification is introduced after observing that the ACI framework performs less well during nighttime operations. A split based directly on demand level would also be possible, but the day-night categorisation is preferred for its simplicity.

The sharp peaks visible in the conformity score distributions in Fig. B.1 reflect the lower demand levels during nighttime operations. The sharp decline to the right of the peak, towards positive conformity scores, possibly results from the zero lower bound. Compared with daytime operations, nighttime demand produces much smaller prediction errors and, consequently, conformity scores closer to zero. This explains the difference in shapes between daytime and nighttime conformity-score distributions.

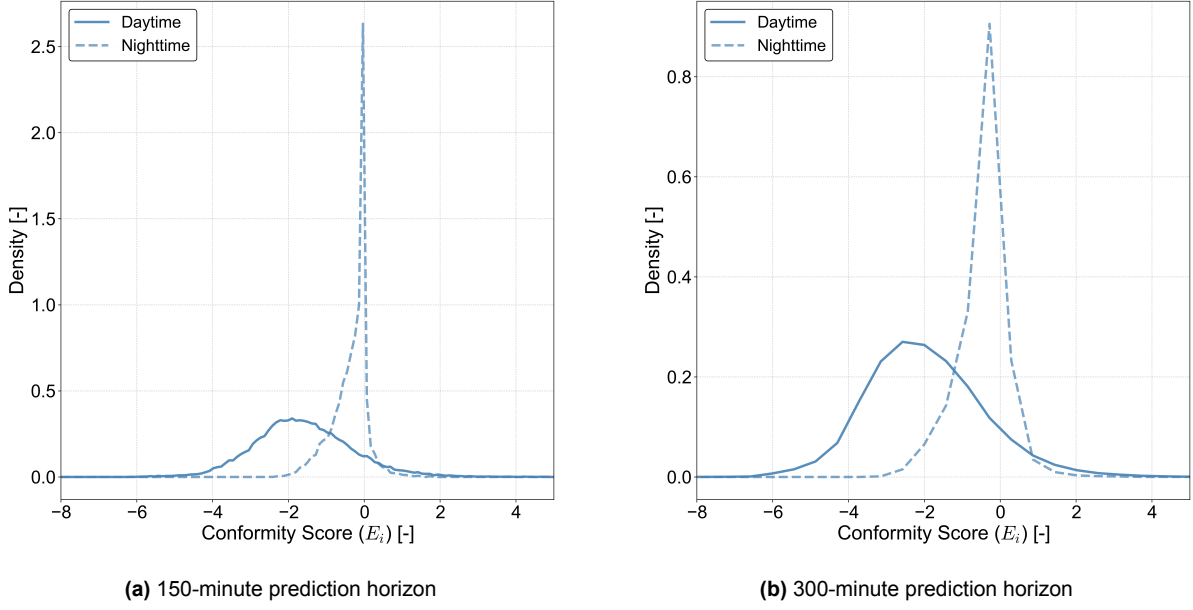


Figure B.1: Overlapping conformity score distributions for both daytime (0400 - 2159 UTC) and nighttime (2200 - 0359 UTC) operations for the 150-minute and 300-minute prediction horizons.

B.2. Effect of Prediction Horizon on Calibration

As shown in Fig. B.2, the conformity score distributions are sensitive to the prediction horizon. The narrower distributions at smaller prediction horizons are the result of lower prediction uncertainty at smaller horizons. Smaller prediction uncertainty should be reflected by smaller prediction intervals, which leads to conformity scores closer to zero and thus narrower distributions.

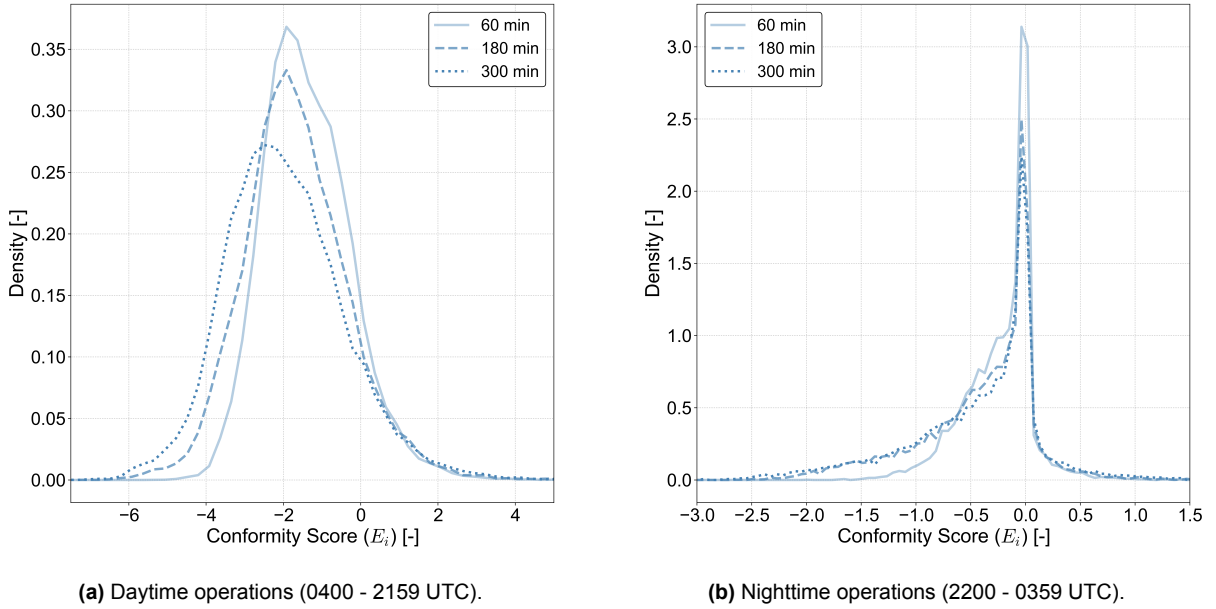


Figure B.2: Conformity score distributions for daytime and nighttime operations across 60-, 180-, and 300-minute prediction horizons.

Fig. B.3 presents the conformal expansion factor ($Q_{1-\alpha}$) used in the CQR framework, split by prediction

horizon for both daytime and nighttime operations. Together, Fig. B.2 and Fig. B.3 demonstrate the necessity of computing $Q_{1-\alpha}$ and $Q_{1-\alpha_t}$ independently for each prediction horizon, rather than using a single aggregate value.

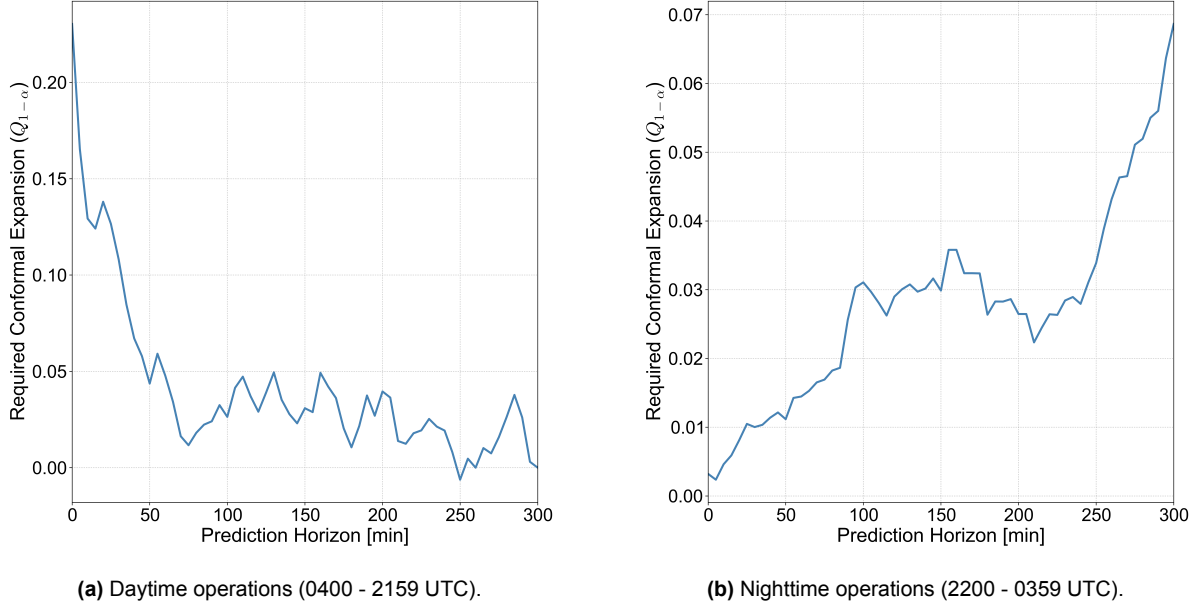


Figure B.3: Comparison of the expansion factor ($Q_{1-\alpha}$) by prediction horizon for daytime (0400 - 2159 UTC) and nighttime (2200 - 0359 UTC) operations resulting from the CQR framework.

By stratifying calibration according to both day-night period and prediction horizon, the resulting conformity score distributions better reflect the underlying uncertainty of the deterministic predictions. This leads to a more accurate calibration step and thus improved prediction interval calibration.

ACI Framework

This chapter presents results concerning the ACI framework. First, Section C.1 discusses the effect of the step size γ on model performance. Thereafter, Section C.2 examines how information lag affects the performance of the ACI framework.

C.1. Effect of Step Size on ACI Performance

The performance of the ACI framework is sensitive to the step size γ . As discussed in Part I, a larger γ produces a highly responsive model that tracks the nominal target coverage level more closely, but at the cost of volatile interval adjustments. Conversely, a smaller γ produces more stable intervals, but risks responding too slowly during distribution shifts. In terms of rolling empirical coverage, this implies that more stable prediction intervals (with a lower γ) may be associated with stronger fluctuations in the rolling mean, whereas a more responsive model (with a higher γ) may produce tighter tracking of the target coverage level. This behaviour is illustrated in Fig. C.1, where $\gamma = 0.01$ tracks the target coverage level more closely than $\gamma = 0.001$. Both models average an empirical coverage level of approximately 90.0% across the full dataset, but they differ substantially in the amplitude of their rolling mean deviations.

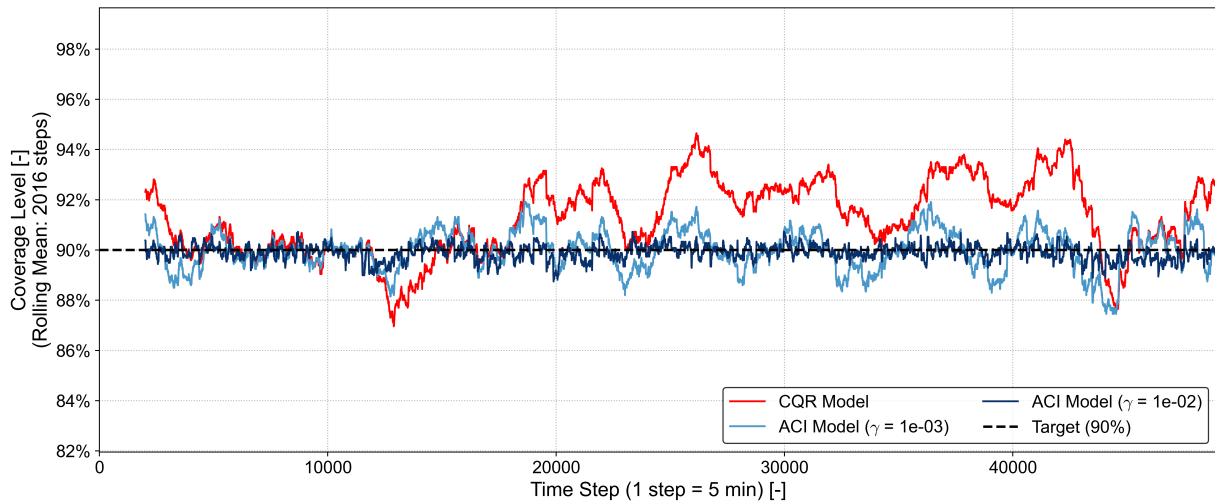


Figure C.1: 2016-step (1 week) rolling mean coverage level of various models for a 60-minute prediction horizon.

At higher prediction horizons, the performance difference between the two step sizes is less pronounced, as shown in Fig. C.2.

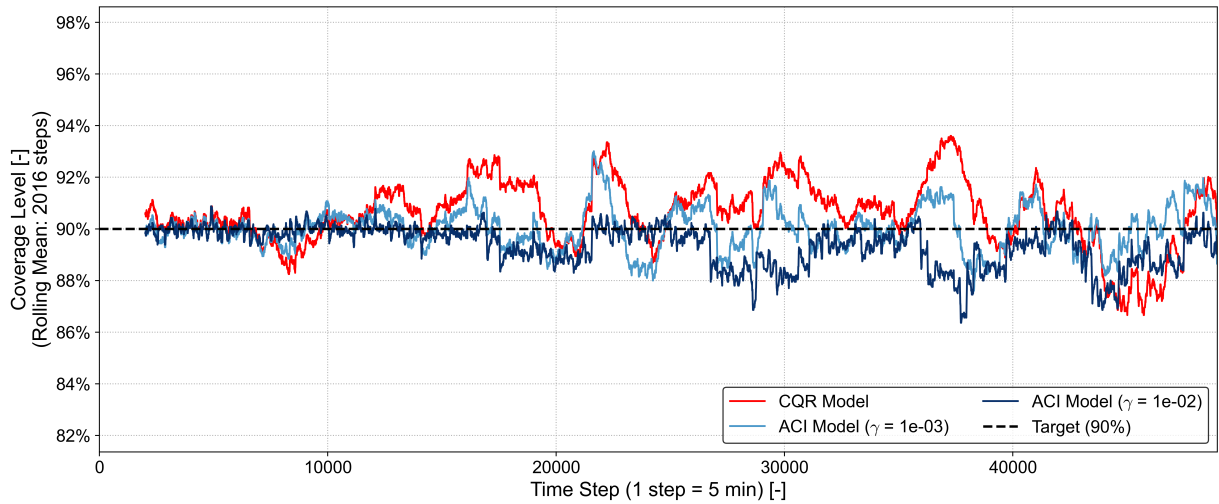


Figure C.2: 2016-step (1 week) rolling mean coverage level of various models for 180-minute prediction horizon.

This pattern is also visible in Fig. C.3 and Fig. C.4. The latter quantifies how closely the models follow the target coverage level by reporting the mean absolute deviation (MAD) of the rolling mean empirical coverage from that target across all evaluated prediction horizons. The $\gamma = 0.01$ model achieves a smaller mean absolute deviation (MAD) than the $\gamma = 0.001$ model at lower prediction horizons. From approximately the 180-minute prediction horizon onwards, however, the performance of both step sizes is similar. This is consistent with the performance differences visible when comparing Fig. C.1 and Fig. C.2. Fig. C.3 suggests that higher step sizes are more susceptible to drift from the target coverage level at longer horizons, whereas smaller step sizes are more stable. An excessively low step size (not shown in the figures), however, will show similar characteristics to the CQR framework and be prone to drifting from the target coverage level too as it does not adequately adapt to distribution shifts.



Figure C.3: Coverage level for various step sizes γ in the Adaptive Conformal Inference (ACI) model.

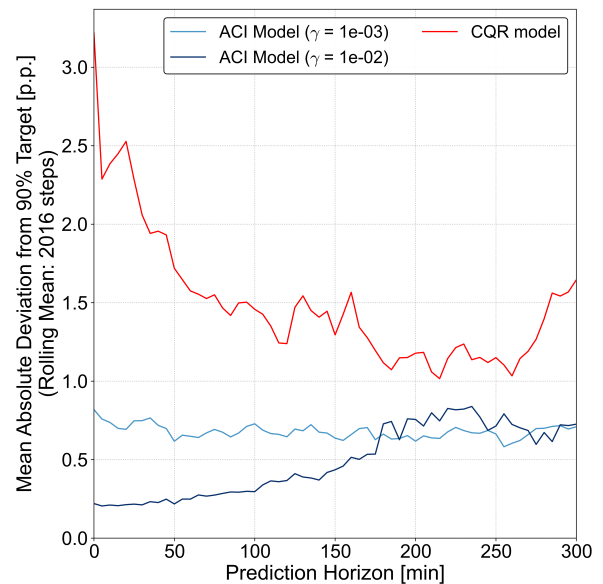


Figure C.4: MAD in percentage points (p.p.) from the target coverage level for various step sizes γ in the Adaptive Conformal Inference (ACI) model.

These trends imply that using different step sizes across the evaluated prediction horizons could improve

model performance relative to using a single global step size. This interpretation is strengthened by Fig. C.5, which shows a downward trend in the optimal step size across the evaluated horizons when minimising either MAD or Mean Winkler Interval Score (MWIS).

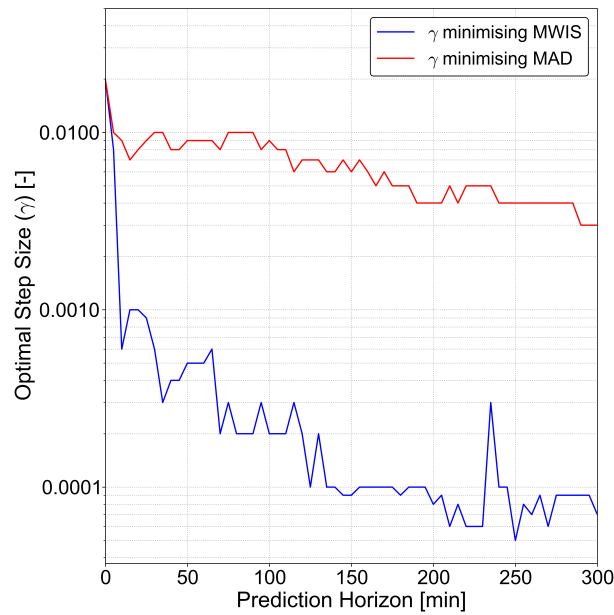


Figure C.5: Step size optimised for mean absolute deviation (MAD) and Mean Winkler Interval Score (MWIS) (2016-step rolling mean), evaluated on the validation dataset using grid search.

These results confirm that no universally optimal step size exists. The choice is a trade-off between responsiveness and stability, and it depends on operational preferences. A user prioritising rapid adaptation may favour a higher γ to minimise absolute deviation from the target, whereas operational contexts requiring stable, predictable interval bounds may prefer a lower step size. Table C.1 summarises the numerical performance for various configurations. In the validation dataset, the lowest MWIS is achieved with $\gamma = 0.0002$, while the lowest MAD is obtained with $\gamma = 0.005$.

Table C.1: Performance of various models using a 2016-step rolling mean evaluated on the test dataset. MAD in percentage points (p.p.).

Model	Coverage Level [-]	Mean PI Width [-]	MWIS [-]	Mean Absolute Deviation [p.p.]
Uncalibrated Model	87.25%	4.4822	5.6988	2.81
Static CQR model	90.88%	4.5585	5.6980	1.49
ACI Model ($\gamma = 0.0002$)	90.02%	4.4789	5.6890	1.03
ACI Model ($\gamma = 0.005$)	89.95%	4.5802	5.7683	0.34
ACI Model ($\gamma = 0.0008$)	90.01%	4.4817	5.6926	0.74
ACI Model (varying γ , minimising MWIS)	90.01%	4.4833	5.6886	0.98
ACI Model (varying γ , minimising MAD)	89.92%	4.5915	5.7724	0.32
Bernoulli Sequence	90.00%	—	—	0.58

$\gamma = 0.0008$ is selected for generating the ACI results in Part I. The choice is not unique, and nearby values

will yield comparable performance. Nevertheless, this configuration reduces the MWIS relative to the static CQR baseline, improves empirical coverage, and lowers the MAD in the 2016-step rolling window, while maintaining stable performance at higher prediction horizons.

To provide a reference for the natural variability of binary coverage outcomes, a Bernoulli sequence is included as a baseline. This sequence, shown in Fig. C.6, represents the variability expected from independent coverage events with a 90% success probability.

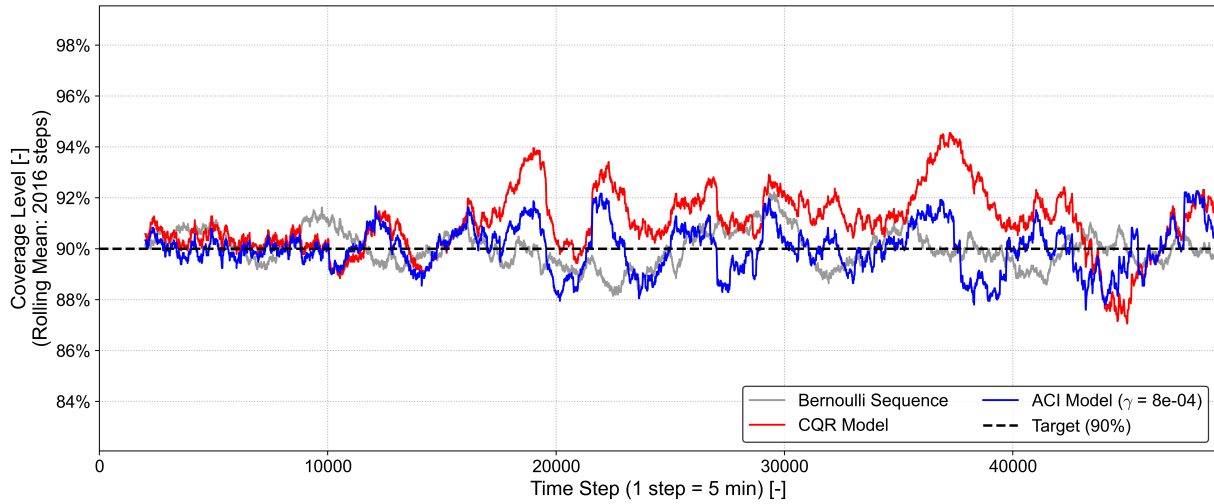


Figure C.6: 2016-step (1 week) rolling mean coverage level of CQR and ACI model and a Bernoulli sequence for a 150-minute prediction horizon evaluated on the test dataset.

It should be noted, however, that the MAD is sensitive to the rolling window size used for its computation. For the results presented above, a 2016-step rolling mean is applied. For windows of different sizes, the relative ranking of step sizes may change, because the window length determines which frequencies of volatility are smoothed over and which are captured in the deviation metric. Consequently, the optimal γ also depends on the user's preferred stability horizon (i.e., the timescale of which consistent coverage behaviour is desired). The smaller the desired timescale, the larger the required step size, and thus the more erratic the prediction intervals, will need to be to keep the rolling mean close to the target coverage level. For example, with a 288-step rolling window (representing 24 hours), the lowest MAD is achieved with a step size of 0.01. Fig. C.7 shows that a 288-step rolling mean shows a downward trend similar to that of a 2016-step rolling mean. Overall, a higher step size is preferred for smaller rolling windows to achieve a low mean absolute deviation (MAD).

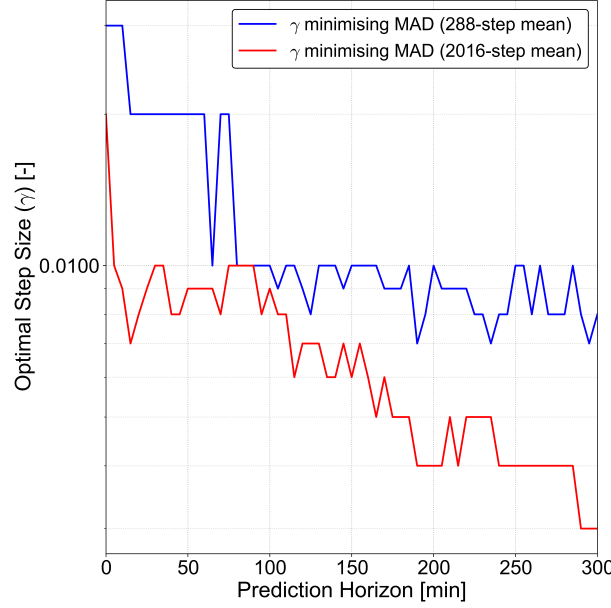


Figure C.7: Step size optimised for mean absolute deviation (MAD) for two different rolling window sizes, evaluated on the validation dataset using grid search.

To summarise, no single step size can be considered optimal, because step-size tuning is inherently a trade-off between stability and reactivity. The observed deviation of the rolling mean from the target depends on the chosen window size and, consequently, on the user's preferred stability horizon. In addition, model performance depends on the target error rate α , which varies with the operational context. Low MAD and MWIS values do not capture every aspect of step-size tuning. A model may perform well on average while still failing during short periods of distribution shift. In such cases, a higher step size may respond more effectively to rapid changes, even if this is not fully reflected in the aggregate metrics. Step-size tuning therefore involves both metric optimisation and operational judgement. The latter is difficult to capture numerically. These findings emphasise the need to focus on advanced adaptive conformal prediction methods, such as Aggregated ACI and Dynamically-tuned ACI, which omit the need to manually tune the step size γ , in future research.

C.2. Effect of Information Lag on ACI Performance

As described in Part I, the ACI framework accounts for information lag when updating the target error rate α_{t+1} . This lag increases with the prediction horizon. For example, for a 200-minute prediction horizon, it totals 220 minutes, comprising both the prediction horizon and the 20-minute time window duration. Information lag affects model performance, as the ACI framework compares its predictions with the observations (which come delayed) to dynamically adjust the prediction intervals. As shown in Fig. C.8, the idealised ACI model without information lag achieves a slightly better empirical coverage across the evaluated prediction horizons, particularly at longer prediction horizons. With a step size of 0.0008 (as used in Part I), the MAD decreases from 0.74 to 0.71 percentage points between the operational ACI model and the idealised model without information lag. This increased performance is mainly obtained at higher prediction horizons, as shown in Fig. C.9. The MWIS similarly improves from 5.6926 to 5.6857.

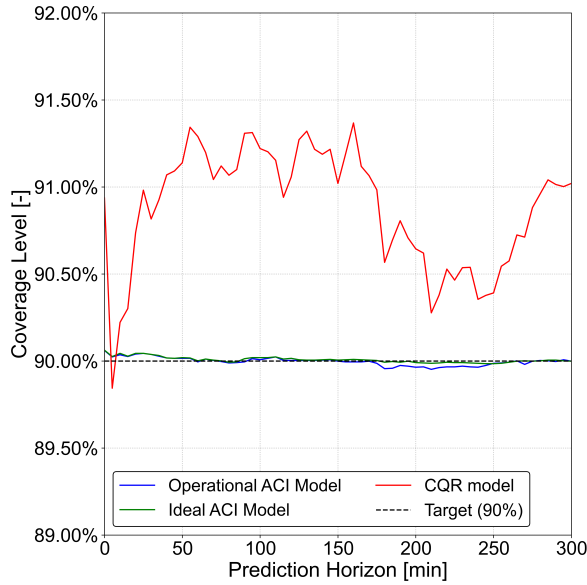


Figure C.8: Coverage level for the CQR, operational ACI ($\gamma = 0.0008$), and ideal ($\gamma = 0.0008$) ACI models evaluated on the test dataset.

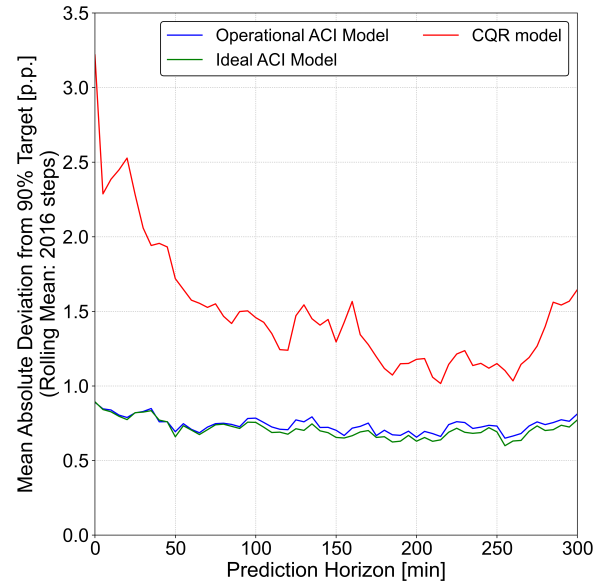


Figure C.9: MAD in percentage points (p.p.) for the CQR, operational ACI, and ideal ACI models evaluated on the test dataset.

Fig. C.10 illustrates the rolling empirical coverage for a 300-minute prediction horizon for the CQR, operational ACI, and ideal (no lag) ACI model. The horizontal axis has been limited to a fraction of the total test dataset to visually enlarge the differences. The idealised model exhibits slightly lower deviations from the target coverage level because it immediately reacts to distribution shifts, whereas the operational model is constrained by delayed feedback on its performance and will therefore adjust to these shifts with a delay. For example, predictions made during periods of rising uncertainty not captured by the machine learning model initially receive too narrow intervals in the operational model.

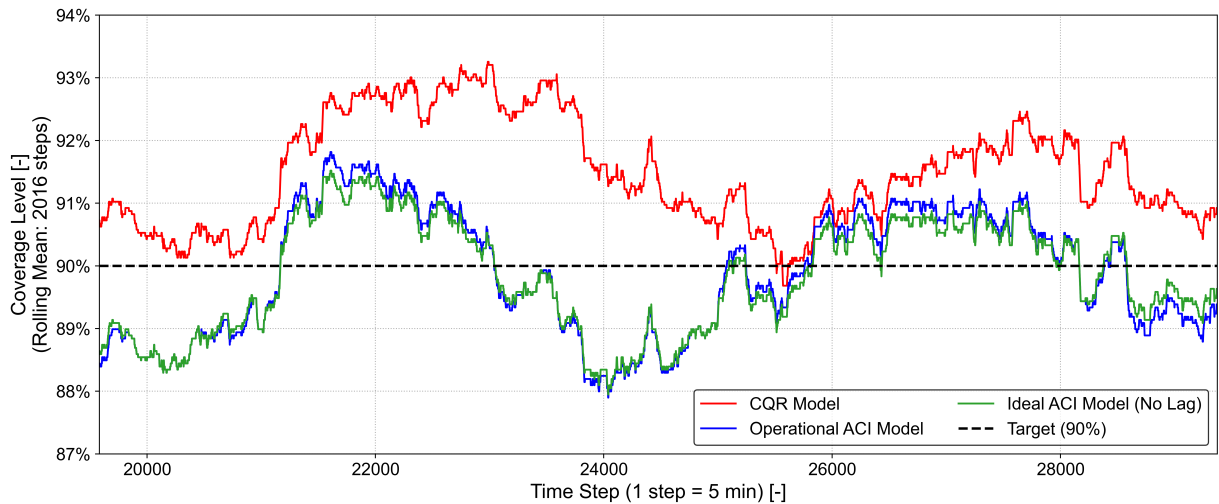
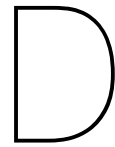


Figure C.10: Coverage level of various models ($\gamma = 0.0008$) for 300-minute prediction horizon with a 2016-step rolling average (1 week) evaluated on the test dataset.

Although the information lag is an inherent limitation of the ACI framework and cannot be removed, these results contribute to understanding the ACI model performance and in that way provide guidance while tuning the step size. For example, the results suggest that larger step sizes may be suitable for shorter

prediction horizons, where information lag is smaller and rapid adaptation is beneficial. Conversely, smaller step sizes may be preferable for longer horizons, where increased uncertainty may already have resolved by the time feedback becomes available. This is consistent with the findings in Section C.1, which show downward trends in the optimal step size as the prediction horizon increases.



Synthetic Inbound Demand

This chapter discusses the synthetic inbound demand. First, Section D.1 explains the difference between the synthetic inbound demand used to train and evaluate the model and a synthetic inbound demand for operational use. Furthermore, this section evaluates the impact of this difference on the prediction intervals. Section D.2 explains the cause of bunch forming visible in the synthetic inbound demand when LVNL applies ATFM regulations.

D.1. Difference Between Offline and Online Synthetic Inbound Demand

Part I describes how the synthetic inbound demand is calculated. In certain cases, this calculation relies on the true flight time, which requires hindsight knowledge. This is acceptable during model training, where such information is available. The synthetic inbound demand is subsequently used in the online phase to track prediction interval performance and to support convergence to the nominal target coverage level using the ACI approach.

In LVNL's operational setting, however, the true flight time is often unavailable. This occurs when a flight arrives after the interval analysis has been conducted. Consequently, the synthetic inbound demand as defined in Part I, which relies on true flight times, cannot be applied in practice. With a minor adjustment, the definition can be made operationally viable. Specifically, for any flight that has received a restrictive ATFM delay originating from LVNL by the time a prediction interval is analysed, the estimated EHAA Cross-Border Area (CBA) entry time from the most recent message prior to the demand prediction for that time window is used as a proxy for the true EHAA CBA entry time.

An analysis is performed to assess the impact of this assumption on the prediction interval width. The model is trained using the offline synthetic demand and evaluated (in the ACI framework) against both the offline definition and the operationally applicable synthetic demand. Fig. D.1 shows the distribution of differences in synthetic demand across three prediction horizons. The vast majority of differences are (close to) zero. Fig. D.2 illustrates the effect of these differences on the dynamic prediction interval width. Of the absolute differences between the adaptively calculated prediction interval widths obtained with the two synthetic demand definitions, 50% are below 0.21, and 99% are smaller than 1.17, indicating that the operational adjustment would have a minimal effect on interval width.

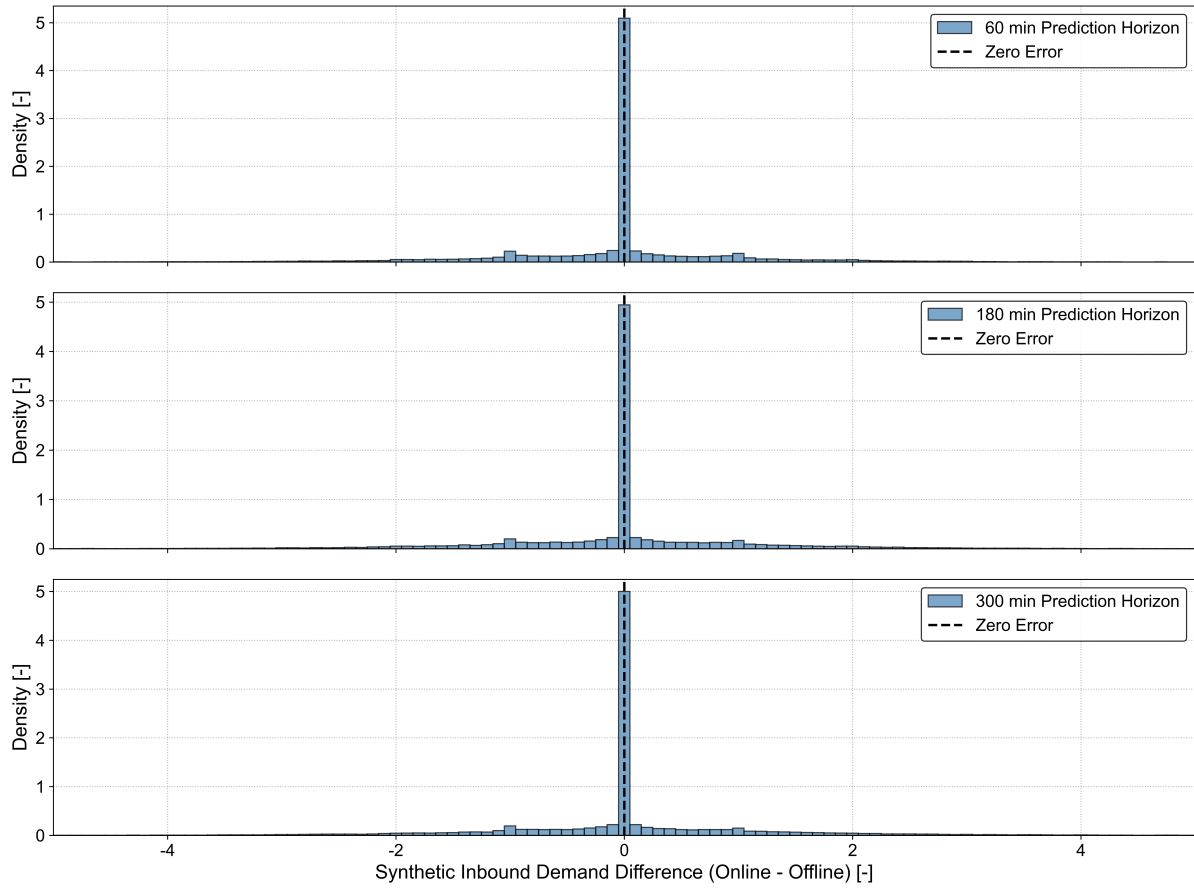


Figure D.1: Distribution of differences between synthetic inbound demand used for training (offline) and operational use (online).

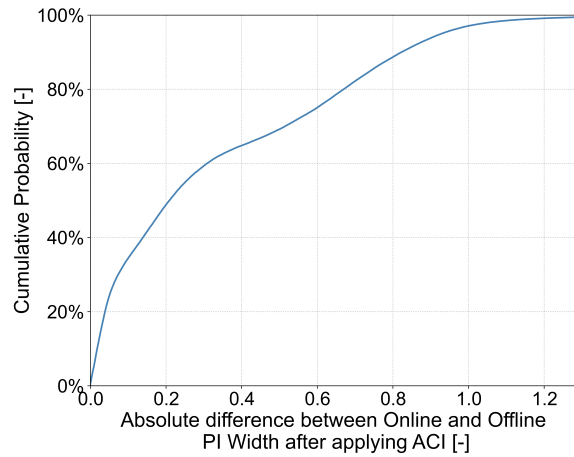


Figure D.2: Distribution of absolute differences between prediction interval widths after applying the ACI framework using the online and offline synthetic inbound demand definitions.

D.2. Bunch Forming Due to LVNL's ATFM Regulations

When LVNL imposes ATFM regulations through EUROCONTROL's Network Manager (NM), bunch forming can occur, as shown in Fig. D.3. Regulations contain fixed start and end times, sometimes causing the

system to delay many aircraft until the moment the capacity constraint ends. This produces demand surges immediately after the regulation expires. If the regulation is subsequently extended, these aircraft are redistributed and the wave dissipates. As ATFM regulations are applied a few hours in advance, this phenomenon mainly occurs at longer prediction horizons.

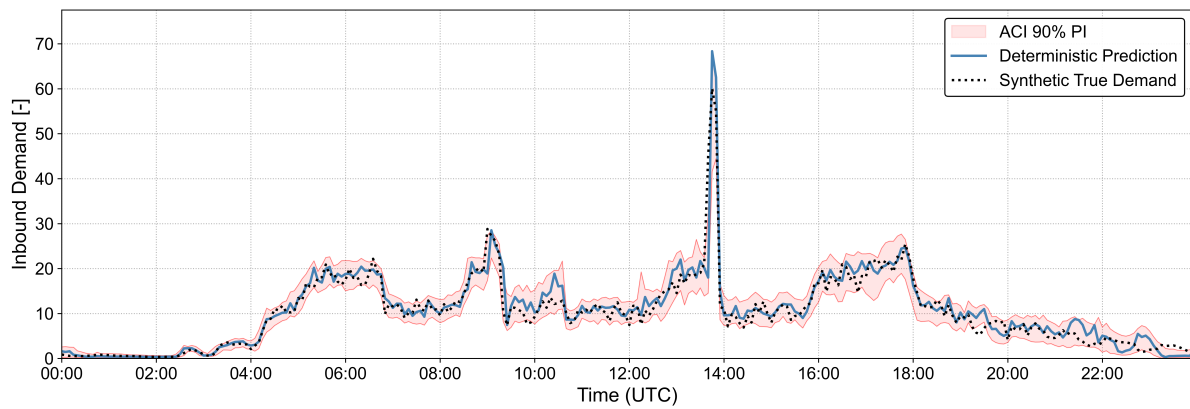
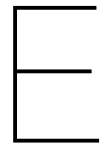


Figure D.3: Example of bunch forming due to ATFM regulations imposed by LVNL on 2025-10-05 with a 300-minute prediction horizon.

The method used to calculate the synthetic inbound demand introduces a subtle discrepancy when bunch forming occurs. To calculate the synthetic demand (which models the unconstrained demand without future LVNL ATFM regulations), the true flight time is used. The DST, however, uses the expected EHAA CBA entry time of the most recent message for its deterministic demand prediction. Because the true flight time can deviate slightly from the expected flight time, the sharp demand increase associated with bunch forming in the synthetic demand can begin and end slightly earlier and later than predicted by the DST. The operationally applicable synthetic demand does not exhibit this discrepancy, because it uses the expected EHAA CBA entry time instead of the true flight time for flights subject to LVNL ATFM regulations.



Effect of Target Error Rate on Prediction Interval Width

The scientific paper in Part I evaluates a 90% prediction interval to analyse the performance of the probabilistic framework. However, no universally optimal target error rate exists, as the appropriate value depends on operational preferences. Therefore, the effect of the target error rate α on prediction interval width is investigated. Fig. E.1 illustrates the results. The target error rate determines the nominal target coverage level, given by $1 - \alpha$. As mentioned in Part I, the lower and upper bounds of the interval correspond to the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles, respectively. For example, $\alpha = 0.4$ corresponds to a 60% target coverage level, bounded by the 20th and 80th percentiles. The impact of α on the interval width is relatively small at short prediction horizons, but its effect increases significantly at longer prediction horizons. Notably, the impact of the target error rate on prediction interval width is non-linear. For example, the increase in mean interval width is much smaller when the target error rate is reduced from 50% to 40% (a 10 percentage point reduction) than when it is reduced from 5% to 2% (a 3 percentage point reduction), despite the larger reduction in percentage points in the former case.

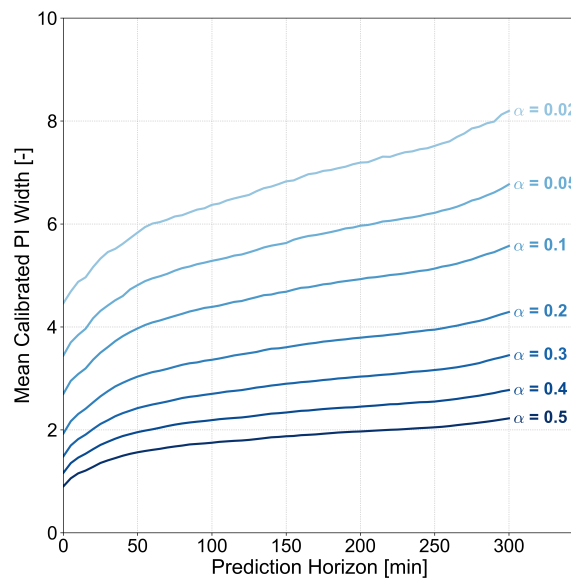


Figure E.1: Effect of target error rate on calibrated prediction interval width evaluated on the test dataset.

Performance of DST vs. Median-Quantile ML Model

While quantile regression is primarily used in this research to construct prediction intervals using lower and upper bounds, computing the median-quantile (i.e., 50th percentile) yields a deterministic point forecast that can be directly compared to the DST. Notably, this median-quantile model outperforms the DST in terms of absolute prediction error, as shown in Fig. F.1. The performance gap is negligible at short prediction horizons, but widens substantially as the prediction horizon increases for both the mean and median error metrics. The median-quantile model consistently outperforms the DST beyond a prediction horizon of approximately 30 minutes, with the median absolute error being 5-9% lower compared to the DST. The improvement is even more pronounced for the mean absolute error, where the performance gap grows from approximately 7.5% at a 60-minute prediction horizon to around 12% at a 300-minute prediction horizon.

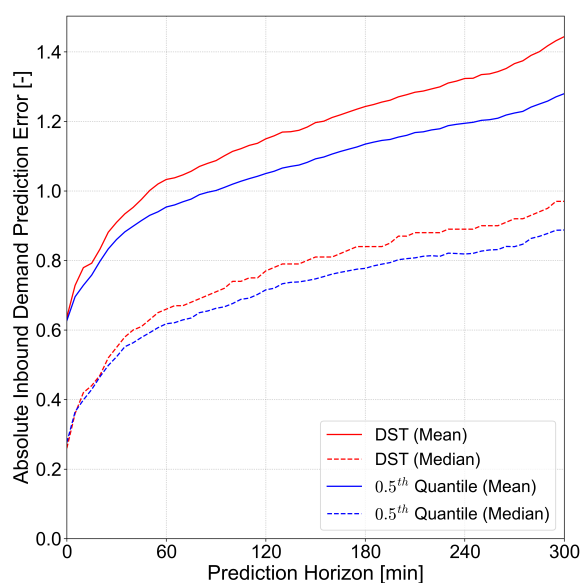


Figure F.1: Mean and median absolute inbound demand prediction error of the DST and the median (50th percentile) quantile regression model, plotted as a function of prediction horizon.

This result shows that the underlying median-quantile machine learning model can capture inbound demand dynamics more accurately than the current DST. Including the median-quantile prediction alongside the DST point forecast in the demand and capacity balancing (DCB) process could provide meaningful insights due to the superior predictive accuracy of the former, particularly at longer prediction horizons.

Part III

Literature Review and Research Proposal

Literature Review

This chapter establishes the theoretical foundation for the research by reviewing the current state of academic knowledge. It begins by defining the operational context of the problem in Section 2.1. Subsequently, previous work on aggregate inbound demand predictions is evaluated in Section 2.2. Section 2.3 discusses methods for quantifying uncertainty in black-box model predictions. Section 2.4 investigates techniques for time series analysis and forecasting, with a focus on conformalised inference. Lastly, Section 2.5 concludes the literature review.

2.1. Background Information

For the aviation industry to uphold its reputation as a safe and efficient mode of transport over long distances, it is of paramount importance to maintain sufficient separation between aircraft. However, an efficient flow of air traffic through the airspace is desired if the airspace is in high demand. Safety is an intrinsic prerequisite for aviation operations, while efficiency is necessary to minimise delays, enhancing passenger comfort, reducing the economic burden on airlines, and limiting the environmental impact of aviation. The safe and efficient flow of air traffic through an airspace is the domain with which Air Traffic Flow Management (ATFM) is concerned. An Air Navigation Service Provider (ANSP), such as LVNL, is responsible for this task.

To ensure safe and efficient management of air traffic, ANSPs set a maximum capacity on an airspace. This is a dynamic measure which is dependent on factors such as runways in use, weather conditions, and the number of air traffic controllers available, among others. The demand, on the other hand, is determined by the number of aircraft wanting to make use of the airspace. It can be defined in various ways. One is the number of aircraft occupying a specific airspace/sector at any given moment in time. Alternatively, demand could be quantified by the number of aircraft flying through the boundary of an airspace within a given time window, or the number of intervals between aircraft flying through the airspace. These are all quite straightforward methods. However, another option is to also take into account the air traffic controller workload required, making the demand calculation more complex. One could consider the planned trajectories of the aircraft passing through the airspace. Two (or even more) aircraft with crossing pathways demand a higher workload from the air traffic controller (as intervention could be needed to maintain a safe operation) than aircraft with diverging paths which will maintain a safe separation without interference from the air traffic controller, for example.

ANSPs can address a predicted demand overload in two ways: by raising the capacity or lowering the demand. The capacity can be increased by, for example, putting additional runways into use, redistributing sectors such that they become more manageable for air traffic controllers, and by having aircraft fly holding patterns. On the other hand, the demand can be lowered by applying ground delay to aircraft that are yet to take off. This process of optimising the airspace demand and capacity is called demand and capacity balancing (DCB), and is achieved using the measures mentioned above. DCB aims to lower the demand and/or raise the capacity when demand peaks are likely to exceed the capacity, and lower the capacity and/or raise the demand when the capacity exceeds the projected demand. An illustration of DCB is provided by Fig. 2.1 (EUROCONTROL, 2019). To lower demand peaks, ground delays can only be applied up to 2 to 3 hours in advance. At too short notice, all aircraft will have already departed, making ground delays impossible. For that reason, interventions to lower a demand peak are preferably taken well in advance.

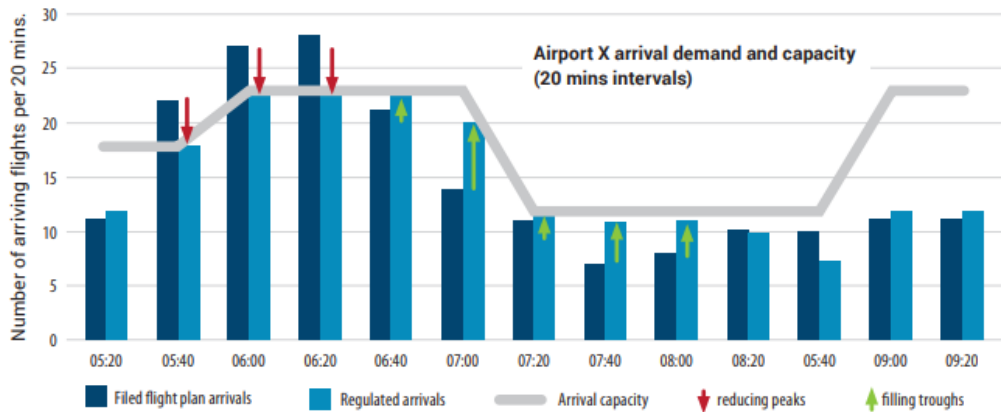


Figure 2.1: Graphical explanation of demand and capacity balancing (DCB) (EUROCONTROL, 2019).

At LVNL, DCB is performed using the Decision Support Tool (DST). One of the many features of the DST for the Area Control Centre (ACC) is forecasting both the inbound capacity LVNL can handle, based on the landing capacity at Schiphol, and the demand it expects, which is based on the number of aircraft crossing the EHAA Cross-Border Area (CBA) inbound to Schiphol Airport (where EHAA is the International Civil Aviation Organization (ICAO) code for the Amsterdam FIR). LVNL is the ANSP responsible for the EHAA below flight level 245. This thesis focuses on the demand side. The DST makes the demand forecasts using ETFMS Flight Data (EFD) obtained from EUROCONTROL's Network Manager. Functionally, the DST examines the flight plans of each flight, takes the entry time of the EHAA CBA, and then sums the number of intervals between aircraft estimated to cross the CBA within a given time window (typically 20 minutes). Thus, the DST effectively calculates the arrival rate of aircraft crossing the EHAA CBA inbound to Schiphol. Fig. 2.2 shows a snapshot of the DST, showing both the demand and capacity. Around 11:00 and 13:00 two instances can be observed where the demand is expected to exceed the capacity.

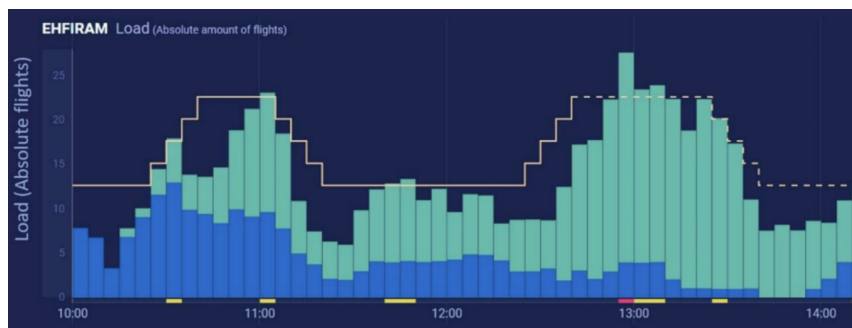


Figure 2.2: Example of EHAM inbound demand prediction by the DST (Vegter et al., 2023).

For the purpose of smoothing, the DST calculates the number of expected intervals between aircraft arriving in a time window rather than the number of aircraft itself. Consider a scenario where one aircraft is expected to arrive (i.e., cross the CBA) at 05:59:59, one at 06:02:30, and another one at 06:05:01. Within the time frame 06:00:00 - 06:04:59, the aircraft count is 1. However, the demand, and therefore the workload for an air traffic controller, is almost identical to a scenario where the first aircraft arrived a few seconds later and the third aircraft a few seconds earlier. If that were the case, the count would be 3. To more accurately represent the workload created by the demand of inbound aircraft and smooth the demand predictions, the DST takes the number of intervals between arriving aircraft into account. This is displayed in Fig. 2.3. The DST accounts for intervals overlapping two time windows by counting the share of the interval that falls in the time window of interest.

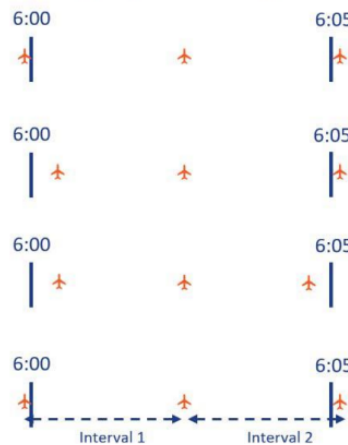


Figure 2.3: Illustration showing the effect of intervals on inbound demand calculation (Dekker, 2023).

It should be noted, however, that the DST has many more functionalities. For Tower/Approach, the DST uses a different arrival time than the CBA entry time it uses for ACC, for example. It also calculates the expected delay and demand for different airspace sectors, among other metrics. However, this research focuses on the DST as used by ACC for inbound air traffic.

It is crucial to be able to distinguish delays imposed by LVNL from the other ATFM regulations. This can be done using the ATFM regulation codes mentioned in the EFD information. ATFM regulations for aircraft inbound to Amsterdam Airport Schiphol (EHAM) start with *EHAMAZzx*, *EHATPzx*, *EHSUGzx*, *EHRIVzx*, where *zx* stands for the day of the month, and *x* stands for the time of day (e.g., E for early mornings, N for night) (Luchtverkeersleiding Nederland (LVNL), n.d.).¹ Besides information about the ATFM regulation(s) imposed (if any), the EFD messages contain information regarding the aircraft's flight plan. This data can be used to extract expected and actual take-off times and EHAA CBA entry times.

If the DST projects that the demand will exceed the capacity, the FMPC can decide to intervene. LVNL cannot control the speed or trajectory of inbound aircraft not yet under its control. Thus, the main method to lower the demand of inbound aircraft is issuing regulations to EUROCONTROL's Network Manager (NM) (if done at least 2 to 3 hours in advance), which the Network Manager (NM) then translates into delays for aircraft that are yet to depart using the Enhanced Tactical Flow Management System (ETFMS). The ETFMS is the tool used by the NM to optimally assign these new delays. The delays are issued by imposing a Calculated Take-Off Time (CTOT), requiring aircraft to depart in a time window of 5 minutes before and up to 10 minutes after the CTOT. Just like a capacity breach, unnecessary ground delays are not desirable either, underscoring the importance of accurate Schiphol inbound demand predictions made by the DST. Currently, however, it is not clear how accurate the demand predictions made by the DST are. For that reason, the FMPC deliberately lowers the capacity graph when using the DST to account for the uncertainty in its demand predictions.¹ It would be highly beneficial to obtain insight into the (in)accuracy of the predictions made by the DST and add probabilistic information to the deterministic forecasts it currently makes. It would allow the FMPC to make more informed decisions and, by doing so, potentially lower the amount of unnecessary ATFM delays imposed and reduce the number of missed capacity breaches. That is the topic this research aims to investigate.

2.2. Previous Work on Aggregate Inbound Demand Predictions

As described above, effective DCB is critical at capacity-constrained airports such as Amsterdam Airport Schiphol (EHAM). When aggregate inbound demand approaches or even exceeds the capacity limits of the Air Navigation Service Provider (ANSP), deterministic schedules become brittle. To effectively manage the air traffic flow, it is important to accurately predict the demand of inbound traffic, and ideally adopt a probabilistic approach.

Relatively few papers discuss the topic of aggregate traffic demand from a top-down perspective. Gilbo and Smith (2007) focused on linear regression to obtain more accurate aggregate demand predictions, taking

¹Follows from a discussion with a Flow Management Position Controller (FMPC) at LVNL

into account the uncertainty of the prediction model. This was achieved by smoothing the predicted demand counts using the time intervals preceding and following the interval of interest. They took a top-down approach to this problem, focusing not on predictions for individual flights, but rather on deterministic aggregate demand predictions. A similar approach was taken by Wanke et al. (2003) and Wanke et al. (2005). The former identified the prediction horizon, demand level, and share of active and proposed flights as important explanatory variables. However, both studies left out time windows with high demand to avoid the self-defeating prophecy. If the demand is projected to exceed the maximum capacity, ANSPs will impose delays to lower the demand peaks. However, this alters the observed demand, and the unconstrained demand is never observed. This problem was tackled by leaving out time windows with high demand, as they are most susceptible to this phenomenon.

The top-down approach mentioned above is not very common in ATFM demand forecasting. A bottom-up approach is more common, relying heavily on the analysis of individual flight trajectories and arrival time estimates (Ruskamp, 2025). Meyn (2002) took a probabilistic approach to aggregate demand forecasting. He used Monte-Carlo simulation using artificial estimated arrival time distributions to shift from deterministic to probabilistic outcomes, and was able to reduce the standard deviation in error of the aggregate demand forecast by 15 to 20%. Following this work, Gilbo and Smith (2009), also taking a bottom-up approach to demand forecasting, observed that arrival time “predictions for active flights are significantly more accurate than for proposed flights, with a much narrower range of uncertainty for active flights”. This observation is supported by the literature discussed below, and it is in line with one of the three explanatory variables used by Wanke et al. (2003) (discussed above). Their work also mentions that the difference in accuracy between shorter (1-2 hours) and longer (2-3 hours) look-ahead times (LAT) is small. Using a probabilistic approach, Gilbo and Smith (2009) sometimes arrived at prediction intervals not encompassing the deterministic prediction, demonstrating that there is a large difference between the deterministic and probabilistic approaches. Chatterji et al. (2004) created prediction bounds using a bottom-up approach. More recently, Gonze et al. (2018) also took a bottom-up approach to arrive at probability curves for sector count, specifically mentioning that their methodology computes “the probability that a flight is in a sector at a given time from the distribution of its probable trajectories” to arrive at the probability curve.

The inbound demand is dependent on the estimated time of arrival (ETA) of inbound aircraft, which is in turn influenced by many factors such as weather, airspace congestion, and above all take-off time variability. Dalmau et al. (2025) identify take-off time variability as “the most significant contributor” making it difficult to accurately predict arrival times. Additionally, approximately 46% of all delay minutes of European air transport are caused by reactionary delay, also called knock-on delay (EUROCONTROL, 2025b). Extensive research has been conducted on estimated time of arrival (ETA) predictions and their variability. For example, studies by Dalmau et al. (2025), Ayhan et al. (2018), Glina et al. (2012), Kern et al. (2015), and Vos et al. (2024) all use machine learning techniques to enhance the ETA predictions of individual flights. Shone et al. (2021) mention that “the question of how to make dynamic updates to predicted aircraft arrival times in response to the latest available information is an important area of study in its own right”. This was a problem explored by, for example, Tielrooij et al. (2015), who attempted to predict the uncertainty of arrival times using historical prediction accuracy and flight information updates.

In summary, a lot of research has been conducted on improving arrival time estimates and thereby improving the aggregate inbound demand calculations, using both deterministic and probabilistic approaches. Using arrival time estimates of individual flights constitutes a bottom-up approach, which has been the dominant methodology in the past as mentioned above. However, significantly fewer studies have been directed toward a top-down framework, viewing the demand as a single entity, and even fewer studies focus on probabilistic aggregate inbound demand predictions using novel machine learning techniques.

2.3. Uncertainty Quantification with a Wrapper Model

To achieve probabilistic inbound demand forecasts, it is necessary to quantify the uncertainty inherent in the predictions. Crucially, the objective of this research is not to improve the point-accuracy of the existing aggregate demand forecasts, but to add probabilistic metrics to the current deterministic output. For that reason, it is pragmatic to treat the model (i.e., the DST) as a black-box model. A black-box model is a model that provides a certain output given an input, but of which the internal workings or logic are unknown. While the internal logic of the DST is known (as detailed in Section 2.1), for the purpose of uncertainty quantification it is treated as a black-box predictor. Thus, the ANSPs aggregate inbound

demand forecasting tool (which for LVNL is the DST) will be deemed a black box, and only the uncertainty of the predictions will be analysed. However, merely analysing how accurate the deterministic predictions of such a tool are, does not directly translate to any benefit in real-world applications. Therefore, the analysis will then be translated to probabilistic forecasting using prediction intervals to make it useful in an operational context. This research employs a residual error modelling approach, where a secondary machine learning model predicts the uncertainty of the DST based on exogenous variables that correlate with prediction uncertainty.

Brando et al. (2020) discuss exactly the type of problem at hand; a black-box algorithm outputting a point-wise estimate. The paper presents a method to estimate the residual errors of the black-box model. It suggests creating a “wrapper model that complements and maintains the original prediction with new information: the uncertainty score” (Brando et al., 2020). It discusses multiple options for a wrapper model and mentions that heteroscedastic methods are preferred. Heteroscedasticity refers to a non-constant variance in a model’s residual errors across observations/time. This is an important property as it is expected that the uncertainty of the inbound demand is dependent on explanatory features such as the prediction horizon. The paper applies this method to, among others, an electrical power demand dataset. Similarly, O’Sullivan et al. (2016) apply quantile regression to a black-box model predicting bus arrival times in Boston, placing emphasis on the inherent heteroscedasticity. They explicitly mention that their “approach is not intended for or capable of improving the quality of the predictions made, merely estimating the uncertainty associated” (O’Sullivan et al., 2016). The paper aims to provide prediction intervals for bus arrival times and defines a prediction interval as “a window within which we expect the next predicted value to fall with some probability” (O’Sullivan et al., 2016). Mathematically, they define a prediction interval (PI) with a confidence level of $1 - \alpha$ by Eq. (2.1) and Eq. (2.2), in which $L(x)$ and $U(x)$ denote the lower and upper bounds of the prediction interval, respectively. α represents the target error rate and Y_{n+1} indicates the true future value.

$$PI = [L(x), U(x)] \quad (2.1)$$

$$P(L(x_{n+1}) \leq Y_{n+1} \leq U(x_{n+1})) = 1 - \alpha \quad (2.2)$$

Their work on prediction intervals for bus arrival times was inspired by the paper by Pereira et al. (2014), which estimated the uncertainty bounds of a real-time traffic prediction system. Importantly, they mention that the methodology can easily be applied to other regression tasks, requiring only “access to the feed of predictions and observations as well as any other variable that is believed to correlate to the prediction error (e.g., traffic conditions and weather status)” (Pereira et al., 2014).

Thus, to model the uncertainty, one does not need to know exactly how the deterministic predictions were made. All that is required are the predictions, the corresponding true observations, and variables that are believed to correlate with the uncertainty of the prediction. Consequently, this approach is well-suited for probabilistic inbound demand forecasting. As this information is measured over time, it can be treated as a time series. Extensive research has already been conducted on suitable models for time series analysis and forecasting. A brief discussion of the existing methods is provided next in Section 2.4.

2.4. Time Series Analysis and Forecasting

Time series can be explained as “a sequence of observations taken sequentially in time” (Box et al., 2016). Because of their broad applicability, they are found in many disciplines. The goal of time series analysis and forecasting is to deepen one’s understanding of patterns over time and predict the future. Inbound demand can be viewed as a time series, as it represents the number of aircraft arriving within a certain time window at discrete, consecutive points in time. When viewed in a graph, one will get a sense of the development of the inbound demand. Various time series forecasting methods have already been applied within the field of aviation. Do et al. (2020) applied both a Long Short-Term Memory model, which is a type of neural network, and Seasonal Autoregressive Integrated Moving Average, which constitutes a statistical method, to forecasting the number of passengers at Incheon International Airport, providing deterministic results. Neural networks have also been used in predicting network delay, for example by Sanaei et al. (2021) and Zhao et al. (2024). Although these models are deterministic and do not focus on inbound demand prediction, these studies indicate that time series analysis is not a foreign concept within the aviation industry.

Choosing the correct model for time series forecasting is of crucial importance, and extensive research has been conducted on the comparison of various types of models. The Makridakis forecasting competitions

(M-competitions) are contests in which participants are tasked with analysing and forecasting time series as accurately as possible. They serve as a benchmark for the performance of various forecasting methods. In 2020, Makridakis et al. (2020) found, based on the M4 competition, that a hybrid approach, combining statistical methods with machine learning, outperformed pure machine learning and pure statistical methods in time series forecasting. Two years later, an investigation into the results of the M5 Accuracy competition, which challenged participants to forecast unit sales of Walmart retail goods, found that the top 50 methods were all machine learning methods and thus outperformed statistical methods. The paper specifically mentions that LightGBM displayed “several advantages over other ML alternatives in forecasting tasks, ... as it allows the effective handling of multiple features ..., is fast to compute compared to typical gradient boosting (GBM) implementations, does not depend on data pre-processing and transformations, and requires the optimization of only a relatively small number of parameters” (Makridakis et al., 2022). One might question why machine learning techniques outperformed other methods in 2022, when in 2020 hybrid methods were superior. Both Makridakis et al. (2022) and Januschowski et al. (2022) attributed this difference to the type of time series used. The M5 competition featured highly correlated time series which facilitated the creation of global models, allowing for cross-learning and the inclusion of explanatory variables. Conversely, the M4 competition featured different types of time series, necessitating local models where cross-series learning was less effective. Thus, the literature suggests that machine learning techniques outperform statistical or hybrid methods when they can efficiently leverage cross-series information and exogenous variables, whereas hybrid or statistical techniques remain superior for isolated, local time series problems.

Grinsztajn et al. (2022) showed that one reason why tree-based models outperform deep learning methods is that neural networks are biased towards smooth functions, whereas tabular data (like the data processed by the DST) can have jagged decision boundaries. As a consequence, tree-based models are better at fitting irregular functions. Furthermore, Elsayed et al. (2021) found that Gradient Boosting Regression Tree models outperformed eight state-of-the-art deep neural networks discussed at top conferences in recent years when tasked with time series forecasting of nine different datasets.

Selecting an appropriate forecasting model requires evaluating its ability to generate prediction intervals. This is a feature that many standard machine learning models lack natively. Beyond basic interval generation, the chosen methodology must address both heteroscedasticity and non-exchangeability. Because it is expected that aircraft inbound demand processes exhibit varying uncertainty over time, the model must be able to capture heteroscedasticity. Furthermore, time-series data such as flight arrivals are expected to be non-exchangeable, meaning that the joint probability distribution of the sequence (i.e., the probability of observing a particular sequence of observations) depends on the order of observations. A sequence is exchangeable only if any reordering of the observations leaves that joint probability distribution unchanged. Consider the example of a coin flip. Each toss is independent, as previous outcomes do not alter the probability of the next. Thus, a sequence of coin flips is exchangeable (e.g., $P(X_1, X_2, X_3) = P(X_2, X_1, X_3)$). It is expected that this does not hold true for inbound demand prediction errors. For instance, a delayed aircraft will initially cause an overprediction, followed by an underprediction in a subsequent time window. This suggests that the order of the data is important, violating the exchangeability condition.

Quantile Regression

Quantile regression was developed by Koenker and Bassett Jr (1978). It uses a loss function called ‘check loss’ (in contemporary literature known as ‘pinball loss’). Eq. (2.3) (Koenker & Bassett Jr, 1978) presents the pinball loss function, in which y is the actual observed value, $\hat{y}$ is the value predicted by the model, and τ is the percentile predicted (e.g., 0.5 for the median, 0.9 for the 90th percentile). The original paper used α , but the notation τ is common to avoid confusion with the error rate used in conformal prediction. This practice will also be applied in this research.

$$\rho_\tau(y, \hat{y}) := \begin{cases} \tau(y - \hat{y}) & \text{if } y - \hat{y} > 0, \\ (1 - \tau)(\hat{y} - y) & \text{otherwise} \end{cases} \quad (2.3)$$

In order to arrive at a prediction interval, two loss functions need to be optimised (i.e., minimised), one for the lower bound and one for the upper bound. For this specific application, this characteristic of optimising two separate functions is desirable, as it allows for handling skewed uncertainty well. If a symmetric interval were imposed, the width around the deterministic prediction would be equally large. By using two separate

functions, the model will be able to capture a higher risk of over- or underprediction. Also, the lower bound is limited to non-negative values, whereas the upper bound does not have a well-defined upper limit.

Standard quantile regression does not provide guaranteed coverage over a finite sample, however. This can be solved through a method called conformalised quantile regression (CQR), which combines quantile regression with conformal inference. It should be noted that other probabilistic forecasting methods exist, such as Bayesian and ensemble methods. However, conformal prediction (such as CQR) is the only approach providing marginal coverage guarantees (Dewolf et al., 2023). Ideally, a forecasting framework would provide conditional coverage guarantees. However, obtaining this is not possible without making assumptions about the underlying data distribution (Barber et al., 2021). Thus, while strict conditional coverage is unattainable, it can be approximated by evaluating the model's empirical coverage across various operational scenarios.

Conformal Prediction

Conformal prediction is concerned with creating prediction intervals with marginal (and ideally conditional) statistical coverage guarantees. Vovk et al. (2005) and Shafer and Vovk (2008) pioneered this field almost two decades ago, providing guaranteed coverage in regression problems. Currently, a widely adopted, state-of-the-art conformal inference method is Conformalized Quantile Regression (CQR) (Luo and Zhou, 2025; Rossellini et al., 2024). Romano et al. (2019) combined quantile regression with conformal prediction into CQR. The benefits of CQR over standard quantile regression are that it guarantees marginal finite-sample coverage (assuming exchangeability) because it requires calibration through a calibration dataset. Additionally, CQR handles heteroscedasticity well and only requires the assumption of exchangeability. Even if the underlying model estimating the quantiles is not very strong, the coverage guarantee still holds (it will, however, potentially lead to larger prediction intervals than if the model was very strong). As CQR is a model combining standard quantile regression with conformalised inference through calibration, it requires a model that can perform quantile regression.

Through the process of calibration, CQR is able to provide statistical guarantees for finite-sample marginal coverage. It is performed using a calibration dataset. For every point in the calibration dataset, a conformity score E_i is calculated, representing how far the true value Y_i lies from one of the two bounds. It is presented by Eq. (2.4), in which $\hat{q}_{\alpha_{lo}}(X_i)$ and $\hat{q}_{\alpha_{hi}}(X_i)$ denote the lower and upper bounds, respectively, given some input X_i . If the true value Y_i is outside the predicted interval, E_i is positive and is a measure of the distance between the true value Y_i and the nearest prediction bound. If E_i is negative, Y_i falls within the prediction interval, and E_i is a measure of the safety margin. (Romano et al., 2019)

$$E_i := \max \{ \hat{q}_{\alpha_{lo}}(X_i) - Y_i, Y_i - \hat{q}_{\alpha_{hi}}(X_i) \} \quad (2.4)$$

Subsequently, the calibrated interval (C) for a new test point X_{n+1} is defined by Eq. (2.5) (Romano et al., 2019), in which $Q_{1-\alpha}(E, I_2)$ is the calibration term where I_2 denotes the calibration dataset and E the set of conformity scores. It is important to note that in this equation α denotes the target error rate (for a 90% interval it is 0.1 for example).

$$C(X_{n+1}) = [\hat{q}_{\alpha_{lo}}(X_{n+1}) - Q_{1-\alpha}(E, I_2), \hat{q}_{\alpha_{hi}}(X_{n+1}) + Q_{1-\alpha}(E, I_2)] \quad (2.5)$$

It should be noted that other conformal prediction techniques exist, such as locally weighted conformalised prediction. However, CQR is currently deemed a common, state-of-the-art method for such applications. It is used for many real-world applications. The Washington Post used it for the 2020 US presidential election for example (Cherian & Bronner, 2020), whereas it is also used for short-term trade optimisation of renewable energy (Backus & Van den Ende, 2024).

One limitation of CQR is that it only provides these finite-sample coverage guarantees under the assumption of exchangeability, a condition that many real-world time series do not satisfy, as explained above. A valuable extension is Adaptive Conformal Inference (ACI).

Handling Non-Exchangeability - Adaptive Conformal Inference

ACI can be seen as an add-on to conformalised quantile regression. In short, ACI is applied in the online phase and dynamically updates the bounds based on live observations. The benefit of ACI is that it improves the empirical coverage behaviour of datasets that violate the exchangeability condition. In the seminal paper, Gibbs and Candès (2021) wrote “our adaptive approach provably achieves the desired

coverage frequency over long-time intervals irrespective of the true data-generating process”, unlike other conformal prediction methods (like CQR) which rely on the assumption of exchangeability. This is achieved by constantly adapting a calibration parameter, ensuring the empirical coverage level converges to the nominal target in the long run.

As mentioned above, real time series often violate the property of exchangeability. The biggest benefit of ACI is that it does not require this exchangeability property and works well in heteroscedastic situations. The coverage of ACI might oscillate locally, but in the long run it oscillates around the target coverage level. Other conformal prediction techniques run the risk of drifting from the target coverage level if real-world conditions change from those that the model was trained on (i.e., distribution shift) and are only valid if the time series possesses exchangeability.

The mathematical reasoning driving the ACI algorithm is presented by Eq. (2.6) and Eq. (2.7) (Gibbs & Candès, 2021). Eq. (2.6) describes the error term. If the observed value falls within the prediction interval, the error term err_t is zero, and 1 if it falls outside. Using this score, Eq. (2.7) adjusts the error rate. If the observed value falls within the prediction interval, the updated target error rate α_{t+1} will be larger than α_t , narrowing the prediction interval. If $\text{err}_t = 1$, α_{t+1} will be reduced, widening the prediction interval. γ denotes the step size or learning rate. Choosing this parameter is a trade-off between reactivity and stability. A larger step size will make the model more reactive, with the risk of becoming erratic. On the other hand, a low learning rate makes the prediction intervals very stable, but less reactive to distribution shifts. The optimal step size will have to be determined iteratively.

$$\text{err}_t := \begin{cases} 1, & \text{if } Y_t \text{ falls outside prediction interval,} \\ 0, & \text{otherwise,} \end{cases} \quad (2.6)$$

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \text{err}_t) \quad (2.7)$$

Other techniques suitable for conformal inference under non-exchangeable conditions exist. These are, for example, Ensemble Batch Prediction Intervals (EnbPI) and Non-Exchangeable Conformal Prediction (NexCP). Standard CQR does not assign different weights based on the age of data. By contrast, NexCP assumes that a recent data point is more important than an older one, and thus assigns higher weights to more recent data points. In addition, special ACI versions exist, like Aggregated ACI (AgACI) and Dynamically-tuned ACI (DtACI), eliminating the need to tune the step size γ (Gibbs & Candès, 2024; Zaffran et al., 2022).

Despite its short existence, ACI has already seen widespread use in probabilistic time series forecasting. For example, Gibbs and Candès (2021) used it for the 2020 US presidential election (on top of CQR), whereas Suresh et al. (2025) used it to forecast the electric power load demand of the Indian state of Tamil Nadu. The former is presented graphically in Fig. 2.4, and the latter by Fig. 2.5.

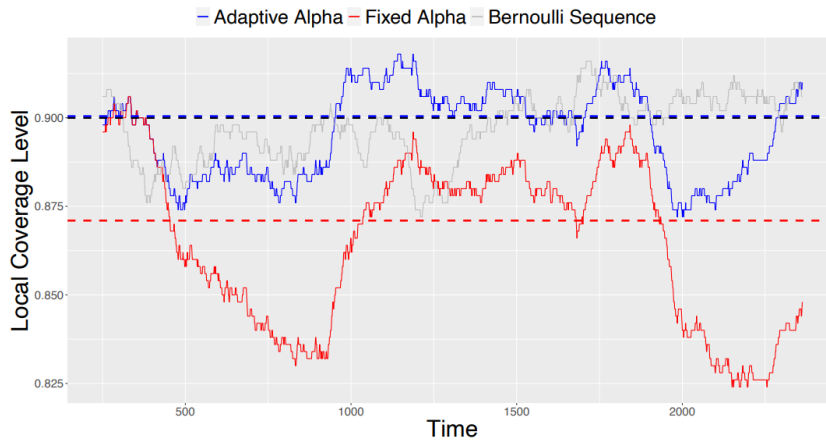


Figure 2.4: ACI applied to the 2020 US presidential election for county-level predictions (Gibbs & Candès, 2021).

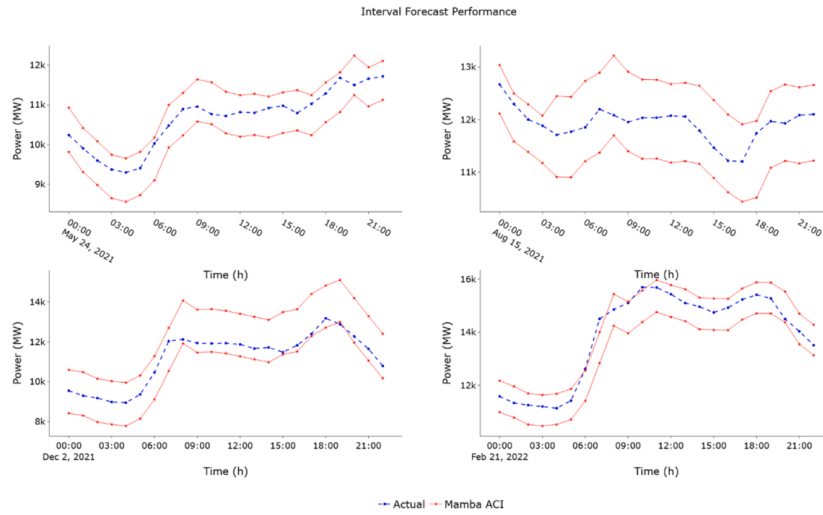


Figure 2.5: Electric power load demand for the Indian state of Tamil Nadu (Suresh et al., 2025).

2.5. Conclusion

As detailed in this literature review, existing ATFM inbound demand prediction methods predominantly rely on bottom-up approaches that aggregate individual flight uncertainty. Conversely, top-down probabilistic methodologies remain under-explored. Consequently, this research proposes a top-down framework to address this gap by using the deterministic predictions from the DST and modelling the residual errors via a secondary quantile regression model. The goal is explicitly not to improve the deterministic predictions made by the DST, but rather to augment them with probabilistic information through prediction intervals.

To ensure statistical rigour under non-exchangeable conditions, Conformalized Quantile Regression (CQR) will be integrated with Adaptive Conformal Inference (ACI). While these are established techniques in other domains, this literature review reveals that such a top-down, probabilistic architecture has not been applied to aircraft inbound demand forecasting before, to the best of the author's knowledge. Therefore, the primary contribution of this research lies in validating the efficacy of these robust uncertainty quantification techniques within the context of air traffic flow management.

Research Objective

This chapter defines the primary research objective in Section 3.1. Thereafter, Section 3.2 discusses the question central to the thesis, followed by Section 3.3, which details the scope of the research.

3.1. Research Objective

The research aims to provide insight into the uncertainty surrounding inbound demand predictions made by an ANSP. This research takes LVNL as a case study, but the concept should be applicable to other ANSPs as well. Synthesising the findings from the literature review, the research objective is defined as follows:

To contribute to the enhancement of demand and capacity balancing by applying a novel probabilistic forecasting approach to inbound air traffic demand.

3.2. Research Question

The following research question is proposed:

How can deterministic inbound air traffic demand forecasts be complemented by probabilistic forecasts using a top-down adaptive conformal prediction framework?

To answer the research question mentioned above, several subquestions should be addressed.

1. Which framework is suitable for a top-down approach to inbound demand uncertainty modelling?
2. How can the interval width be minimised while still maintaining statistical coverage?
3. How can the target prediction interval coverage be maintained under distribution shifts?

3.3. Scope

- **Definition of Arrival Time**

The DST has many functionalities; ACC uses a different definition of arrival time from tower/approach (TWR/APP). ACC defines the arrival time as the entry time of the CBA. After an aircraft has entered the CBA, the arrival manager becomes relevant, as it can alter the arrival times used by TWR/APP by, for example, putting aircraft in holding patterns. This complicates the calculation of true arrival times because of a self-defeating prophecy, where the prediction influences the actual outcome. This research focuses on the DST as used by ACC, and thus takes CBA entry time as arrival time, omitting effects of the arrival manager on arrival times. This does not remove the self-defeating nature of demand forecasts completely, as the FMPC can impose ATFM delays on inbound aircraft based on the predictions made by the DST.

- **Focus on Schiphol Inbound Aircraft**

This research focuses exclusively on aircraft inbound to Schiphol. Inbound traffic to other airports and all outbound aircraft are excluded. While these flights are controlled by ACC, they are not taken into account when calculating the inbound demand predictions for ACC. This is similar to the approach LVNL uses in its DST calculations.

- **Limited Prediction Horizon**

Prediction interval width generally increases with prediction horizon. Too large prediction horizons result in intervals that are too large to provide any operational value. Conversely, an overly short prediction horizon offers little operational value as ATFM delay (i.e., assigning a CTOT) can only be applied to aircraft that have not departed yet. Although a very short prediction horizon is not useful for requesting ATFM regulations, the smallest horizon considered is 0 hours. The highest horizon analysed will be 5 hours (300 minutes).

Project Plan

This chapter details the steps necessary to achieve the goals outlined in Chapter 3. First, Section 4.1 discusses the methods used throughout the project. Thereafter, Section 4.2 describes the results the author expects to obtain from this research. Lastly, a project schedule is provided in Section 4.3. The reader should be aware that a more detailed description of the methodology and experimental setup is presented in the research paper, which can be found in Part I.

4.1. Methodology

Achieving the research objective and answering the research question, presented in Section 3.1 and Section 3.2 respectively, will require the steps detailed below. This research builds upon the literature review presented in Chapter 2.

Gathering Data

Before any data analysis can be performed, it is crucial to have sufficient data to analyse. This research uses ETFMS Flight Data (EFD). Each flight regularly sends updates about its intended and/or actual route, which is included in the EFD dataset. This data is used by LVNL's DST to predict the inbound demand. Thus, it will be necessary to gather this data and analyse it to calculate the predictions the DST would have made, together with the corresponding true values of inbound demand. For this specific research, mainly the estimated and actual entry times of the EHAA CBA are of importance, together with flight specific information like flight plan identity, airline, aircraft type, and ATFM delay imposed (if any), among others.

Once this data has been obtained, it will be necessary to analyse the data set for potential erroneous inputs. Furthermore, the self-defeating nature of high inbound demand will be addressed using the ATFM delay information provided by the EFD.

Developing the Wrapper Model

The residual errors are defined as differences between the inbound demand predictions made by the DST and actual inbound demand. The next step is to find correlations between exogenous variables and these residual errors. As a starting point, the author intends to examine the impact of the following exogenous variables on the uncertainty of demand predictions:

- Traffic volume
- Prediction horizon
- Ratio of airborne vs. pre-departure flights
- Weather
- Day of the week
- Hour of the day
- ATFM delay imposed (if any)
- Jaggedness of predictions between different time-windows

These variables serve as a starting point and it is likely that the final model will use variables not mentioned

in this list. Using these variables will result in the development of a secondary model predicting the residual errors of the black-box model (i.e., the DST inbound demand predictions), incorporating the heteroscedastic nature of the problem at hand into its predictions. The objective is to produce a 90% prediction interval, for example, which indicates that, for specified model inputs, there is a 90% probability that the actual inbound demand will lie within that range.

The exact wrapper model used will need to be determined while performing the research. However, it should be well suited as a wrapper model on top of a black-box model (i.e., the DST inbound demand predictions) and be able to perform quantile regression.

Calibration

Once quantile regression has been performed using the wrapper model, CQR is applied. This process calibrates the prediction intervals. It essentially fixes the mistakes made by the quantile regression process using a calibration dataset, producing statistically guaranteed marginal coverage. CQR works by calculating a conformity score. The conformity score is negative if the true value falls inside the prediction interval, and positive if it is outside of the interval. It then adjusts the interval width depending on the conformity score corresponding to the, for example, 90th percentile. Through this process, marginal statistical coverage is guaranteed. The mathematical equations and definitions of this process have been discussed in Section 2.4 and will therefore not be repeated here.

Coverage under Non-Exchangeable Conditions

The issue with CQR is that it assumes exchangeability (as explained in Section 2.4). However, real-world data often do not possess this property. This problem can be solved through a technique called Adaptive Conformal Inference (ACI). ACI is an addition to CQR and monitors the real-time success of the model. ACI is applied in the online phase, whereas CQR is applied beforehand while preparing (i.e., training/calibrating) the model. The goal of ACI is to enhance the coverage behaviour of the model over the long-term in non-exchangeable circumstances. The mathematical workings of ACI are discussed in the literature review in Section 2.4 and will therefore be omitted from this section. ACI adjusts the prediction interval based on past observations. If it sees that observations fall outside of the prediction interval, it widens the future prediction interval. If observations fall inside the prediction interval, it narrows the prediction interval.

In total, four datasets are required: one for training the wrapper model, one for calibration, one for tuning the model, and one for testing/evaluation purposes. Since the data is serially correlated, the dataset cannot be split randomly. Instead, the chronological order must be preserved. Keeping these four subsets strictly separate is crucial to ensure valid results; testing the model on data it has already 'seen' would yield meaningless performance metrics.

The research will rely on Python for analysing the EFD, training the model, and making forecasts. MAPIE, which is an acronym for Model Agnostic Prediction Interval Estimator, is a Python library frequently used for CQR, and could potentially be of value for this research. Besides software tools like Python, it is not expected that any specific tools are required to conduct this research.

4.2. Expected Results

It is expected that, with the required steps mentioned above in Section 4.1, a model will be created that provides prediction intervals on top of the point estimate of demand made by the DST model.

How well it will be able to capture model uncertainty will likely depend on the correlation of the residual errors of the demand predictions with exogenous variables. The extent to which this is the case is difficult to predict upfront. Making small prediction intervals is only desirable if it is statistically sound. For the model to be useful for LVNL, knowing that the uncertainty in expected demand is large might be just as useful as knowing that it is small. However, decision-making becomes easier with smaller prediction intervals. Furthermore, it is expected that this research will arrive at unforeseen results, making this research an exciting endeavour.

4.3. Project Schedule

Based on the methodology detailed in Section 4.1, the research can be divided into a few main tasks. This section addresses the order and time frame in which these tasks should be conducted.

The goal is to first analyse the accuracy of the aggregate demand predictions made by the DST used by LVNL. This will provide insight into the uncertainty of the forecasts. Thereafter, the results will be visualised. It is expected that this will take approximately 8 weeks. These steps should be finished well before the midterm review.

Once these tasks have been completed, the next task is to make probabilistic forecasts of the aggregate inbound demand of Schiphol. First, sufficient processed data is required. This is subsequently split into four datasets: a training dataset, calibration dataset, validation dataset, and a dataset for testing/evaluation. Thereafter, a model will be trained using quantile regression to obtain uncalibrated prediction intervals. This will take approximately 7 weeks. This also includes finding explanatory variables that correlate with the prediction uncertainty.

Subsequently, the prediction intervals will be calibrated through CQR and in the online phase they will be adaptively adjusted using ACI. It is expected that creating these algorithms will require approximately 5 and 4 weeks, respectively.

It is important to note that this process will likely be iterative. For example, after analysing the data in Python, it could turn out that more data are needed before the analysis is useful, or that redundant information was used during the analysis. In the former case, it will be necessary to go a step back and obtain more data. The creation of the model performing the quantile regression will likewise be highly iterative. Given the complexity of the data, developing an optimal model will likely require multiple iterations of training and tuning.

The methodology and results will be combined into a scientific article. To ensure clarity and maintain momentum, the drafting process will be iterative and conducted concurrently with the research itself. Documenting findings continuously, rather than leaving the writing for the final phase of the project, is deemed preferable. Consequently, sufficient time is allocated for writing and revision of the scientific article at various points in the project timeline.

While the preceding sections outline the core objectives, a more comprehensive project schedule is illustrated in Fig. 4.1. This Gantt chart serves as a living document, iteratively updated to reflect ongoing research progress. To ensure the milestones are achieved, the schedule will constantly be translated into daily objectives.

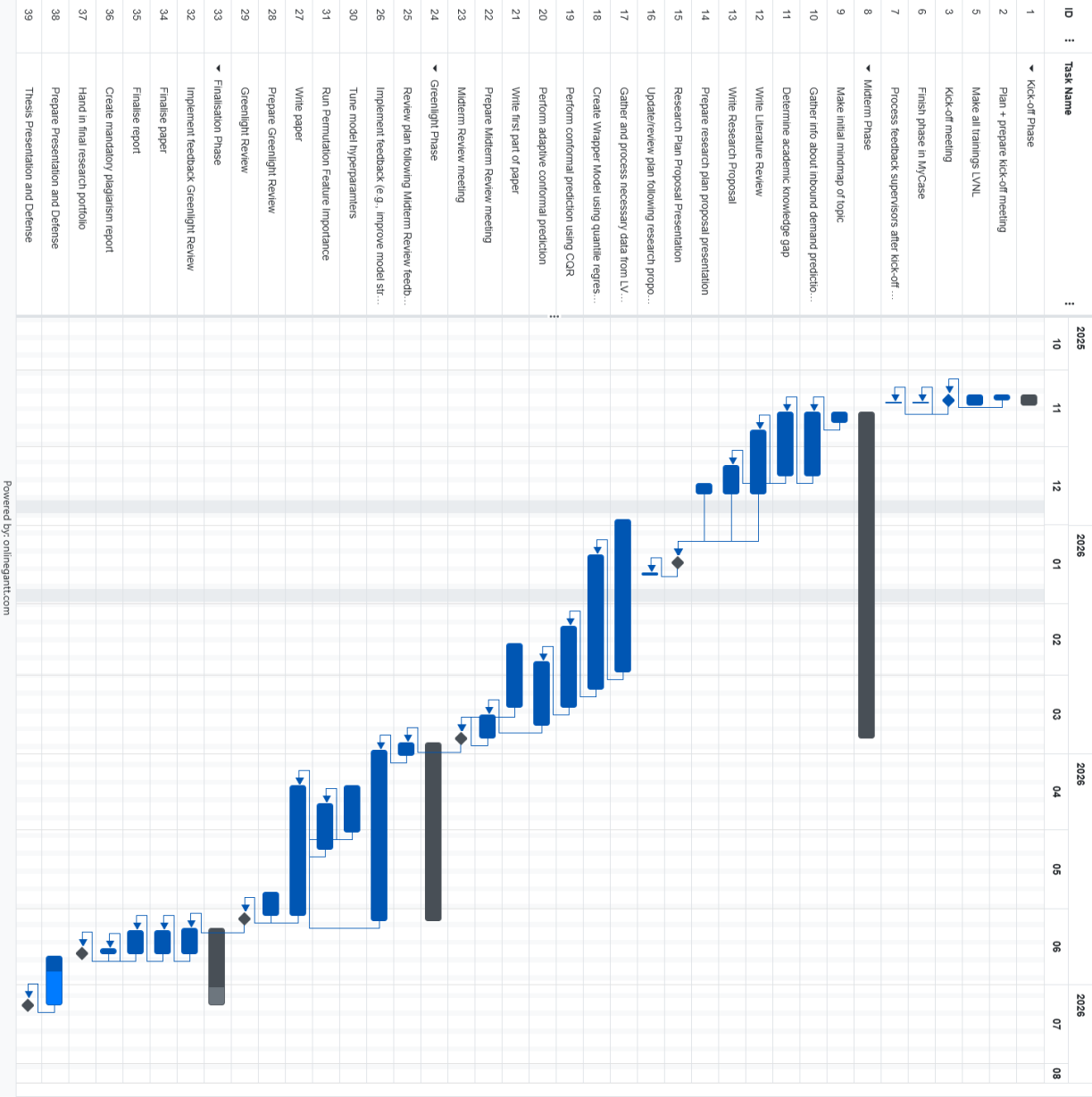


Figure 4.1: Gantt chart, last updated on 2026-06-19.

5

Conclusion

This research proposal serves as the blueprint for the subsequent study. It has established the research objectives, framed the research question, and detailed the steps required to conduct the research.

The literature study discussed the current state of research in the field of inbound demand forecasting. By investigating uncertainty quantification techniques, specifically Conformalised Quantile Regression and Adaptive Conformal Inference, a distinct knowledge gap was identified concerning the application of a top-down, probabilistic approach to inbound demand forecasting.

The methodology outlined is designed to directly address this gap. The project plan provides a structured but flexible timeline that accommodates the iterative nature of the research. Overall, the study aims to advance inbound demand forecasting by using a novel probabilistic approach.

References

- Ayhan, S., Costas, P., & Samet, H. (2018). Predicting estimated time of arrival for commercial flights. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 33–42. <https://doi.org/10.1145/3219819.3219874>
- Backus, A., & Van den Ende, I. (2024). *Probabilistic price forecasts for short-term trade optimization*. Dexter Energy. Retrieved December 15, 2025, from <https://dexterenergy.ai/news/probabilistic-price-forecasts-for-short-term-trade-optimization/>
- Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2), 455–482. <https://doi.org/10.1093/imaiai/iaaa017>
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time series analysis*. Wiley.
- Brando, A., Torres-Latorre, C., Rodriguez-Serrano, J. A., & Vitria, J. (2020). Building uncertainty models on top of black-box predictive APIs. *IEEE Access*, 8, 121344–121356. <https://doi.org/10.1109/ACCESS.2020.3006711>
- Chatterji, G., Sridhar, B., Sheth, K., Kim, D., & Mulfinger, D. (2004). Methods for establishing confidence bounds on sector demand forecasts. *AIAA Guidance, Navigation, and Control Conference and Exhibit*. <https://doi.org/10.2514/6.2004-5232>
- Cherian, J., & Bronner, L. (2020). *How the washington post estimates outstanding votes for the 2020 presidential election*. The Washington Post. Retrieved April 13, 2026, from <https://washpost.engineering/how-the-washington-post-estimates-outstanding-votes-for-the-2020-presidential-election-3f82f8415eda>
- Dalmau, R., Trzmiel, A., & Kirby, S. (2025). Combining machine learning models to improve estimated time of arrival predictions. *European Journal of Transport and Infrastructure Research*, 25(1). <https://doi.org/10.59490/ejtir.2025.25.1.7488>
- Dekker, R. (2023). *Decision support tool - ontwerpdocument* (Internal Document).
- Dewolf, N., Baets, B. D., & Waegeman, W. (2023). Valid prediction intervals for regression problems. *Artificial Intelligence Review*, 56(1), 577–613. <https://doi.org/10.48550/arXiv.2107.00363>
- Do, Q. H., Lo, S.-K., Chen, J.-F., Le, C.-L., & Anh, L. H. (2020). Forecasting air passenger demand: A comparison of LSTM and SARIMA. *Journal of Computer Science*, 16(7), 1063–1084. <https://doi.org/10.3844/jcssp.2020.1063.1084>
- Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H. S., & Schmidt-Thieme, L. (2021). Do we really need deep learning models for time series forecasting? *arXiv preprint arXiv:2101.02118*. <https://doi.org/10.48550/arXiv.2101.02118>
- EUROCONTROL. (2019). *Think paper #2: Air traffic flow management (atfm) regulations: A power for good*. EUROCONTROL. Retrieved April 13, 2026, from <https://www.eurocontrol.int/sites/default/files/2020-01/eurocontrol-think-paper-2-atfm-regulation.pdf>
- EUROCONTROL. (2025a). *Air traffic in the european network in summer 2025*. EUROCONTROL. Retrieved November 28, 2025, from <https://www.eurocontrol.int/summer-2025>
- EUROCONTROL. (2025b). *Coda digest: Annual report 2024*. EUROCONTROL. Retrieved April 13, 2026, from <https://www.eurocontrol.int/sites/default/files/2025-07/eurocontrol-coda-digest-annual-report-2024.pdf>
- European Union Aviation Safety Agency, European Environment Agency, & EUROCONTROL. (2025). *European aviation environmental report 2025*. European Union Aviation Safety Agency. Retrieved

- April 13, 2026, from https://www.easa.europa.eu/sites/default/files/eaer-downloads/EASA_EAER_2025_Book_v5.pdf
- Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34, 1660–1672. <https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html>
- Gibbs, I., & Candès, E. (2024). Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162), 1–36. <http://jmlr.org/papers/v25/22-1218.html>
- Gilbo, E., & Smith, S. (2007). A new model to improve aggregate air traffic demand predictions. *AIAA Guidance, Navigation and Control Conference and Exhibit*. <https://doi.org/10.2514/6.2007-6450>
- Gilbo, E., & Smith, S. (2009). Probabilistic prediction of aggregate traffic demand using uncertainty in individual flight predictions. *AIAA Guidance, Navigation, and Control Conference*. <https://doi.org/10.2514/6.2009-6194>
- Glina, Y., Jordan, R., & Ishutkina, M. (2012). A tree-based ensemble method for the prediction and uncertainty quantification of aircraft landing times. *American Meteorological Society–10th Conference on Artificial Intelligence Applications to Environmental Science*.
- Gonze, F., Huens, E., Jungers, R. M., Simonetto, A., & Boucquey, J. (2018). Probabilistic occupancy counts and flight criticality measures in air traffic management. *Journal of Air Transportation*, 26(3), 94–103. <https://doi.org/10.2514/1.D0087>
- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 507–520. https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cda dc-Abstract-Datasets_and_Benchmarks.html
- Januschowski, T., Wang, Y., Torkkola, K., Erkkilä, T., Hasson, H., & Gasthaus, J. (2022). Forecasting with trees. *International Journal of Forecasting*, 38(4), 1473–1481. <https://doi.org/10.1016/j.ijforecast.2021.10.004>
- Kern, C. S., de Medeiros, I. P., & Yoneyama, T. (2015). Data-driven aircraft estimated time of arrival prediction. *2015 Annual IEEE Systems Conference (SysCon) Proceedings*, 727–733. <https://doi.org/10.1109/SYSCON.2015.7116837>
- Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50. <https://doi.org/10.2307/1913643>
- Luchtverkeersleiding Nederland (LVNL). (n.d.). *QRC 084 - Richtlijnen ATFM regulations* (Uitgave 015) [Internal Document. Beheer: P&D]. LVNL.
- Luo, R., & Zhou, Z. (2025). Conformal thresholded intervals for efficient regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18), 19216–19223. <https://doi.org/10.48550/arXiv.2407.14495>
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74. <https://doi.org/10.1016/j.ijforecast.2019.04.014>
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4), 1346–1364. <https://doi.org/10.1016/j.ijforecast.2021.11.013>
- Meyn, L. (2002). Probabilistic methods for air traffic demand forecasting. *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 4766. <https://doi.org/10.2514/6.2002-4766>
- O’Sullivan, A., Pereira, F. C., Zhao, J., & Koutsopoulos, H. N. (2016). Uncertainty in bus arrival time predictions: Treating heteroscedasticity with a metamodel approach. *IEEE Transactions on Intelligent Transportation Systems*, 17(11), 3286–3296. <https://doi.org/10.1109/TITS.2016.2547184>

- Pereira, F. C., Antoniou, C., Fargas, J. A., & Ben-Akiva, M. (2014). A metamodel for estimating error bounds in real-time traffic prediction systems. *IEEE Transactions on Intelligent Transportation Systems*, 15(3), 1310–1322. <https://doi.org/10.1109/TITS.2014.2300103>
- Romano, Y., Patterson, E., & Candès, E. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32. <https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html>
- Rossellini, R., Barber, R. F., & Willett, R. (2024). Integrating uncertainty awareness into conformalized quantile regression. *International Conference on Artificial Intelligence and Statistics*, 1540–1548. <https://doi.org/10.48550/arXiv.2306.08693>
- Royal Schiphol Group. (2025). *Schiphol's performance and finances improve in 2024*. Royal Schiphol Group. Retrieved April 13, 2026, from <https://nieuws.schiphol.nl/download/4600cb97-bdd0-4af6-822a-68af1d986eba/schiphol039sperformanceandfinancesimprovein2024.pdf>
- Ruskamp, T. M. (2025). *Prediction of traffic take-off times at out-stations* [Master's thesis, Delft University of Technology].
- Sanaei, R., Pinto, B. A., & Gollnick, V. (2021). Toward ATM resiliency: A deep CNN to predict number of delayed flights and ATFM delay. *Aerospace*, 8(2), 28. <https://doi.org/10.3390/aerospace8020028>
- Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421. <https://jmlr.org/papers/v9/shafer08a.html>
- Shone, R., Glazebrook, K., & Zografos, K. G. (2021). Applications of stochastic modeling in air traffic management: Methods, challenges and opportunities for solving air traffic problems under uncertainty. *European Journal of Operational Research*, 292(1), 1–26. <https://doi.org/10.1016/j.ejor.2020.10.039>
- Suresh, V., Swain, A., Revathi, B. S., & Guerrero, J. M. (2025). Mamba based adaptive conformal inference for probabilistic short-term load forecasting. *Knowledge-Based Systems*, 114222. <https://doi.org/10.1016/j.knosys.2025.114222>
- Tielrooij, M., Borst, C., Van Paassen, M. M., & Mulder, M. (2015). Predicting arrival time uncertainty from actual flight information. *Proceedings of the 11th USA/Europe Air Traffic Management Research and Development Seminar*, 577–586.
- Vegter, S., Murrieta-Mendoza, A., Zuñiga, C., Noordeloos, K., & Westerveld, E. (2023). Sector traffic balancing using tactical demand tailoring. *13th SESAR Innovation Days 2023, SIDS 2023*. https://www.sesarju.eu/sites/default/files/documents/sid/2023/Papers/SIDs_2023_paper_21%20final.pdf
- Vos, R., Sun, J., & Hoekstra, J. (2024). A transformer-based trajectory prediction model to support air traffic demand forecasting. *Proceedings of the 11th International Conference for Research in Air Transportation (ICRAT)*.
- Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.
- Wanke, C., Callaham, M., Greenbaum, D., & Masalonis, A. (2003). Measuring uncertainty in airspace demand predictions for traffic flow management applications. *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 5708. <https://doi.org/10.2514/6.2003-5708>
- Wanke, C., Song, L., Zobell, S., Greenbaum, D., & Mulgund, S. (2005). Probabilistic congestion management. *Proceedings of the 6th Europe-USA ATM Seminar*. Baltimore. US.
- Wijngaard, J., Vogelesang, D., van Bruggen, H., Maat, N., de Rover, C. J., Smit, L., Heijstek, J., Keet, M., & ten Hove, R. (2008). *Improved low visibility and ceiling forecasts at schiphol airport: Final report, part 1* (tech. rep. No. KNMI publication 222 (KDC/2008/0089)). Royal Netherlands Meteorological Institute (KNMI). De Bilt, The Netherlands. Retrieved April 13, 2026, from <https://cdn.knmi.nl/knmi/pdf/bibliotheek/knmipubmetnummer/knmipub222.pdf>
- Zaffran, M., Dieuleveut, A., Féron, O., Goude, Y., & Josse, J. (2022). Adaptive conformal predictions for time series. *International Conference on Machine Learning*, 25834–25866. <https://proceedings.mlr.press/v162/zaffran22a.html>

- Zhao, Z., Yuan, J., & Chen, L. (2024). Air traffic flow management delay prediction based on feature extraction and an optimization algorithm. *Aerospace*, 11(2), 168. <https://doi.org/10.3390/aerospace11020168>