STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, August 2026, pp 1060–1085.
Published online in International Academic Press (www.IAPress.org)

Rolling-Calibrated Conformal Decision Support for Seasonal Airport Taxi-Time Planning under Distribution Shift

Soufiane Momtaz

ENSET Mohammedia, Hassan II University of Casablanca, Morocco

Abstract Seasonal airport taxi-time planning is considered to be an asymmetric-risk decision problem, under-planning in operations can propagate delay and disrupt airline schedules, whereas excessive padding can reduce aircraft utilization and contribute to schedule creep which explains the importance of smart balance. This paper formulates seasonal taxi-time planning as a one-sided risk-control problem and studies a rolling-calibrated conformal decision-support framework for airline scheduling under non-stationary operating conditions. The proposed framework combines a pooled seasonal point forecaster with rolling one-sided conformal calibration to convert point predictions into planning recommendations with an explicit marginal under-planning target. This paper operationalizes existing one-sided conformal calibration for a real aviation planning problem, evaluates its empirical behavior over 17 IATA seasons and 722 airports in the EUROCONTROL CODA dataset, and provides deployment-oriented diagnostics for airport-level monitoring, cohort instability, sparse-history airports, and wake-turbulence-category heterogeneity.Results show that traditional persistence heuristics can fail during the disruption/recovery period, whereas rolling conformal calibration substantially reduces under-planning toward the nominal planning target, at the cost of additional padding. We explicitly distinguish formal conformal guarantees under exchangeability from empirical rolling adaptation under distribution shift. The framework is therefore best interpreted as an auditable, human-supervised decision-support layer for seasonal planning, not as an autonomous scheduling rule or a guarantee-restoration mechanism under arbitrary drift.

Keywords Conformal prediction, uncertainty quantification, statistical decision support, rolling calibration, empirical adaptation, risk-controlled planning, taxi time, airport surface operations.

AMS 2020 subject classifications 90B06, 62M20, 62G15

DOI: 10.19139/soic-2310-5070-3782

1. Introduction

Planning under non-stationarity in operational context is a challenging problem that stands in the intersection of statistics, optimization and information computing. Airport seasonal taxi-time planning is a concrete aviation example because seasonal benchmark values are used later in airline scheduling, block-time design, and airport performance analysis. The task is treated as a *risk-controlled statistical decision problem* rather than as point forecasting alone and the recommendation is to protect against operationally costly shortfalls while keeping additional padding visible and controllable.

Airport surface operations provide the application setting studied here. Taxi-in and taxi-out times affect block-time design, on-time performance, network reliability, cost, and engine-on emissions. Hence airline schedulers and performance analysts often rely on *seasonal planning values* rather than on single-flight predictions.

EUROCONTROL's Central Office for Delay Analysis (CODA) publishes such seasonal taxi-time statistics from standardized airline-reported timestamps, including the 90th percentile ("$P90$") planning value [1]. The underlying timestamps are standardized by the Air Transport Operator Data Flow (AODF) specification (AOBT, ATOT, ALDT, AIBT), enabling consistent benchmarking across reporting airlines [2]. Seasonal planning also aligns naturally with Airport Collaborative Decision-Making (A-CDM) practices that coordinate surface processes and milestone times [3].

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Methodologically, the main challenge is that the planner needs a recommendation with a one-sided service-level interpretation. The realized seasonal benchmark should not normally exceed the recommended value, at a user-selected rate α, while unnecessary padding should remain controlled and the trad-off is questionable over long horizons because demand, capacity, airline strategy, and disruption regimes evolve across IATA seasons. In this paper, we study how rolling conformal calibration can operationalize a transparent risk–padding trade-off under observed non-stationarity, without claiming a formal guarantee under arbitrary distribution shift.

1.1. Why point prediction alone is insufficient

Seasonal taxi planning is inherently a decision problem. A scheduler selects a planning value (or buffer) that trades off two competing effects:

- **Under-planning risk** (planning too low): infeasible rotations, missed connections, spillover delays, and stronger disruption propagation.
- **Over-planning cost** (planning too high): inflated schedules, lower asset utilization, distorted benchmarking, and pressure toward schedule creep.

A point forecast may achieve acceptable mean absolute error while still failing on the one-sided loss that matters in practice and should be considered for reliability and it is significant under distribution shift because regime changes can preserve average error at some airports and alter the frequency of operational shortfalls. The methodological target is *auditable control of under-planning frequency* with an explicit accounting of the added margin.

In this study, the main challenge is to consider and manage the fact that seasonal taxi-time planning is not only a forecasting problem but it is also a decision problem with asymmetric consequences because planning too little taxi time can amplify delay propagation and disrupt aircraft rotations, whereas planning too much taxi time consumes schedule slack, reduces asset utilization, and may gradually inflate scheduled block times. This is why the related asymmetry justifies and explains a one-sided risk-control formulation rather than an evaluation based only on point-forecast accuracy.

An advisory agent, not autonomous control. Throughout the paper, the term *advisory agent* is used in a narrow decision-support sense: a policy π that maps a planning state (airport, season, recent history) to (i) a recommended planning value b, (ii) an uncertainty margin Δ in minutes, and (iii) explicit guardrail triggers (refer for review or cap season-to-season changes). We do *not* claim an autonomous controller or a reinforcement-learning agent interacting with a live environment; the setting is offline seasonal planning.

Scope of the study. We study a single planning artifact: CODA airport–season benchmark values (here, taxi-in/out $P90$) used in scheduling and performance workflows. The agent is offline and advisory. This study does not *not* claim any of the folowing:(i) flight-level optimization or real-time A-CDM control, (ii) causal effects on emissions or delay propagation, (iii) *conditional* per-airport coverage—risk control is marginal over the CODA-reported airport cohort in each season, (iv) distribution-free guarantees under arbitrary drift, or (v) fully local cold-start solutions without sufficient history. These exclusions define the intended scope of the paper as an offline, human-supervised decision-support study.

This paper should be considered as an operational and empirical contribution, the study explains how rolling one-sided conformal calibration can be operationalized for seasonal airport taxi-time planning, how it behaves empirically during disruption and recovery regimes, and what monitoring safeguards are required for responsible deployment in airline scheduling.

1.2. Contributions

The main contributions of this study are fourfold.

- **One-sided statistical decision formulation:** we formulate seasonal airport taxi-time planning as a one-sided risk-control problem, where the operational objective is to limit under-planning rather than merely minimize average prediction error. This formulation aligns the statistical task with the asymmetric cost structure of airline scheduling.

- **Rolling-calibrated conformal planning layer:** we operationalize rolling one-sided conformal calibration as an auditable planning layer on top of seasonal point forecasts. The conformal margin is interpreted as an empirical residual-quantile safety margin, not as a causal explanation of the operational factors driving taxi-time variability.
- **Large-scale empirical evaluation:** we evaluate the framework on a large EUROCONTROL CODA seasonal panel covering 17 IATA seasons and 722 airports, with particular attention to the disruption/recovery window in which persistence-based planning heuristics lose under-planning control.
- **Deployment safeguards:** we provide deployment-oriented safeguards for responsible use: explicit clarification of the scope of the conformal guarantee under distribution shift, airport-level monitoring, cohort-stability diagnostics, sparse-history and cold-start handling, wake-turbulence-category stratification, and operational interpretation of the padding–risk trade-off.

Decision contract. For each airport–season planning state, the agent outputs a conservative recommendation b (in minutes) and a margin Δ such that the under-planning event $\{y > b\}$ is targeted at rate α under approximate exchangeability within the rolling calibration window of length K. We interpret this as a *marginal service-level expectation* over the evaluated CODA airport–season cohort, not as a conditional per-airport guarantee. The decision contract is therefore audited through aggregate under-planning, per-airport diagnostics, margin inflation, and cohort-stability monitoring.

2. Related Work

The proposed framework can be situated at the intersection of four research streams. First, operational taxi-time benchmarking, second, airport surface-time modeling, third, airline schedule robustness, and last, uncertainty quantification under distribution shift. The emphasis is on converting seasonal airport benchmark forecasts into auditable one-sided planning recommendations with explicit under-planning and padding diagnostics.

2.1. Operational Taxi-Time Benchmarks and A-CDM Context

Literature demonstrates how operational taxi-time planning relies on standardized definitions of airport surface milestones and on comparable seasonal statistics. EUROCONTROL CODA taxi-time planning values are built from airline-reported timestamps, while the AODF specification defines the operational timestamp semantics that make cross-airport benchmarking meaningful [1, 2]. A-CDM practice further motivates the use of shared, auditable time estimates in surface decision processes and network coordination [3]. Related EUROCONTROL performance-indicator documentation treats additional taxi-out and taxi-in time as operational measures linked to surface queuing, congestion, and reference-time estimation [4, 5]. These operational sources motivate our focus on taxi-time planning values as decision-support objects rather than as purely statistical response variables.

2.2. Taxi-Time Modeling and Airport Surface Operations

Literature presents how airport taxi-time modeling has a long queueing and surface-operations foundation. Early taxi-out queueing models represent the departure process as a congestion-sensitive system in which surface traffic and runway queues drive taxi-out variability [6, 7]. More recent integrated surface–airspace models extend this perspective by coupling airport surface states with terminal-airspace departure constraints [8]. In parallel, data-driven methods have studied regression, machine learning, reinforcement learning, and airport-specific model structures for taxi-time prediction [9–12]. Feature-importance analyses show that taxi-time drivers can vary materially across airports, movement types, and operating regimes [13], and recent reviews organize the field around taxi-time type, movement type, and modeling methodology [14]. Surface-management interventions such as pushback allocation and stand holding address congestion through operational control decisions [15]. This study deliberately remains at the seasonal planning layer: it uses sparse airport-season histories and exposes one-sided risk margins instead of optimizing flight-level surface trajectories or real-time control actions.

2.3. Airline Schedule Design, Block-Time Reliability, and Delay Propagation

Literature explains why and how taxi-time planning is connected to airline schedule robustness because taxi buffers enter scheduled block times, aircraft rotations, connection feasibility, and delay propagation. Robust airline-operation models show how routing and departure-time decisions can be adjusted to reduce passenger disruption and propagated delay [16], while robust block-time scheduling explicitly treats uncertainty in planned flight durations [17]. Also integrated schedule-planning models and slack re-allocation methods further show that robustness depends not only on the amount of buffer, but also on where that buffer is placed in the network [18,19]. In the other side empirical block-time studies document how reliability incentives shape scheduled block-time setting and later adjustments [20, 21], and long-run schedule creep reflects systematic growth in scheduled times under operational and strategic pressures [22]. Delay propagation studies and disruption-management reviews show why upstream surface-time under-planning can have downstream network consequences [23,24]. This contribution is intentionally narrower than network-wide optimization models because it provides a calibrated seasonal taxi-time component that enrich schedule-design workflows while keeping the risk–padding trade-off transparent.

2.4. Dataset Shift and Drift in Operational AI

Literature justifies why seasonal taxi-time planning is exposed to distribution shift because traffic demand, airport capacity, airline schedules, runway-use patterns, reporting cohorts, and disruption regimes evolve over time. The machine-learning literature distinguishes several forms of dataset shift, including covariate shift, prior shift, and concept drift [25]. Concept-drift surveys emphasize that models deployed in evolving environments require temporal evaluation, monitoring, and adaptation rather than one-time validation [26]. This point is especially important for airport planning because an airport-season panel is short, heterogeneous, and affected by exogenous disruption and recovery periods. We therefore treat non-stationarity as a primary evaluation condition and use rolling calibration as an empirical adaptation mechanism to recent residual behavior; we do not claim that rolling windows identify drift causes or restore distribution-free validity under arbitrary non-exchangeability.

2.5. Uncertainty Quantification and Conformal Prediction

Literature that treats conformal prediction provides a distribution-free route to finite-sample marginal prediction sets under exchangeability [27, 28]. In regression, split conformal inference and related model-agnostic methods provide predictive intervals without assuming a parametric residual distribution [29]. Conformalized quantile regression improves adaptivity under heteroskedasticity [30], weighted conformal prediction addresses covariate shift under appropriate reweighting assumptions [31], and jackknife+ methods provide resampling-based finite-sample predictive inference [32]. Broader related work synthesize and review these guarantees and clarify both their assumptions and limitations [33, 34].

Recent literature also treats conformal inference beyond the standard exchangeable setting. Time-series and sequential conformal methods adapt calibration to temporally ordered data [35, 36]. Adaptive conformal inference updates the effective miscoverage level online under distribution shift [37], and subsequent online conformal work studies adaptation to arbitrary distribution shifts with local-regret guarantees [38]. Non-exchangeable conformal prediction generalizes conformal ideas through weighted quantiles and robustness analyses when exchangeability is violated [39], while robust validation studies predictive coverage under distributional perturbations around a reference population [40]. The proposed method is operationally aligned with this literature but intentionally simpler: it applies one-sided split conformal calibration on rolling airport-season residuals to produce seasonal planning recommendations. The resulting guarantee should be interpreted as the standard conformal marginal guarantee under exchangeability, plus empirical rolling adaptation diagnostics under observed shift.

3. Data: CODA Taxi-Time Planning Values

CODA seasonal taxi-time statistics are computed from airline-reported operational timestamps [1]. Taxi-out is defined as ATOT−AOBT and taxi-in as AIBT−ALDT, where AOBT/ATOT/ALDT/AIBT are standardized by

AODF [2]. CODA provides seasonal benchmark values for both summer and winter IATA seasons, including mean, standard deviation, 10th percentile, median, and 90th percentile. IATA summer seasons cover 31 weeks starting the last Sunday of March, and winter seasons cover 21 weeks starting the last Sunday of October [1]. Extreme outliers are filtered (taxi-in >120 min, taxi-out >180 min), and airports with fewer than 100 observations per season are excluded [1].

3.1. Panel structure

An *instance* is an airport–season pair (a, t). We denote by $y_{a,t}^{\mathrm{TXO}}$ and $y_{a,t}^{\mathrm{TXI}}$ the CODA taxi-out and taxi-in $P90$ planning values (minutes). The dataset spans 17 IATA seasons (S14–S22, W14–W21) and covers 722 airports with 8,659 airport-season records (not all airports appear in every season). Because CODA applies minimum-observation thresholds, the reported airport cohort varies by season (TXO: 373–570 airports; TXI: 368–574 in our panel), which we treat as an operational population shift and explicitly monitor as a cohort-stability signal.

3.2. Wake turbulence category subset

For additional granularity, CODA provides taxi-out planning values by wake turbulence category (WTC) for a subset of airports, based on M (medium), H (heavy), and J (super), excluding L (light) [1]. We treat each (airport, WTC) pair as a separate seasonal series and report WTC-stratified results on this 53-airport subset.

4. Problem Formulation

For each airport a and future season $t+1$, the scheduler seeks a recommended planning value $b_{a,t+1}$ for a target planning statistic (taxi-out $P90$). Let $y_{a,t+1}$ denote the CODA planning value observed after the season closes.

We focus on a *one-sided* under-planning risk constraint:

$$\mathbb{P}\left(y_{a,t+1} \le b_{a,t+1}\right) \ge 1 - \alpha, \tag{1}$$

where $\alpha \in (0, 1)$ is a user-defined risk level ($\alpha = 0.1$). Violations of (1) correspond to *under-planning* events (recommendation too low).

Among solutions satisfying (1), a natural objective is to minimize expected *padding*:

$$\min\ \mathbb{E}\left[\max\left(0,\ b_{a,t+1} - y_{a,t+1}\right)\right], \tag{2}$$

The objective in (2) penalizes unnecessary buffer. This framing is related to the classic *newsvendor/quantile* decision problem: with asymmetric costs C_u (under) and C_o (over), the optimal risk level is $\alpha = C_o/(C_u + C_o)$. In practice, schedulers can select α to reflect service levels or operational tolerance, while the agent should expose the induced padding.

5. Method: Rolling-Calibrated Conformal Taxi-Time Planning Agent

The proposed agent has two layers: a global seasonal forecaster for point predictions and a rolling one-sided conformal layer to convert point predictions into risk-controlled planning recommendations.

5.1. Seasonal forecaster

We learn a global forecaster $\hat{y}_{a,t} = f(x_{a,t})$ for each target planning value, trained across airports. Features $x_{a,t}$ are available at planning time and are constructed from lagged seasonal statistics for the same airport, including:

- last-season lag $(t-1)$ and same-season lag $(t-2)$,
- rolling aggregates (mean/std over the last four seasons),
- season type (summer/winter) and an integer season index t,

• airport identity (one-hot encoded).

This design favors interpretability and reproducibility, while still exploiting cross-airport pooling to stabilize estimates for airports with sparse season coverage. In our experiments, $f(\cdot)$ is ridge regression with standardized numeric features and one-hot categorical encoding. This choice is deliberately conservative for auditability and to reduce overfitting on the short seasonal horizon; the conformal layer is model-agnostic and can wrap alternative regressors without changing the risk-control contract.

5.2. Rolling one-sided conformal calibration

To transform point predictions into a risk-controlled recommendation, we apply one-sided split conformal calibration [27, 29]. Given a calibration set $\mathcal{C}$, we compute signed residuals

$$r_{a,t} = y_{a,t} - \hat{y}_{a,t}. \tag{3}$$

Let $q_{1-\alpha}$ be the empirical $(1-\alpha)$-quantile of $\{r_{a,t} : (a,t) \in \mathcal{C}\}$ using the conservative conformal index

$$q_{1-\alpha} = \text{Quantile}_k\left(\{r_{a,t}\}\right), \quad k = \lceil (|\mathcal{C}|+1)(1-\alpha) \rceil. \tag{4}$$

The conformal upper recommendation is then

$$b_{a,t} = \hat{y}_{a,t} + q_{1-\alpha}. \tag{5}$$

Coverage contract and interpretation. Calibration residuals are pooled across airports. Thus, the risk constraint in (1) is best read as *marginal* risk control over the CODA-reported airport cohort in the relevant seasons (and marginal within a stratum when we stratify by season type or WTC). This paper reports both aggregate under-planning and per-airport diagnostics to expose heterogeneity, conditional per-airport guarantees would require substantially longer histories, stratified calibration, or hierarchical calibration and are left for future work.

5.3. Scope of the Conformal Guarantee under Distribution Shift

It is important to distinguish the standard conformal guarantee from the empirical behavior of the proposed rolling calibration procedure under non-stationary conditions. Under exchangeability, conformal prediction provides finite-sample marginal coverage guarantees. However, this assumption is generally violated when airport operating conditions, traffic levels, reporting cohorts, or post-disruption recovery regimes change over time.

Under arbitrary distribution shift, rolling conformal calibration does not restore finite-sample coverage guarantees. The rolling calibration window is therefore not used here as a guarantee-restoration device. It is used as an empirical adaptation mechanism that gives more weight to recent residual behavior and reduces reliance on older seasonal errors that may no longer represent current operating conditions.

Consequently, the under-planning rates reported during the disruption/recovery window should be interpreted as empirical findings for the observed CODA shift patterns, not as universal distribution-free guarantees for all future drift scenarios. This distinction is important to the use of the proposed framework by considering that the method provides an auditable, human-supervised risk-control layer that must be accompanied by monitoring diagnostics, rather than an autonomous guarantee of performance under arbitrary non-stationarity.

Algorithm 1 summarizes the rolling one-sided conformal procedure used in each backtest step. Operationally, K acts as a robustness parameter because smaller K adapts more rapidly to recent regimes but yields noisier quantiles, whereas larger K stabilizes quantiles but may lag during abrupt shifts.

Algorithm 1 Rolling one-sided conformal taxi-time planning

Require: Target planning statistic y, season index t, risk level α, calibration window K

1: Train forecaster f using seasons $< t-K$
2: Predict calibration seasons $[t-K, t-1]$ and compute residuals $r = y - \hat{y}$
3: Compute $q_{1-\alpha}$ using (4)
4: Predict season t and output $b = \hat{y} + q_{1-\alpha}$ (and margin $\Delta = q_{1-\alpha}$)

5.4. Advisory policy, margins, and guardrails

The agent outputs $(b_{a,t+1}, \Delta_{t+1})$, where $\Delta_{t+1} = q_{1-\alpha}$ is an auditable statistical margin in minutes. This supports a set of guardrails designed for human-supervised seasonal planning:

- **High-uncertainty deferral:** if Δ_{t+1} exceeds a threshold, refer the case for operational review or apply a conservative recent-worst-case fallback, rather than reverting to uncalibrated persistence.
- **Human-in-the-loop overrides:** the margin helps justify unusually large buffers as data-driven responses to recent residual volatility, while leaving the final scheduling decision to the human planning process.
- **Monitoring:** deviations of realized under-planning from α provide a direct monitoring signal, especially when analyzed jointly with airport-level diagnostics and cohort-stability flags.

Figs. 1 and 2 summarize the decision output exposed to planners and the overall advisory architecture.

Illustrative decision card.
Target: Taxi-out $P90$ (min)
Airport: a (ICAO) Season: $t+1$
Point forecast: $\hat{y}_{a,t+1}$
Risk level: α Calibration window: K seasons
Conformal margin: $\Delta_{t+1} = q_{1-\alpha}$
Recommended planning value: $b_{a,t+1} = \hat{y}_{a,t+1} + \Delta_{t+1}$
Guardrail: if Δ_{t+1} exceeds threshold $\rightarrow$ review or conservative fallback

Figure 1. Illustrative decision card showing the recommendation and its conformal margin.

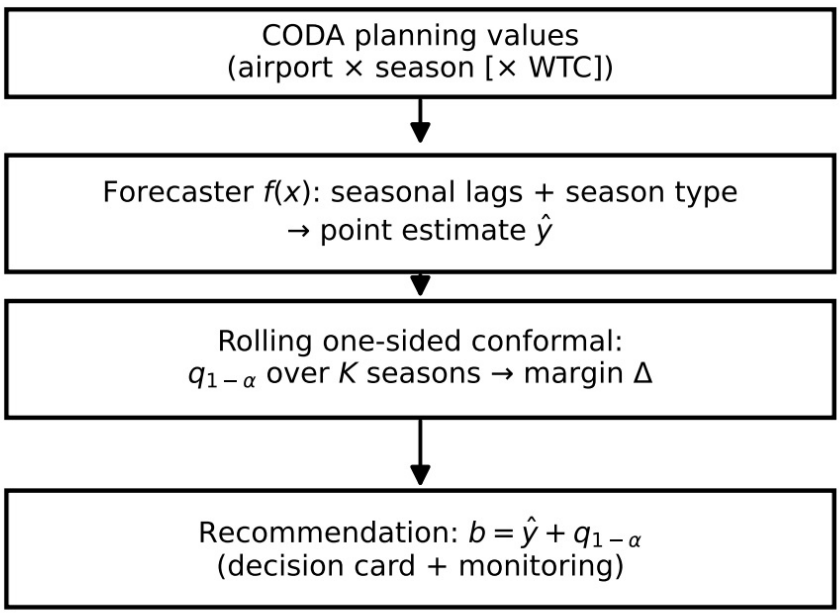


Figure 2. Architecture of the seasonal planning advisory agent.

5.5. Computational complexity

For each season, the forecaster is trained on a few thousand airport-season records and the conformal layer computes a single quantile over calibration residuals. Both steps are computationally modest and compatible with seasonal deployment cycles. In a CPU-only audit run of the default Ridge+CP rolling evaluation over the TXO/TXI

panels (9,919 evaluated airport-season decisions), the full rolling-origin loop completed in 15.4 seconds, including repeated seasonal refits and conformal-margin computation. This timing is hardware-dependent and is reported only to document operational latency order of magnitude.

6. Experimental Protocol

6.1. Rolling-origin backtesting

We evaluate with rolling-origin backtesting across IATA seasons. For each test season t, we:

1. train the forecaster on seasons $< t - K$,
2. calibrate $q_{1-\alpha}$ on the previous K seasons $[t - K, t - 1]$,
3. test on season t using (5).

Unless specified otherwise, we set $\alpha = 0.1$ and $K = 6$.

6.2. Baselines

We compare against transparent heuristics that are commonly used as *planning* proxies when high-resolution covariates are unavailable:

- **Last-season persistence:** $\hat{y}_{a,t} = y_{a,t-1}$.
- **Same-season persistence:** $\hat{y}_{a,t} = y_{a,t-2}$ (one year prior; two-season lag), with fallback to last-season when the same-season lag is missing.
- **Rolling mean (4):** the mean of the last four seasons for the same airport (when available).

Same-season persistence corresponds to the *seasonal-naïve* forecast with period 2, a widely used benchmark in seasonal time-series forecasting.

Why we do not use per-airport ARIMA/ETS baselines. Per airport univariate models are not reliable because airport series are short (at most 17 seasons, often fewer) and missingness changes with reporting thresholds. We therefore prefer auditable seasonal heuristics and a pooled forecaster, and the conformal layer is model-agnostic. These baselines are deliberately transparent. The aim is not to increase the accuracy of point-forecasts by themselves, but to determine whether explicit risk control by conformal calibration improves the operational suitability of seasonal planning recommendations.

6.3. Metrics

We report: (i) point prediction accuracy (MAE), (ii) under-planning rate $\mathbb{P}(y > b)$, and (iii) average padding $\mathbb{E}[\max(0, b - y)]$. Under-planning is the primary risk metric, aligned with the decision constraint (1).

6.4. Disruption/recovery evaluation window and ablations

To highlight drift, we report results for the disruption/recovery period (S20–S22, W20–W21) as the focused evaluation window. We include ablations over $K \in \{2, 4, 6, 8\}$ and sensitivity to $\alpha \in \{0.05, 0.1, 0.2\}$. For WTC, we evaluate taxi-out $P90$ stratified by category.

7. Results

Tables 1, 2, 3, and 4 summarize overall late-season performance, the focused disruption/recovery window, baseline comparisons, and feature ablations.

Table 1. Rolling-origin evaluation on late seasons (S19–S22, W19–W21), $K = 6$, $\alpha = 0.1$. Under is the under-planning rate; Pad is average padding in minutes. (TXO: $n = 3412$ airport-season records; TXI: $n = 3498$).

Target	Method	Under	Pad	MAE
TXO $P90$	Same-season	0.355	1.011	1.803
	Conformal	0.105	2.893	2.072
TXI $P90$	Same-season	0.254	0.461	0.863
	Conformal	0.100	1.348	0.916

Table 2. Disruption/recovery window (S20–S22, W20–W21), $K = 6$, $\alpha = 0.1$. (TXO: $n = 2351$ airport-season records; TXI: $n = 2437$).

Target	Method	Under	Pad	MAE
TXO $P90$	Same-season	0.397	1.050	1.903
	Conformal	0.095	3.054	2.046
TXI $P90$	Same-season	0.267	0.452	0.863
	Conformal	0.092	1.382	0.916

Table 3. Reinforced comparison with planning baselines on the disruption/recovery seasons (S20–S22, W20–W21). CP denotes the identical rolling one-sided conformal calibration layer with $K = 6$ and $\alpha = 0.1$. The row "Same-season + CP" is the conformalized same-season persistence baseline added to disentangle the value of conformal calibration from the pooled seasonal forecaster.

Target	Method	Under	Pad	MAE
TXO $P90$	Last-season	0.375	1.117	2.069
	Same-season	0.398	1.042	1.900
	Roll-mean(4)	0.413	1.165	1.918
	Same-season + CP	0.105	2.547	2.868
	Full agent	0.095	3.054	2.046
TXI $P90$	Last-season	0.256	0.348	0.728
	Same-season	0.267	0.450	0.863
	Roll-mean(4)	0.348	0.467	0.830
	Same-season + CP	0.078	1.238	1.431
	Full agent	0.092	1.382	0.916

Table 4. Feature ablations for the conformal agent on the disruption/recovery seasons (S20–S22, W20–W21), $K = 6$, $\alpha = 0.1$. "Full" uses airport identity and cross-target lags; "No airport" removes ICAO; "No cross" removes cross-target lag features. Pad is minutes.

Variant	TXO $P90$			TXI $P90$		
	Under	Pad	MAE	Under	Pad	MAE
Full	0.095	3.054	2.046	0.092	1.382	0.916
No airport	0.136	2.367	1.748	0.118	1.168	0.804
No cross	0.093	3.083	2.076	0.087	1.380	0.911

Table 5. Forecaster ablation on the disruption/recovery seasons (S20–S22, W20–W21). Point-only rows use the raw point forecast as the planning recommendation. CP rows apply the identical rolling one-sided conformal calibration layer with $K = 6$ and $\alpha = 0.1$.

Target	Method	Under	Pad	Rec. MAE	Point MAE
TXO $P90$	Ridge point-only	0.324	1.411	1.911	1.911
	Ridge + CP	0.101	2.839	3.035	1.911
	GradBoost point-only	0.373	1.291	1.913	1.913
	GradBoost + CP	0.111	2.762	3.012	1.913
TXI $P90$	Ridge point-only	0.437	0.522	0.851	0.851
	Ridge + CP	0.112	1.245	1.352	0.851
	GradBoost point-only	0.414	0.498	0.825	0.825
	GradBoost + CP	0.107	1.250	1.366	0.825

This study uses a nonlinear Gradient Boosting ablation with and without the identical rolling conformal layer. The point-only comparison separates predictive flexibility from risk control, while the CP comparison tests whether a more expressive forecaster reduces the conformal margin, padding, or recommendation error. The results show that nonlinear forecasting can improve point accuracy for some targets, but point-only forecasts still do not provide an explicit under-planning contract. When wrapped by the same conformal layer, Gradient Boosting remains close to the nominal risk target and provides an operationally useful comparator. The interpretation is model-agnostic, the conformal layer supplies the one-sided planning contract, whereas the point forecaster determines how efficiently that contract is achieved in terms of padding and recommendation MAE.

The conformalized same-season persistence baseline provides the key decomposition. It shows that the conformal layer is the primary mechanism responsible for reducing under-planning, applying the same rolling conformal calibration to a same-season persistence forecast already moves empirical under-planning close to the configured target. This supports the revised framing of the paper as an operationalization of one-sided conformal risk control rather than as a claim of new forecasting theory. At the same time, the full pooled forecaster remains useful because it improves the accuracy of the resulting recommendation. On the disruption/recovery window, the full agent reduces recommendation MAE relative to conformalized persistence for both TXO and TXI, while keeping under-planning in the vicinity of the nominal target. Thus, the forecasting layer contributes efficiency and stabilization, whereas the conformal layer provides the explicit one-sided risk-control contract.

7.1. Conformal calibration variants

Beyond uniform rolling calibration, practitioners may prefer stratification (by season type) or recency weighting to emphasize the most recent regimes. Table 6 compares several calibration variants on the disruption/recovery window.

Table 6. Conformal calibration variants on the disruption/recovery seasons (S20–S22, W20–W21), $K = 6$, $\alpha = 0.1$. "Season-type" calibrates separate margins for summer versus winter; "Weighted" uses exponential recency weights (half-life 2 seasons).

Variant	TXO $P90$		TXI $P90$	
	Under	Pad	Under	Pad
Uniform	0.095	3.054	0.092	1.382
Season-type	0.103	2.990	0.092	1.370
Weighted	0.098	2.987	0.096	1.361

Two patterns are consistent across targets. First, same-season persistence can exhibit competitive MAE while remaining unreliable under drift, with substantial under-planning during recovery seasons. Second, the conformal layer trades moderate padding for substantial risk reduction, keeping under-planning close to the nominal level.

These differences are unlikely to be attributable to sampling variability alone: on the disruption/recovery window, TXO has $N = 2351$ airport-season decisions and TXI has $N = 2437$, so a binomial standard error at $\alpha = 0.1$ is approximately $\sqrt{\alpha(1-\alpha)/N} \approx 0.006$.

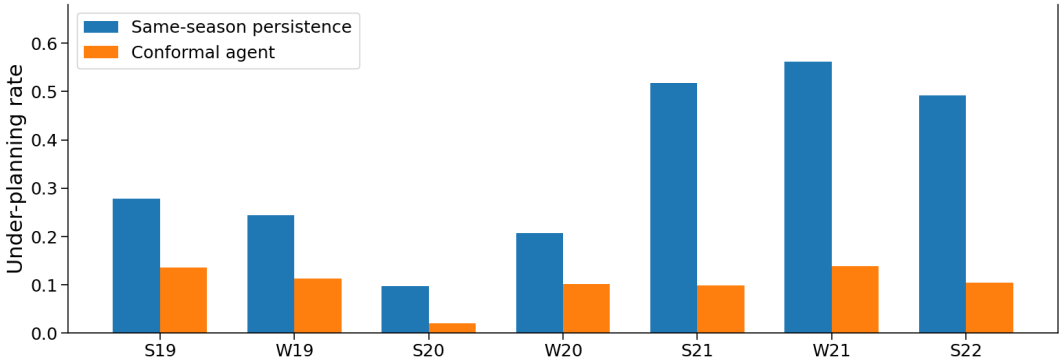
7.2. Season-by-season risk dynamics

Fig. 3 visualizes under-planning by season for taxi-out and taxi-in. The disruption/recovery seasons highlight the failure mode of same-season persistence and the stabilizing role of rolling conformal calibration.

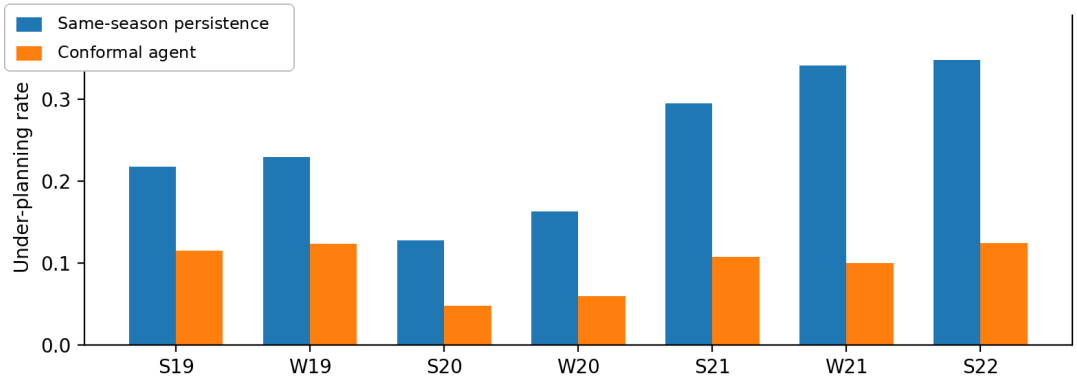
7.3. Airport-cohort stability and balanced-panel sensitivity

CODA reporting is an unbalanced operational panel, airports can enter or exit the published cohort when reporting coverage or minimum observation thresholds change. We add two diagnostics for population-shift monitoring. First, Table 7 and Fig. 4 report seasonal cohort turnover and flag seasons where the share of entering or exiting airports exceeds 20%. Second, Table 8 compares the full unbalanced panel with a balanced subset of airports present in all 17 seasons for the corresponding target.

We notice that the sensitivity analysis shows that the balanced subset is materially smaller than the full CODA panel, especially for taxi-out, and should be interpreted as a robustness check rather than as the primary deployment



(a) Taxi-out $P90$ under-planning.



(b) Taxi-in $P90$ under-planning.

Figure 3. Under-planning by IATA season: same-season persistence versus the conformal agent ($K = 6$, $\alpha = 0.1$).

Table 7. CODA airport-cohort stability diagnostics. Turnover is the share of entering or exiting airports relative to the union of the current and previous seasonal cohorts; seasons above 20% are flagged for recalibration review.

Target	Airport range	Balanced	Mean turnover	Max turnover	Flagged seasons (disruption/recovery)
TXO $P90$	373–570	196	0.208	0.371	S20; W20; S21; W21; S22
TXI $P90$	368–574	260	0.190	0.378	S20; W20; S21

Table 8. Balanced-panel sensitivity on the disruption/recovery seasons (S20–S22, W20–W21). The full panel uses all available CODA airport–season records; the balanced panel keeps only airports present in all 17 seasons for the corresponding target. CP denotes rolling one-sided conformal calibration with $K = 6$ and $\alpha = 0.1$.

Target	Method / panel	N	Airports	Under	Pad
TXO $P90$	Same-season + CP / full	2351	593	0.105	2.547
	Same-season + CP / balanced	980	196	0.111	2.466
	Ridge + CP / full	2351	593	0.101	2.839
	Ridge + CP / balanced	980	196	0.118	2.748
TXI $P90$	Same-season + CP / full	2437	592	0.078	1.238
	Same-season + CP / balanced	1300	260	0.075	1.211
	Ridge + CP / full	2437	592	0.112	1.245
	Ridge + CP / balanced	1300	260	0.137	1.087

population. We conlude that conformal calibration substantially reduces under-planning relative to uncalibrated

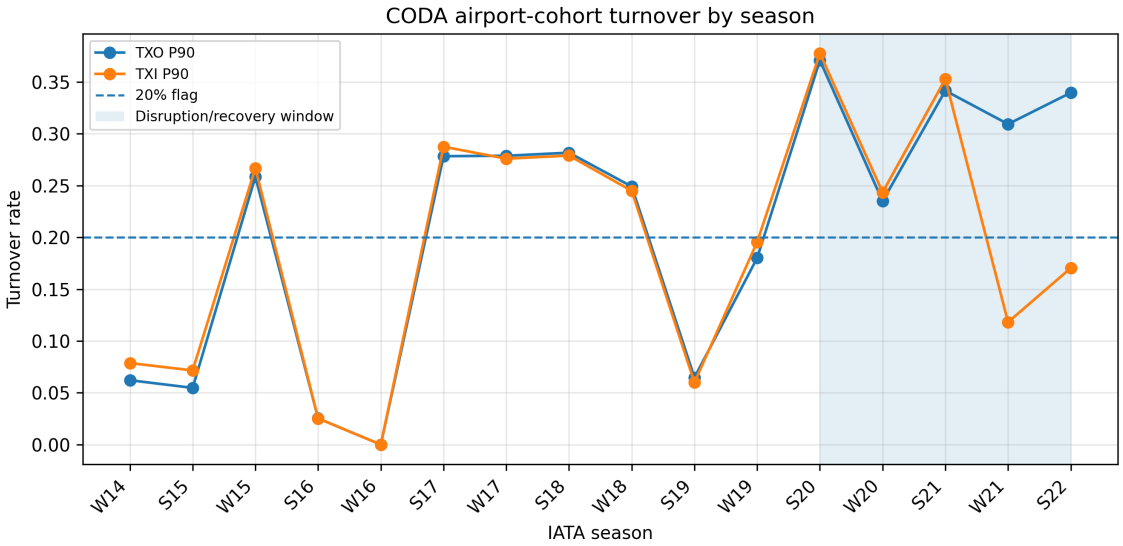


Figure 4. CODA airport-cohort turnover by IATA season. The dashed line marks the 20% deployment-monitoring flag, and the shaded region indicates the disruption/recovery window.

persistence, while cohort turnover must be monitored because the population on which marginal risk is evaluated can change from season to season.

7.4. Conformal margins as residual-volatility monitoring signals

The conformal margin $\Delta = q_{1-\alpha}$ should be interpreted as an empirical one-sided residual quantile: it summarizes recent forecast error volatility in the same units as the planning value (minutes), while remaining a statistical margin rather than a causal explanation of the underlying operational drivers. Large margins can therefore serve as an operational signal of recent instability: inflation in Δ indicates that recent seasons were harder to predict under the chosen model and calibration window, and may justify review or conservative fallback policies. Fig. 5 shows the seasonal evolution of this conformal margin for taxi-out and taxi-in planning targets.

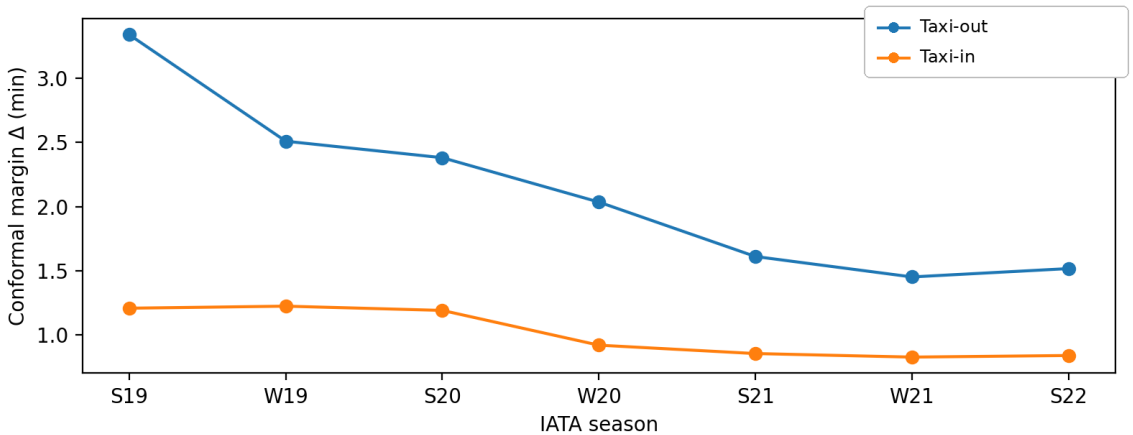
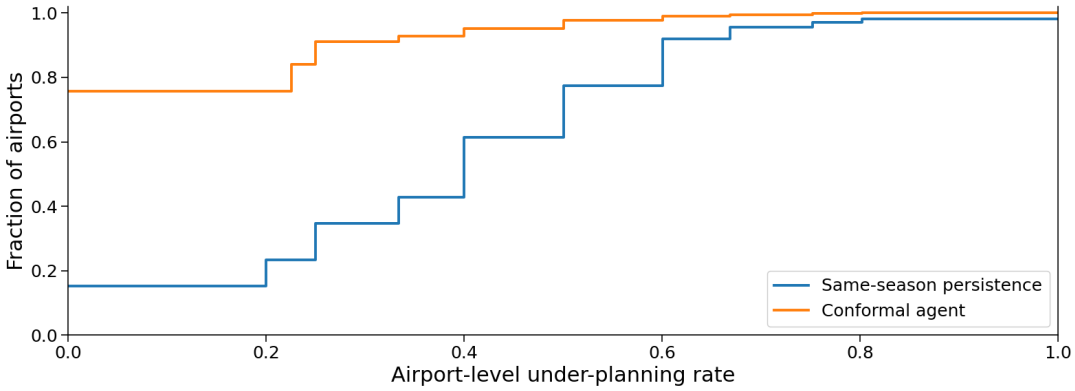


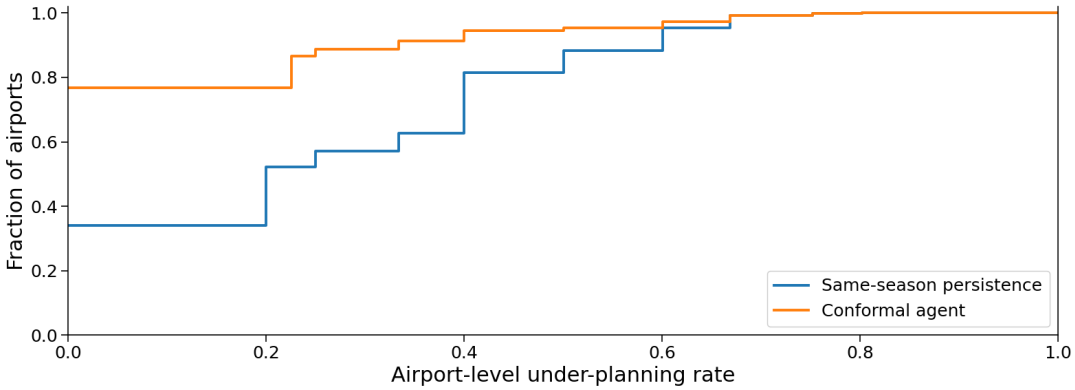
Figure 5. Seasonal conformal margins Δ (minutes) for taxi-out and taxi-in planning targets under $K = 6$, $\alpha = 0.1$.

7.5. Airport-level tail risk

Aggregate results can mask heterogeneity. Fig. 6 shows the distribution of airport-level under-planning rates over the disruption/recovery seasons. The agent shifts the CDF left, reducing the fraction of airports that experience repeated under-planning events across seasons.



(a) Taxi-out $P90$.



(b) Taxi-in $P90$.

Figure 6. CDF of airport-level under-planning rates during the disruption/recovery seasons.

Table 9 lists illustrative airports with the largest reductions in under-planning during the disruption/recovery seasons.

Table 9. Illustrative airports with the largest reduction in under-planning rate for taxi-out $P90$ during the disruption/recovery seasons (airports with at least 3 observed seasons).

ICAO	n	Under (base)	Under (agent)	Reduction
LIBC	3	1.000	0.000	1.000
LTCM	3	1.000	0.000	1.000
GCLP	5	0.800	0.000	0.800
LTCI	5	0.800	0.000	0.800
ESKN	5	0.800	0.000	0.800
EGSS	4	0.750	0.000	0.750
BKPR	4	0.750	0.000	0.750
VGHS	3	1.000	0.333	0.667
LGKF	3	1.000	0.333	0.667
GBYD	3	1.000	0.333	0.667

7.6. *Case study: a major hub airport*

To illustrate airport-level decision behavior, Fig. 7 reports a case study for EGLL (London Heathrow) on the late-season evaluation window. During the disruption period, same-season persistence may become either overly conservative (excess padding) or insufficiently responsive (elevated risk), whereas the conformal recommendation adapts through its rolling residual quantile. The recommendation is always accompanied by an explicit margin, thereby supporting human review when the implied buffer is operationally nontrivial.

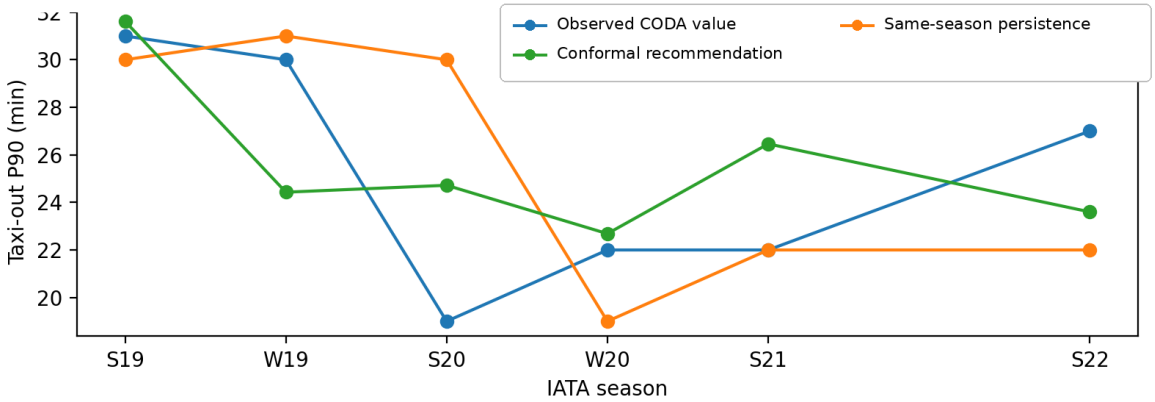


Figure 7. Case study (EGLL): taxi-out $P90$ CODA value, same-season persistence, and conformal recommendation ($K = 6$, $\alpha = 0.1$).

7.7. *Ablation: calibration window size* K

Table 10 reports under-planning and padding as K varies, while Fig. 8 summarizes the corresponding risk, padding, and MAE sensitivity in three panels.

Table 10. Calibration-window sensitivity and operational reading on the disruption/recovery window (S20–S22, W20–W21), with $\alpha = 0.1$. Under is empirical under-planning; Pad is average realized surplus in minutes.

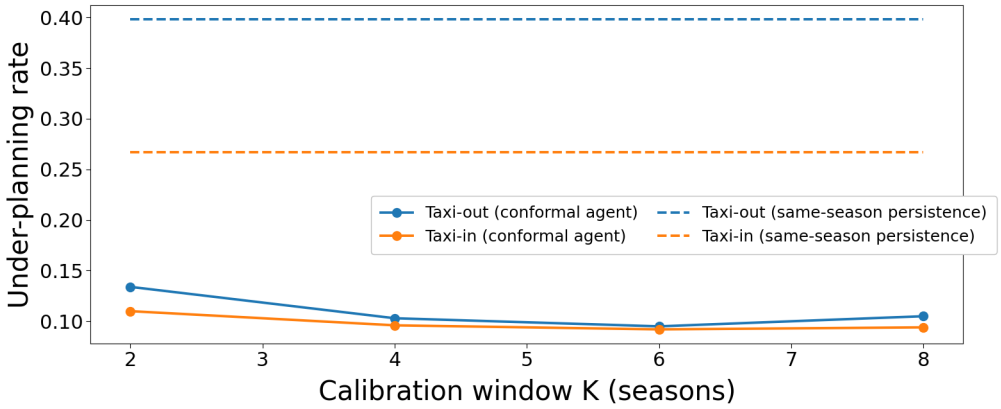
K	Operational reading	Under(TXO)	Pad(TXO)	Under(TXI)	Pad(TXI)
2	Fast, responsive but noisier	0.134	2.602	0.110	1.244
4	Responsive compromise	0.103	2.867	0.096	1.310
6	Balanced default	0.095	3.054	0.092	1.382
8	Stable but slower to adapt	0.105	2.911	0.094	1.393

Increasing K generally stabilizes calibration but affects padding according to the residual dynamics; thus, K should be interpreted as an operational tuning parameter rather than as a parameter for which larger values are uniformly preferable. Smaller windows are more responsive to abrupt regime changes but can be noisier, whereas larger windows are more stable but may lag after a sudden disruption. In this dataset, $K = 4$–6 provides the most practical compromise between empirical under-planning control and padding, while $K = 8$ should be interpreted as a more stable but potentially slower adaptation setting.

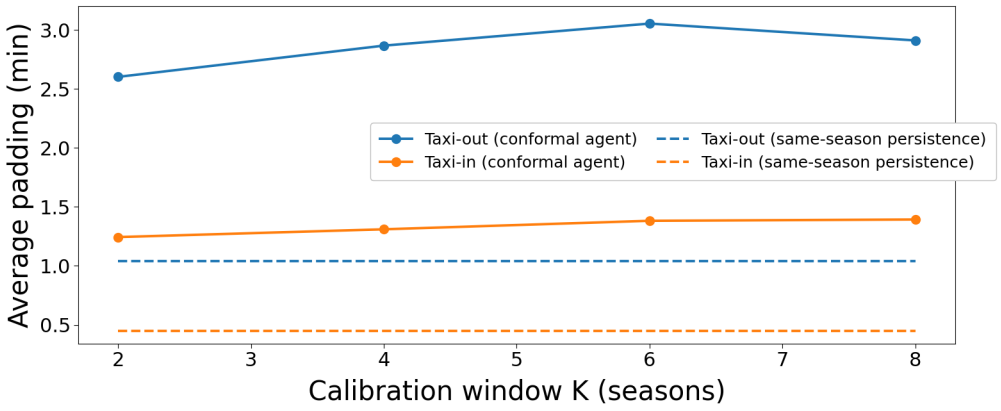
7.8. *Risk-level sensitivity*

Fig. 9 shows that empirical under-planning tracks the chosen α closely on the disruption/recovery window, with the expected monotonic trade-off between risk and padding.

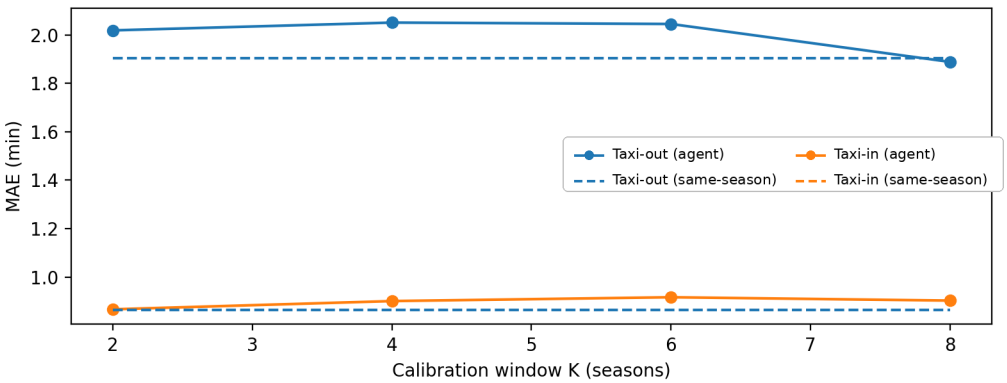
Table 12 translates the same sensitivity analysis into service-tier language. This addition is intended to make α usable when precise monetary values for delay propagation are unavailable, planners can select a conservative, balanced, or utilization-preserving posture and then verify the resulting under-planning–padding trade-off through seasonal backtesting.



(a) Under-planning versus K.



(b) Padding versus K.

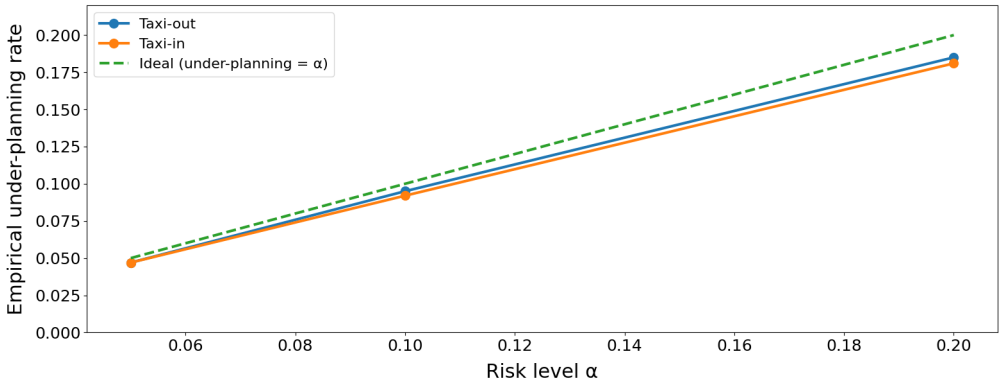


(c) Point-prediction MAE versus K.

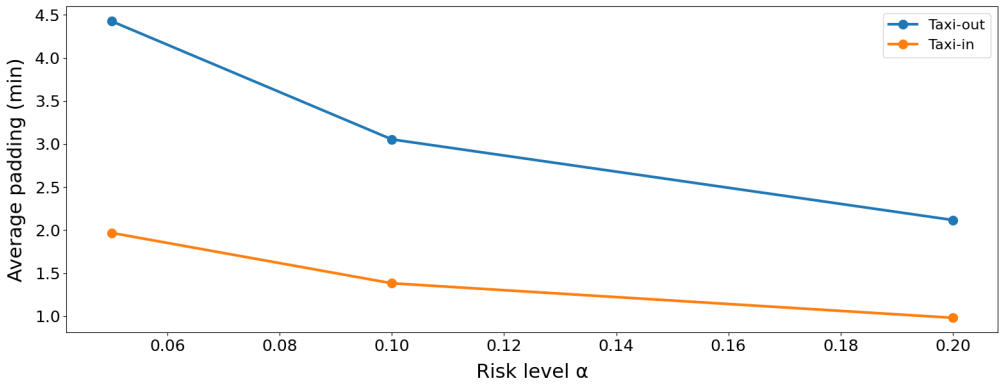
Figure 8. Disruption/recovery sensitivity to calibration-window size K ($\alpha = 0.1$). Dashed lines show same-season persistence where applicable.

7.9. Wake turbulence category (WTC) analysis

Table 11 strengthens the WTC analysis by comparing three policies on the disruption/recovery window: same-season persistence, pooled conformal calibration, and Mondrian-style WTC-stratified conformal calibration. The purpose is not to claim a per-WTC formal guarantee, but to test whether the marginal residual pool masks category-specific residual behavior. The results support that Heavy and Medium traffic should not automatically be assumed to share the same operational noise regime.



(a) Empirical under-planning versus α.



(b) Average padding versus α.

Figure 9. Risk-level sensitivity on the disruption/recovery seasons ($K = 6$). Panel (a) compares empirical under-planning with the configured risk level; panel (b) shows the induced padding.

Table 11. WTC-stratified taxi-out disruption/recovery sensitivity ($K = 6$, $\alpha = 0.1$). Pooled CP uses one residual calibration pool across WTC categories; WTC CP applies Mondrian-style WTC calibration when calibration support is sufficient and otherwise falls back to the pooled margin. Super (J) has no observations in the disruption/recovery evaluation window.

WTC	n	Under-planning			Padding (min)		
		Baseline	Pooled CP	WTC CP	Baseline	Pooled CP	WTC CP
H	148	0.466	0.101	0.068	2.155	8.796	9.783
M	248	0.528	0.024	0.060	1.504	8.285	6.765

Operationally, the WTC subset shows why stratification should be treated as a deployment safeguard rather than a cosmetic detail. Heavy traffic remains closer to the target under WTC-specific calibration, while Medium traffic receives a smaller margin than under the pooled calibration, reducing excessive padding. This supports a practical rule that considers that at mixed-traffic airports, WTC-aware calibration should be preferred when same-stratum calibration support is sufficient; otherwise, pooled or partially pooled margins should be retained to avoid unstable small-sample quantiles.

8. Extended Analyses

This section reports additional quantitative summaries that may be useful to practitioners when tuning and deploying the agent.

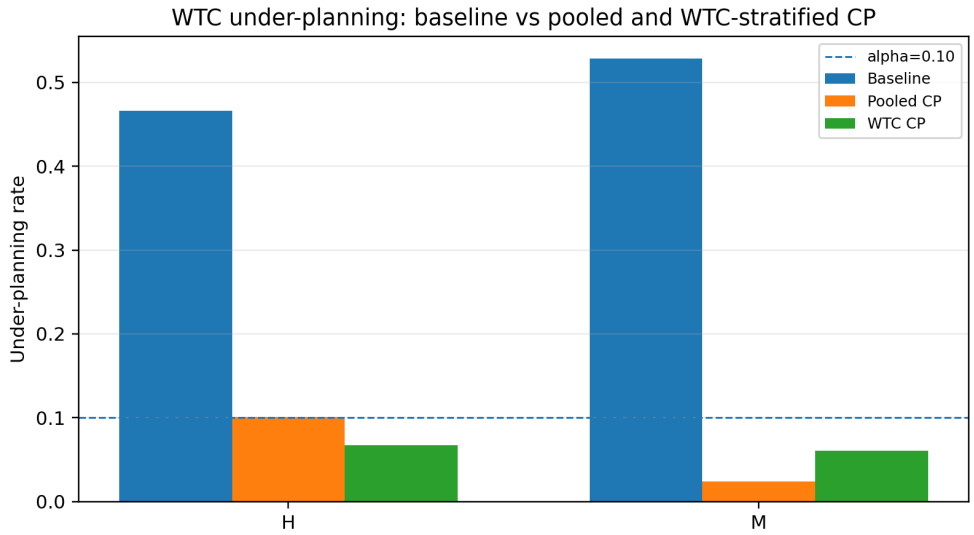


Figure 10. WTC disruption/recovery under-planning under baseline persistence, pooled conformal calibration, and WTC-stratified conformal calibration ($K = 6$, $\alpha = 0.1$).

8.1. Tabular risk–padding sensitivity to α

Panel (a) of Fig. 9 visualizes the relationship between the configured risk level and empirical under-planning. Table 12 complements the figure with both under-planning and padding.

Table 12. Risk-level sensitivity and service-tier interpretation on the disruption/recovery seasons ($K = 6$). Under is empirical under-planning; Pad is average realized surplus in minutes; $C_u/C_o = (1-\alpha)/\alpha$ is the implied under-planning-to-padding cost ratio under the simplified linear cost model.

α	Service-tier reading	C_u/C_o	Under(TXO)	Pad(TXO)	Under(TXI)	Pad(TXI)
0.05	Conservative disruption-protection	19.0	0.047	4.428	0.047	1.968
0.10	Balanced default risk-control	9.0	0.095	3.054	0.092	1.382
0.20	Utilization-preserving planning	4.0	0.185	2.117	0.181	0.981

Table 12 should be read as a practical decision aid rather than as a claim of cost optimality. The implied C_u/C_o ratio follows from the simplified linear cost model, but airline delay costs are often nonlinear and small taxi-time shortfalls may be absorbed by available slack, whereas larger shortfalls may propagate through aircraft rotations, passenger connections, and hub-bank structures. For this reason, the table provides service-level tiers that can be selected and audited even when precise cost accounting is unavailable.

8.2. Fallback and sparse-history guardrail stress test

A high-uncertainty fallback based on uncalibrated persistence would be inconsistent with the risk-averse logic of the framework, because persistence is precisely the heuristic that can fail during distribution shift. We evaluate a revised deployment guardrail. A high-margin flag is triggered when the season-level conformal margin exceeds an illustrative top-quartile threshold computed separately for TXO and TXI in the rolling backtest. For flagged cases, the conservative policy either routes the airport-season to manual review or applies a recent-worst-case overlay based on the maximum observed taxi-time over the most recent K available seasons. The overlay is deliberately monotone because it never lowers the base conformal recommendation.

The diagnostic confirms the logic of the revision. Replacing flagged high-margin cases by uncalibrated persistence increases disruption/recovery under-planning, whereas the recent-worst-case overlay preserves or slightly improves under-planning at the cost of additional padding. The result supports the revised guidance,

Table 13. High-uncertainty fallback stress test on the disruption/recovery window. The flag is an illustrative top-quartile season-level margin threshold, applied separately for TXO and TXI. The recent-worst-case rule is an overlay: it never lowers the base conformal recommendation.

Target	Policy	Δ flag threshold	Flagged share	Under rate	Padding (min)	Flagged under rate
TXO P90	Base rolling conformal agent	2.33	0.190	0.103	2.834	0.018
TXO P90	Flagged to uncalibrated persistence	2.33	0.190	0.118	2.416	0.096
TXO P90	Flagged to recent-worst-case overlay	2.33	0.190	0.102	2.999	0.013
TXI P90	Base rolling conformal agent	0.82	0.181	0.090	1.359	0.034
TXI P90	Flagged to uncalibrated persistence	0.82	0.181	0.105	1.157	0.118
TXI P90	Flagged to recent-worst-case overlay	0.82	0.181	0.089	1.391	0.027

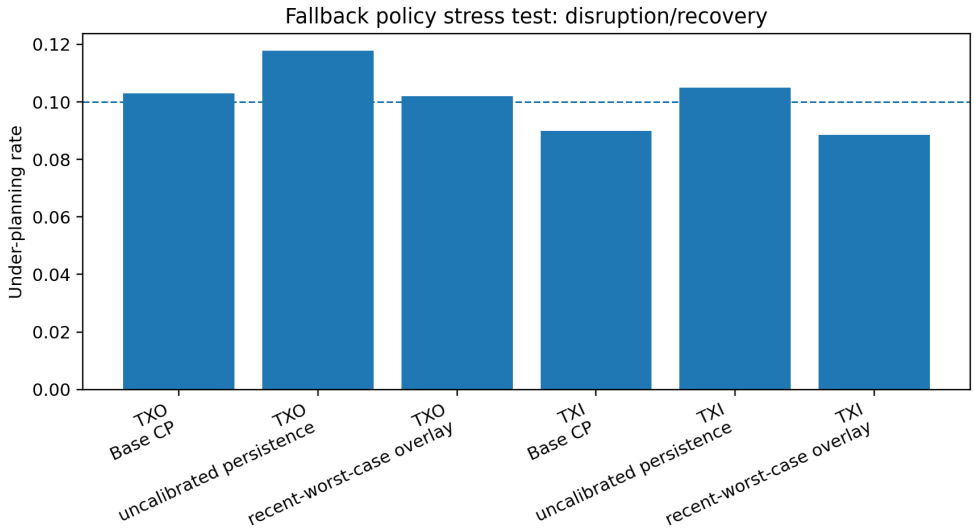


Figure 11. Fallback policy stress test on the disruption/recovery window. The figure compares the base conformal recommendation, an uncalibrated persistence fallback for flagged high-margin cases, and the recent-worst-case overlay.

persistence should not be used as the high-uncertainty fallback; flagged cases should instead trigger manual review or a conservative worst-case overlay.

Table 14. Sparse-history and cold-start diagnostic on the disruption/recovery window. The conservative overlay is applied only to cold-start records and uses a past-only cohort/season-type 90th percentile as a guardrail; established airports remain governed by the base conformal recommendation.

Target	History group	N	Base under rate	Base pad (min)	Overlay under rate	Overlay pad (min)
TXO P90	Cold start (0-2 seasons)	140	0.107	3.465	0.014	10.619
TXO P90	Limited history (3-5 seasons)	160	0.106	2.482	0.106	2.482
TXO P90	Established (6+ seasons)	2051	0.102	2.819	0.102	2.819
TXI P90	Cold start (0-2 seasons)	135	0.074	2.042	0.022	5.055
TXI P90	Limited history (3-5 seasons)	165	0.055	1.599	0.055	1.599
TXI P90	Established (6+ seasons)	2137	0.094	1.297	0.094	1.297

Sparse-history airports are handled as a separate deployment regime rather than as ordinary well-calibrated airports. Airports with at most two prior seasons are flagged as cold-start cases. For these cases, a conservative initialization can use cohort-level or season-type calibration until enough local history accumulates. Table 14 shows the expected operational trade-off, the conservative overlay strongly reduces under-planning for cold-start records,

but it can substantially increase padding. This is why the overlay is presented as a guardrail and monitoring policy, not as the default planning rule for established airports.

9. Interpretation of the Coverage Guarantee and the Conformal Margin

Conformal prediction is important for safety-critical domains because it provides a simple, testable statement, under exchangeability, the probability that the realized target is below the one-sided recommendation is at least $1-\alpha$ in finite samples [33, 34]. The proper justification of this property comes from the rank structure of the calibration residuals rather than from strong parametric assumptions. The next considerations are essential to the aviation-planning interpretation:

- **Marginal vs conditional reliability.** Conformal guarantees are marginal over the population represented by the calibration set. They do not imply identical error rates across airports, traffic strata, or operating regimes, especially when sample sizes are small. This motivates per-airport monitoring and stratified (Mondrian) calibration across airports and traffic strata when the data support meaningful groups.
- **Distribution shift breaks exchangeability.** When the data-generating process changes, the formal exchangeability-based guarantee no longer applies. Rolling calibration is an empirical adaptation mechanism because it focuses on recent residual behavior, but it does not restore distribution-free finite-sample coverage under arbitrary drift.
- **Population shift.** CODA only reports planning values for airports that meet observation thresholds, so the calibration population can change by season. Monitoring should therefore track cohort stability alongside under-planning and margin diagnostics.

The conformal margin Δ is interpretable as an empirical one-sided residual quantile and auditable as a statistical safety margin. It should not be interpreted as a causal operational explanation. A high margin indicates that recent forecasts have underestimated taxi time more often or more severely than desired and it does not identify whether the cause is weather volatility, runway configuration, ATC procedures, airport works, stand allocation, recovery dynamics, or traffic-mix changes. Causal diagnosis requires additional exogenous operational data beyond the seasonal CODA panel used in this study.

For operational deployment, the margin should be displayed together with diagnostic context, including recent residual trends, airport-level under-planning rates, cohort-stability flags, sparse-history warnings, and traffic-stratum indicators. This turns Δ into part of an auditable decision card rather than a standalone explanatory signal.

10. Discussion and Operational Implications

The goal of the proposed formulation is to move evaluation from point-forecast accuracy to decision reliability in seasonal planning. In seasonal taxi-time planning, the main failure mode is systematic under-planning: the planning value is too low to match the actual seasonal target. Such shortfalls also emerge in block-time assumptions, robustness analyses, and schedule design downstream. On the late season evaluation window, same season persistence can still achieve competitive MAE, but it is not enough to meet the operational risk target, whereas the conformal agent is able to trade off a few padding points for a substantial reduction in under-planning (Table 1).

10.1. Decision-oriented evaluation under-planning as the primary criterion

A key element to consider is that point-error metrics alone do not determine the success of the planning process. In Table 1, same-season persistence achieves competitive MAE and low padding, but its under-planning rates remain far above the nominal 10% risk target (0.355 for TXO $P90$ and 0.254 for TXI $P90$). This gap is important because seasonal planning values are *commitments* in the planning process, not just probabilistic forecasts. In contrast, the conformal agent keeps empirical under-planning close to the configured level (0.105 for TXO $P90$ and 0.100 for TXI $P90$) [27, 28, 33]. This is a practical example of a broader lesson from dataset shift and concept drift: stable average accuracy can mask substantial changes in operational risk [25, 26].

10.2. Margins as auditable statistical uncertainty, not causal diagnosis

The conformal margin Δ is the $(1-\alpha)$ quantile of recent residuals and therefore summarizes *recent volatility* in a unit that is operationally meaningful (minutes). Fig. 5 shows that Δ inflates during the most disrupted seasons and contracts as operations stabilize. Margin inflation is not a drift detector and does not identify the physical cause of the residual change. It is instead a residual-quantile monitoring signal related to the drift literature on distributional changes through performance degradation and residual dynamics [26]. In practice, we notice that margin inflation can trigger (i) domain review, (ii) feature or model upgrades, or (iii) conservative fallback policies that preserve service-level protection.

10.3. Operational tuning of α, K, and recency weighting

The method exposes three transparent tuning parameters. The first is α, which encodes the user's tolerance for under-planning: decreasing α reduces risk at the expense of higher padding (Fig. 9 and Table 12). Second, the rolling calibration window size K governs responsiveness to regime changes: smaller K adapts faster but can be noisier, while larger K is more stable and may lag after an abrupt shift (Table 10 and Fig. 8). Third, recency weighting (Table 6) provides a continuous alternative to shrinking K by emphasizing the most recent seasons [31, 34].

A reasonable default in the proposed setting is $\alpha = 0.1$ and $K \in [4, 8]$. In practice, the best decision is to select the smallest K that keeps under-planning close to the target without excessive padding inflation on stable airports, and to tighten α only when the operational cost of under-planning dominates the cost of padding. Padding is not a neutral statistical correction and additional planned taxi time can reduce aircraft utilization, consume schedule slack, and contribute to schedule creep if applied mechanically. The conformal recommendation should therefore be interpreted as decision support rather than as an automatic schedule-inflation rule.

A minimal cost model for selecting α. This study proposes a practical way to choose α which is to make the over/under asymmetry explicit. Let C_o denote the marginal cost of *one minute of scheduled padding* (lost utilization and pressure toward schedule inflation) and C_u the marginal cost of *one minute of shortfall* (reactionary delay and propagation through rotations). Under a standard piecewise-linear loss, minimizing $\mathbb{E}[C_o(b-y)_+ + C_u(y-b)_+]$ yields a quantile decision rule with $(1-\alpha) = \frac{C_u}{C_u+C_o}$ (equivalently $\alpha = \frac{C_o}{C_u+C_o}$), consistent with classic robust block-time setting insights [17, 20, 23]. Thus, the common default $\alpha = 0.1$ corresponds to $C_u \approx 9\,C_o$: under-planning is treated as roughly an order of magnitude more expensive than padding. In practice, operators can estimate a plausible C_u/C_o range from their own disruption accounting (misconnection penalties and hub-banking sensitivity) and then select α accordingly, while using the under-planning and padding sensitivity curves (Fig. 9 and Table 12) to anticipate the induced padding. When precise monetary costs are unavailable, planners can instead select α through service-level tiers, such as conservative, balanced, or utilization-preserving planning, and validate the resulting padding–risk trade-off through backtesting. The cost model is a simplified linear approximation; in real airline operations, under-planning costs may be nonlinear because small deviations can be absorbed by slack whereas larger deviations can propagate through rotations.

10.4. WTC stratification: heterogeneity, data scarcity, and pooling opportunities

The WTC subset provides an instructive view of heterogeneity, but it also illustrates the statistical cost of stratification. Within the disruption/recovery seasons, baseline persistence under-plans heavily for both Heavy and Medium categories, while conformal calibration substantially improves marginal under-planning control (Table 11). WTC-stratified calibration changes the allocation of the conformal margin across categories: it increases protection for Heavy traffic and reduces excessive padding for Medium traffic relative to the pooled margin. Figure 12 visualizes the corresponding padding trade-off.

These results justify a cautious deployment recommendation. First, *stratified (Mondrian) calibration* is desirable when operational strata such as WTC have different residual distributions [27, 33]. Second, stratification should be protected by minimum-support checks; where strata are data-poor, *partial pooling* or fallback to a pooled residual margin is safer than unstable category-specific quantiles. Finally, practitioners should remember that the CODA

Table 15. WTC calibration-support diagnostics on the disruption/recovery window. The support column reports the mean number of same-WTC calibration residuals available to compute the Mondrian margin across evaluated seasons.

WTC	n	Mean WTC cal. n	Fallback-to-pooled rate	Mean WTC margin (min)
H	148	193.8	0.000	6.288
M	248	303.2	0.000	4.085

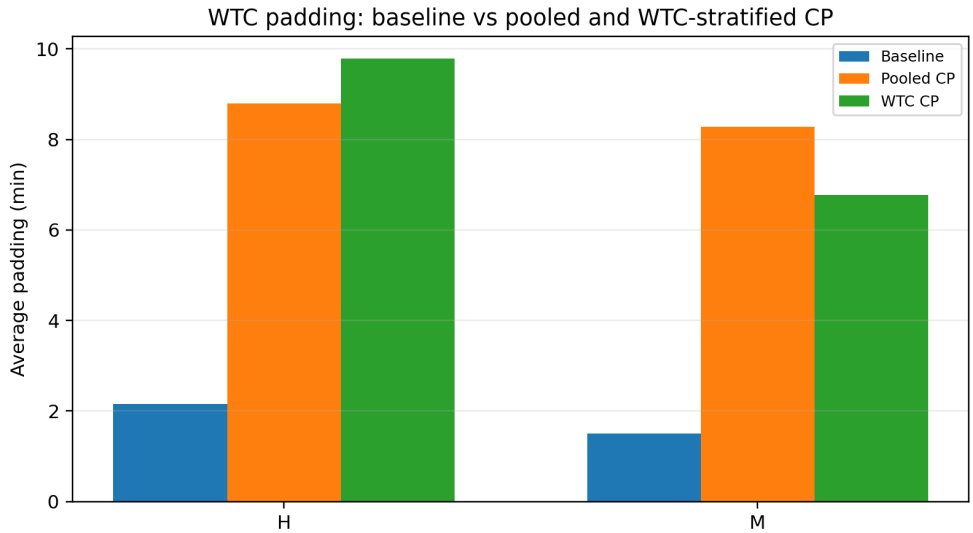


Figure 12. WTC disruption/recovery padding under baseline persistence, pooled conformal calibration, and WTC-stratified conformal calibration ($K = 6$, $\alpha = 0.1$).

WTC report excludes light aircraft and focuses on M/H/J categories [1]. In our disruption/recovery evaluation, Super (J) has no observations, so the WTC conclusions are indicative rather than definitive.

10.5. From taxi planning values to integrated schedule robustness

Taxi planning values are only one component of schedule robustness, but they are a component that is both measurable and frequently updated in airline planning cycles. The proposed agent can act as a modular "risk-controlled planning value" service for larger processes such as block-time setting, schedule design, and robustness analysis [20–22]. Because the conformal layer is model-agnostic, teams can iterate on richer covariates or stronger predictive models while keeping the risk-control contract stable. More broadly, the approach demonstrates how uncertainty quantification can be operationalized into auditable decision rules, which is increasingly important as AI components are integrated into safety- and performance-critical aviation workflows.

11. Deployment, Monitoring, and Governance Considerations

While the proposed agent is computationally efficient, deploying risk-controlled planning recommendations in operational contexts benefits from explicit monitoring and governance. The risk target considered in this study is marginal over the evaluated airport-season cohort. It does not imply that each individual airport will experience under-planning at or below α. Airports with unusual residual distributions, sparse seasonal history, systematic forecast bias, or changing traffic mix may exhibit local under-planning rates above the nominal target without contradicting the marginal cohort-level objective.

11.1. Monitoring signals

The agent produces three naturally auditable signals:

- **Realized under-planning rate.** Over time, practitioners can track the realized fraction of airports for which $y > b$ and compare it to the configured α. Persistent deviation suggests either drift beyond the rolling window's ability to adapt, or a mismatch between the forecaster's feature set and the evolving operational regime. As a reference point, for N airport-season decisions in a monitoring window, binomial variability suggests fluctuations on the order of $\sqrt{\alpha(1-\alpha)/N}$; sustained deviations beyond such bands warrant investigation.
- **Margin inflation.** The margin $\Delta = q_{1-\alpha}$ summarizes recent residual volatility. Sudden increases in Δ should be treated as monitoring signals of residual instability that may accompany drift, and can be used to trigger domain review, feature updates, or conservative fallbacks.
- **Cohort stability.** CODA reports only airports meeting minimum observation thresholds; hence the monitored cohort size N_t can change season-to-season. Sharp changes in N_t should be treated as a selection-shift trigger and interpreted alongside coverage and margin diagnostics. In deployment, a practical cohort-stability flag can be raised when the seasonal share of entering or exiting airports exceeds a pre-specified threshold, such as 20%, prompting recalibration review.

11.2. Guardrail policies

Because planning recommendations affect downstream business processes (block-time refresh, robustness checks, KPI baselines), we advocate explicit guardrails that preserve human oversight:

- **Deferral thresholds** on Δ (or on the implied padding) for airports where the agent is least certain.
- **Change limits** to prevent abrupt planning updates (cap season-to-season change) unless supported by large and persistent residual shifts.
- **Airport exception handling:** if an airport repeatedly violates the target risk (under-planning above α for multiple consecutive seasons), route it to review and apply a more conservative fallback or a dedicated calibration stratum.
- **Audit logs** capturing $(\hat{y}, \Delta, b, \alpha, K)$ for each airport-season decision, enabling post-season analysis and accountability.

The high-uncertainty fallback should be aligned with the risk-averse logic of the framework. When the uncertainty flag is triggered, the system should not revert to uncalibrated persistence, since persistence-based rules are precisely those that can fail during distribution shift. Instead, flagged cases should either be routed to manual operational review or assigned a conservative recent-worst-case recommendation, such as the maximum observed taxi time over the most recent K available seasons, subject to data-sufficiency checks. The stress test in Table 13 supports this choice: an uncalibrated persistence fallback increases under-planning, whereas the recent-worst-case overlay preserves the risk-control logic at the cost of additional padding.

11.3. Toward integrated schedule optimization

The agent outputs a controllable safety margin for a single block-time component. A natural next step is to include these margins within schedule optimization loops (bank structure design, fleet utilization constraints, passenger connection risk) where the choice of α is aligned with explicit cost models and service-level targets. This is also where multi-airport coupling and network effects come into play: while our current formulation is per airport-season, airline schedules couple uncertainty across legs and rotations, which motivates and paves the way for future network-level extensions.

12. Limitations and Threats to Validity

Some limitations of this study should be considered when it comes to the validity. CODA planning values are seasonal and do not directly encode weather, runway configuration, airport works, ATC procedures, or demand–capacity imbalance, all of which can affect taxi-time variability and the interpretation of high margins. Second, CODA uses reporting filters and minimum observation thresholds, which can change the airport cohort across seasons and create population shift. Third, airports with limited or no history require separate cold-start handling: when only one or two historical seasons are available, local rolling calibration is statistically fragile and should be

initialized using cohort-level or stratum-level calibration, conservative α settings, and mandatory local monitoring until sufficient seasonal history accumulates. Fourth, conformal guarantees are exact only under exchangeability and are marginal for the calibration population; rolling calibration does not restore finite-sample guarantees under arbitrary distribution shift. Finally, the WTC subsets cover fewer airports, do not include light aircraft, and contain no Super (J) observations in the disruption/recovery evaluation window, so WTC-stratified results should be interpreted as indicative deployment diagnostics rather than definitive category-level guarantees. Our default forecaster is deliberately simple for auditability, and future work should test richer learners and domain-informed covariates while preserving the explicit risk-control layer.

13. Implementation Details and Reproducibility

This study is intentionally designed to be reproducible with computationally modest tooling. CODA data are access-controlled; we provide the full pipeline and derived artifacts but do not redistribute the raw dataset. All reported results are produced by deterministic scripts (fixed random seeds and no per-season hyperparameter tuning loops) using rolling-origin splits to prevent look-ahead.

13.1. Preprocessing and feature construction

For each airport and season, we construct lag features using only past seasons for the same airport (or for the same (airport, WTC) pair in the WTC subset). We use one-year (same-season) lags and one-season lags, plus short rolling summaries (mean and standard deviation over the last four seasons). Missing lag values (for airports with sparse historical coverage) are imputed using median statistics computed only from the training split available before the target season. For any target season t, imputation medians are computed using seasons $s < t$ only, thereby preserving temporal ordering and preventing look-ahead leakage while allowing the forecaster to emit a point prediction and the conformal layer to calibrate a margin. Categorical variables (airport identity, season type, WTC) are one-hot encoded.

13.2. Forecaster training

We use ridge regression as the default forecaster because it is stable under multicollinearity, computationally inexpensive, and easy to audit. Numeric features are standardized to zero mean and unit variance on the training split. The regularization hyperparameter is selected for robustness (we use a fixed value rather than a per-season tuning loop to avoid leakage across time). In principle, any regression model can be substituted, including non-linear learners and hierarchical models.

13.3. Conformal calibration implementation

For each test season t, the calibration set consists of the previous K seasons. We compute signed residuals and apply the conservative conformal quantile index in (4) to obtain a single margin Δ shared across airports within the relevant calibration pool. This design keeps the recommendation statistically auditable and avoids unstable airport-specific margins when seasonal histories are short. When stratifying (by season type or WTC) or weighting residuals by recency, we use the same quantile logic within each stratum or with weighted ranks.

13.4. Recommended reporting for practitioners

In addition to the aggregate metrics in this paper, a deployment should also record and review the following diagnostics: (i) per-airport under-planning rates (to detect heterogeneity), (ii) the distribution of margins Δ (to detect volatility spikes), and (iii) the stability of recommendations for different K and α settings. Such diagnostics help ensure that the agent remains a transparent decision-support tool rather than a hidden driver of systematic schedule inflation.

14. Conclusion

This paper develops seasonal airport taxi-time planning under distribution shift as a risk-controlled statistical decision problem. The proposed study combines a pooled seasonal point forecaster with rolling one-sided conformal calibration to create recommendations of the form $b = \hat{y} + \Delta$, where Δ is an explicit residual-quantile safety margin. In the EUROCONTROL CODA application, this lightweight statistical-computing pipeline substantially reduces under-planning during disruption and recovery seasons while making the implied padding explicit and auditable.

The most significant conclusion for the aviation application is not that rolling calibration solves distribution shift in the formal sense, but that conformal uncertainty quantification can convert point forecasts into decision-oriented planning recommendations with explicit asymmetric risk control. Future work should test richer covariates, non-linear or hierarchical forecasters, stratified calibration, balanced-cohort diagnostics, and direct integration of these margins into block-time optimization and network-level schedule robustness models.

Table 16 provides a complete list of abbreviations used throughout the paper.

Table 16. Complete list of abbreviations and shorthand notations used in the paper.

Abbreviation	Meaning in this paper
A-CDM	Airport Collaborative Decision-Making.
AI	Artificial intelligence.
AIBT	Actual in-block time.
ALDT	Actual landing time.
AOBT	Actual off-block time.
AODF	Air Transport Operator Data Flow.
ARIMA	Autoregressive integrated moving average.
ATM	Air traffic management.
ATOT	Actual take-off time.
CDF	Cumulative distribution function.
CODA	Central Office for Delay Analysis (EUROCONTROL).
ETS	Error–Trend–Seasonality exponential smoothing state-space model family.
EUROCONTROL	European Organisation for the Safety of Air Navigation.
IATA	International Air Transport Association.
ICAO	International Civil Aviation Organization; the paper also uses ICAO four-letter airport identifiers (EGLL for London Heathrow).
KPI	Key performance indicator.
MAE	Mean absolute error.
P90	90th percentile planning value used for seasonal taxi-time benchmarking and decision support.
Pad	Average padding in minutes (table shorthand).
Sxx / Wxx	Generic IATA season labels for summer and winter seasons, respectively (S20, W20).
TXI	Taxi-in.
TXO	Taxi-out.
Under	Under-planning rate (table shorthand).
WTC	Wake turbulence category.
L / M / H / J	Light, Medium, Heavy, and Super wake turbulence categories, respectively.

REFERENCES

1. EUROCONTROL, "Taxi-time planning values," *Aviation Intelligence Portal (CODA)*. [Online]. Available: https://ansperformance.eu/reference/dataset/planning-taxi-times/ (accessed 6 Apr. 2026).
2. EUROCONTROL, *EUROCONTROL Specification for Operational ANS Performance Monitoring—Air Transport Operator Data Flow*, ed. 1.0, 6 Aug. 2020. Available: https://www.eurocontrol.int/publication/eurocontrol-specification-aodf.
3. Airports Council International (ACI), EUROCONTROL, and IATA, *Airport Collaborative Decision-Making (A-CDM) Implementation Manual*, ver. 5.0, 31 Mar. 2017. Available: https://www.eurocontrol.int/publication/airport-collaborative-dec

ision-making-cdm-implementation-manual.
4. EUROCONTROL, *Additional Taxi-Out Time Performance Indicator Document*, ed. 01.00, 16 Mar. 2023. Available: https://ansperformance.eu/library/ATXOT_indicator_documentation_mar23.pdf (accessed 15 Jun. 2026).
5. EUROCONTROL, *Additional Taxi-In Time Performance Indicator Document*, ed. 01.00, 16 Mar. 2023. Available: https://ansperformance.eu/library/ATXIT_indicator_documentation_mar23.pdf (accessed 15 Jun. 2026).
6. H. Idris, J.-P. Clarke, R. Bhuva, and L. Kang, "Queuing model for taxi-out time estimation," *Air Traffic Control Quarterly*, vol. 10, no. 1, pp. 1–22, 2002. doi: 10.2514/atcq.10.1.1.
7. I. Simaiakis and H. Balakrishnan, "A queuing model of the airport departure process," *Transportation Science*, vol. 50, no. 1, pp. 94–109, 2016. doi: 10.1287/trsc.2015.0603.
8. S. Badrinath, M. Z. Li, and H. Balakrishnan, "Integrated surface–airspace model of airport departures," *Journal of Guidance, Control, and Dynamics*, vol. 42, no. 5, pp. 1049–1063, 2019. doi: 10.2514/1.G003964.
9. S. Ravizza, J. Chen, J. A. D. Atkin, P. Stewart, and E. K. Burke, "Aircraft taxi time prediction: Comparisons and insights," *Applied Soft Computing*, vol. 14, pp. 397–406, 2014. doi: 10.1016/j.asoc.2013.10.004.
10. H. Lee, W. Malik, and Y. C. Jung, "Taxi-out time prediction for departures at Charlotte Airport using machine learning techniques," in *16th AIAA Aviation Technology, Integration, and Operations Conference*, 2016, AIAA 2016-3910. doi: 10.2514/6.2016-3910.
11. O. Lordan, J. M. Sallán, and M. Valenzuela-Arroyo, "Forecasting of taxi times: The case of Barcelona-El Prat airport," *Journal of Air Transport Management*, vol. 56, pp. 118–122, 2016. doi: 10.1016/j.jairtraman.2016.04.015.
12. P. Balakrishna, R. Ganesan, and L. Sherry, "Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case study of Tampa Bay departures," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 6, pp. 950–962, 2010. doi: 10.1016/j.trc.2010.03.003.
13. X. Wang, A. E. I. Brownlee, J. R. Woodward, M. Weiszer, and S. Ravizza, "Aircraft taxi time prediction: Feature importance and their implications," *Transportation Research Part C: Emerging Technologies*, vol. 124, Art. no. 102892, 2021. doi: 10.1016/j.trc.2020.102892.
14. J. Yin, M. Zhang, Y. Ma, W. Wu, H. Li, and P. Chen, "Prediction and analysis of airport surface taxi time: Classification, features, and methodology," *Applied Sciences*, vol. 14, no. 3, Art. no. 1306, 2024. doi: 10.3390/app14031306.
15. J. A. D. Atkin, E. K. Burke, and S. Ravizza, "Addressing the pushback time allocation problem at Heathrow airport," *Transportation Science*, vol. 47, no. 4, pp. 584–602, 2013. doi: 10.1287/trsc.1120.0446.
16. S. Lan, J.-P. Clarke, and C. Barnhart, "Planning for robust airline operations: Optimizing aircraft routings and flight departure times to minimize passenger disruptions," *Transportation Science*, vol. 40, no. 1, pp. 15–28, 2006. doi: 10.1287/trsc.1050.0134.
17. M. G. Sohoni, S. Bali, and R. L. Shanbhag, "Robust airline scheduling under block-time uncertainty," *Transportation Science*, vol. 45, no. 4, pp. 451–464, 2011. doi: 10.1287/TRSC.1100.0361.
18. M. Dunbar, G. Froyland, and C.-L. Wu, "Robust airline schedule planning: Minimizing propagated delay in an integrated routing and crewing framework," *Transportation Science*, vol. 46, no. 2, pp. 204–216, 2012. doi: 10.1287/trsc.1110.0395.
19. V. Chiraphadhanakul and C. Barnhart, "Robust flight schedules through slack re-allocation," *EURO Journal on Transportation and Logistics*, vol. 2, no. 4, pp. 277–306, 2013. doi: 10.1007/s13676-013-0028-y.
20. L. Hao and M. Hansen, "Block time reliability and scheduled block time setting," *Transportation Research Part B: Methodological*, vol. 69, pp. 98–111, 2014. doi: 10.1016/j.trb.2014.08.008.
21. L. Kang and M. Hansen, "Behavioral analysis of airline scheduled block time adjustment," *Transportation Research Part E: Logistics and Transportation Review*, vol. 103, pp. 56–68, 2017. doi: 10.1016/j.tre.2017.04.004.
22. T. P. C. Fan, "Schedule creep: In search of an uncongested baseline block time by examining scheduled flight block times worldwide 1986–2016," *Transportation Research Part A: Policy and Practice*, vol. 121, pp. 192–217, 2019. doi: 10.1016/j.tra.2019.01.006.
23. M. Jetzki, *The Propagation of Air Transport Delays in Europe*. RWTH Aachen University and EUROCONTROL, 2009. Available: https://www.eurocontrol.int/publication/propagation-air-transport-delays-europe.
24. Y. Su, K. Xie, H. Wang, Z. Liang, W. A. Chaovalitwongse, and P. M. Pardalos, "Airline disruption management: A review of models and solution methods," *Engineering*, vol. 7, no. 4, pp. 435–447, 2021. doi: 10.1016/j.eng.2020.08.021.
25. J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., *Dataset Shift in Machine Learning*. MIT Press, 2009. ISBN 9780262170055. Available: https://direct.mit.edu/books/edited-volume/3841/Dataset-Shift-in-Machine-Learning.
26. J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, Art. no. 44, 2014. doi: 10.1145/2523813.
27. V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.
28. G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, pp. 371–421, 2008. Available: https://jmlr.org/papers/v9/shafer08a.html.
29. J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.
30. Y. Romano, E. Patterson, and E. Candès, "Conformalized quantile regression," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019. Available: https://papers.nips.cc/paper/8613-conformalized-quantile-regression.
31. R. J. Tibshirani, R. Foygel Barber, E. J. Candès, and A. Ramdas, "Conformal prediction under covariate shift," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019. Available: https://papers.neurips.cc/paper/8522-conformal-prediction-under-covariate-shift.
32. R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *The Annals of Statistics*, vol. 49, no. 1, pp. 486–507, 2021. doi: 10.1214/20-AOS1965.
33. A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023. doi: 10.1561/2200000101.
34. M. Fontana, G. Zeni, and S. Vantini, "Conformal prediction: A unified review of theory and new challenges," *Bernoulli*, vol. 29, no. 1, pp. 1–23, 2023. doi: 10.3150/21-BEJ1447.
35. M. Zaffran, O. Feron, Y. Goude, J. Josse, and A. Dieuleveut, "Adaptive conformal predictions for time series," in *Proceedings of the 39th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 162, pp. 25834–25866, 2022.

Available: https://proceedings.mlr.press/v162/zaffran22a.html.
36. C. Xu and Y. Xie, "Conformal prediction interval for dynamic time-series," in *Proceedings of the 38th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 139, pp. 11559–11569, 2021. Available: https://proceedings.mlr.press/v139/xu21h.html.
37. I. Gibbs and E. J. Candès, "Adaptive conformal inference under distribution shift," in *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021. Available: https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html.
38. I. Gibbs and E. J. Candès, "Conformal inference for online prediction with arbitrary distribution shifts," *Journal of Machine Learning Research*, vol. 25, no. 162, pp. 1–36, 2024. Available: https://jmlr.org/papers/v25/22-1218.html.
39. R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *The Annals of Statistics*, vol. 51, no. 2, pp. 816–845, 2023. doi: 10.1214/23-AOS2276.
40. M. Cauchois, S. Gupta, A. Ali, and J. C. Duchi, "Robust validation: Confident predictions even when distributions shift," *Journal of the American Statistical Association*, vol. 119, no. 548, pp. 3033–3044, 2024. doi: 10.1080/01621459.2023.2298037.