

Decision-Calibrated Graph Learning for Airline Schedule Buffering

Mohamed Abdelmagid

Department of Management Science and Engineering
Khalifa University
Abu Dhabi, United Arab Emirates
100059835@ku.ac.ae

Abstract

Airlines require delay forecasts that translate into scheduled actions because late arrivals can propagate through aircraft rotations. This study links airport-network forecasting with flight-level schedule-buffer design. Using 2024 Bureau of Transportation Statistics on-time records for 49 continental U.S. hubs merged with hourly METAR weather, we construct a leakage-controlled airport-hour panel and represent airport dependence with geographic and scheduled-flow graphs. A spatiotemporal propagation network estimates the destination-hour arrival-delay risk, split-conformal calibration audits residual uncertainty, and a rotation-aware newsvendor rule converts the forecast into a scheduled-arrival buffer based on idle slack, primary delay, and same-tail downstream exposure. In the held-out November-December operations, STPN achieves the lowest airport-hour MAE among the recurrent and graph baselines, although the RMSE and F1 are mixed. Static conformal intervals are sharp but show undercoverage during the holiday shift; a rolling conformal variant restores near-nominal coverage. Under literature-based delay-cost assumptions and signed latent delay-exposure accounting, the policy shifts the median buffer from 9.0 to 22.9 min and reduces a retrospective proxy fleet delay-exposure cost by 16.0% across 650,091 mapped test flights, with positive savings in 33 of 36 cost scenarios.

Keywords

Flight delay prediction, graph neural networks, conformal prediction, schedule buffers, and airline operations.

1. Introduction

Airline delays are network and scheduling problems, not just prediction problems. A late arrival can consume turn slack, delay the aircraft's next departure, increase crew duty pressure, and break passenger connections several legs later. Analytical and empirical studies have shown that propagated delays can dominate the original disturbance when airport networks are dense and aircraft rotations are tight (Brueckner et al. 2021; Pyrgiotis et al. 2013). Airline schedules deliberately include published buffer time: scheduled block times are longer than typical gate-to-gate elapsed times so that stochastic disruptions can be absorbed before they become reported arrival delays or downstream disruptions (Abdelghany et al. 2024; Wu 2005).

This creates a decision problem that accuracy metrics alone do not answer: when a forecast signals that a destination hour is risky, how should the scheduled-arrival buffer of a particular flight change? Recent graph-learning studies forecast delay over airport networks (Bao et al. 2021; Cai et al. 2022; Li et al. 2024a; Wu et al. 2024; Zheng et al. 2024), but their evaluations usually stop at MAE, RMSE, AUC, or F1. Schedule buffer models optimize reliability but often use historical or unconditional delay distributions. The missing link is a calibrated mapping from the conditional network risk to the flight-level buffer economics.

We studied this link by separating the method into forecast, calibration, and decision layers. The forecast layer estimates the airport-hour arrival delay using an STPN architecture. The calibration layer measures the residual

uncertainty and exposes coverage errors under temporal shifts. The decision layer converts each destination-hour forecast into a flight-specific scheduled-arrival buffer whose service level changes with the idle-time cost, primary-delay cost, and downstream same-tail propagation exposure.

1.1 Objectives and Contributions

This study asks how a calibrated airport-network forecast can be converted into a flight-level buffer decision that reduces the realized delay-exposure cost. The objectives are to construct a reproducible BTS-METAR airport-hour panel, benchmark network forecasting under chronological splits, measure calibration under temporal shift, and evaluate the resulting buffer policy under transparent cost assumptions. This study makes three contributions. First, network delay forecasting is reframed as a decision problem: forecasts are evaluated not only by MAE, RMSE, AUC, or F1, but also by the cost of the schedule-buffer policies they induce. Second, it builds a leakage-controlled 2024 BTS-METAR airport-hour panel for 49 U.S. hubs and uses it to benchmark spatiotemporal graph learning under chronological train, validation, and test splits. Third, it introduces a corrected latent delay exposure audit and a rotation-aware buffer rule, allowing destination-hour forecasts to be translated into flight-specific schedule-buffer recommendations while accounting for idle slack, primary delay, and downstream same-tail propagation costs. This combination connects graph forecasting, calibrated uncertainty, and schedule-buffer economics within a single evaluation pipeline.

2. Literature Review

Prior graph-learning studies have mainly addressed the question of how well delays can be forecasted over an airport network. Cai et al. (2022) model time-evolving graphs for flight-delay prediction, and Bao et al. (2021) use graph-to-sequence attention for multi-step network forecasts. Wu et al. (2024) propose spatio-temporal propagation learning with multiple airport graphs and temporal attention. More recent architectures have added adaptive external-impact, gated, or coupled-attention mechanisms to represent spatial-temporal dependence (Li et al. 2024a; Zheng et al. 2024). These models improve network state estimation and can detect conditional airport-hour risk; however, their evaluations generally emphasize point accuracy or classification metrics such as MAE, RMSE, AUC, and F1. They rarely ask how a forecast should change the schedule-buffer decision for a particular flight. Deterministic and explainable machine learning studies offer additional predictors and diagnostics, but they also tend to stop at prediction performance rather than schedule-buffer economics (Choi et al. 2016; Pineda-Jaramillo et al. 2024).

Prior schedule reliability studies examine how slack should trade off efficiency and reliability. Pyrgiotis et al. (2013) modeled delay propagation analytically, Du et al. (2018) inferred delay-causality networks, and Li et al. (2024b) reviewed data sources, propagation mechanisms, and modeling opportunities. On the decision side, Wu (2005) established the trade-off between schedule efficiency and reliability, Brueckner et al. (2021) analyzed propagated-delay mitigation through buffers, and Abdelghany et al. (2024) optimized flight block times in a data-driven schedule-level framework. This stream shows why buffer decisions matter, but it usually relies on historical or unconditional delay distributions, rather than calibrated network-state forecasts. As a result, schedule-buffer optimization and graph-based delay forecasting remain only partially connected: one stream produces conditional forecasts without costly flight-level decisions, while the other optimizes slack without fully using conditional airport-network information.

The third stream addresses forecast uncertainty. Conformal prediction provides distribution-free marginal coverage under exchangeability for arbitrary predictors (Angelopoulos and Bates 2021; Vovk et al. 2005), and weighted conformal prediction extends this idea to covariate shifts (Tibshirani et al. 2019). These tools are relevant because a buffer decision depends on tail risk, not just on the conditional mean. Recent reviews of flight-delay prediction also identified calibrated uncertainty and repeatable evaluation as open needs (Wandelt et al. 2025). However, uncertainty estimates are rarely translated into a flight-level schedule buffer rule or evaluated through realized operating costs.

This study connects these three research streams by treating forecasts as an input to a decision rather than as a final output. A destination-hour graph forecast is calibrated, mapped to individual flights, adjusted for downstream same-tail rotation exposure, and evaluated through a corrected latent delay-exposure audit that avoids double-counting the buffer already embedded in reported arrival-delay minutes. Therefore, this contribution is not another isolated forecasting comparison; it is a decision-calibrated link between airport-network prediction, uncertainty measurement, and airline schedule-buffer economics. This positioning addresses the combined question that remains underdeveloped in prior work: how should a conditional network forecast change a flight-level-schedule-buffer decision?

3. Methods

The method comprises three separable layers: forecast, calibration, and decision. Figure 1 shows the flow from public flight weather data to a flight-level buffer recommendation and a realized cost audit. This separation is important because the forecasting model can be replaced without changing the buffer rule, and the cost rule can be updated without retraining the graph model. BTS on-time records and METAR weather produce a leakage-controlled airport-hour panel and two airport graphs; STPN estimates destination-hour mean delay; split-conformal calibration measures residual uncertainty; and the decision layer converts the calibrated forecast into a rotation-aware buffer β^* .

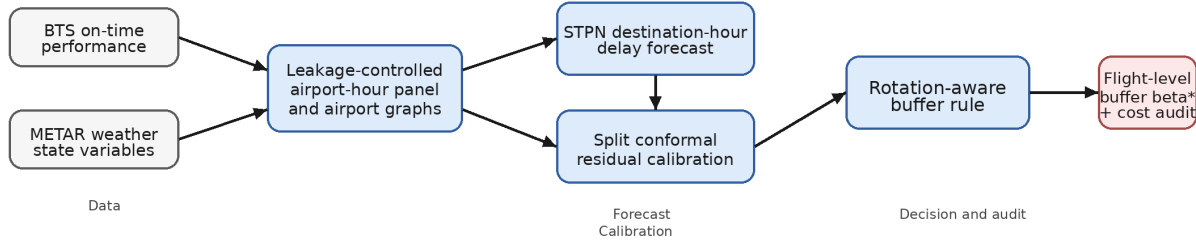


Figure 1. Decision-calibrated workflow from public flight-weather data to flight-level buffer recommendations and realized latent delay-exposure cost audit.

3.1 Spatio-temporal forecast model

Let the network comprise N airports. Time is discretized into hourly buckets, and $y_{i,t}$ denotes the mean non-negative arrival delay of flights arriving at airport i during hour t . The model consumes a lookback tensor $X_{t-L+1:t}$ whose features include the lagged mean delay, scheduled flight count, temperature, visibility, wind speed, hour, day of the week, and month. Two fixed airport graphs enter the spatial convolution: a geographic graph based on a Gaussian kernel over the great-circle distance, sparsified to the eight nearest neighbors, and a scheduled-flow graph based on row-normalized training-period flight counts between airports. The same-tail rotation is reserved for the decision layer through the downstream count k_f rather than being used as a third spatial convolution relation.

We adapt the STPN architecture by applying graph propagation separately to the geographic and scheduled-flow graphs and then aggregating the resulting airport states with a temporal-attention module. For the lookback step s and graph $g \in \{\text{geo}, \text{flow}\}$, the spatial propagation is written as

$$h_{i,s}^{(g)} = \text{ReLU} \left(\sum_j A_{ij}^{(g)} x_{j,s} W_g \right), \quad g \in \{\text{geo}, \text{flow}\}.$$

The model concatenates the propagated states with local airport features and applies temporal attention over the lookback window as follows:

$$z_{i,s} = \text{concat} \left(h_{i,s}^{(\text{geo})}, h_{i,s}^{(\text{flow})}, x_{i,s} \right), \quad \hat{y}_{i,t+1} = f_{\text{attn}}(z_{i,t-L+1}, \dots, z_{i,t}).$$

Four temporal attention heads were used. The mean head was trained using L_1 loss. The scale head is retained only as a diagnostic; the interval widths used in the decision layer originate from conformal residuals; therefore, the forecast model is not treated as a fully specified Gaussian likelihood.

3.2 Split-conformal calibration

For the validation cells $c = (i, t)$, the absolute residual is defined as follows:

$$r_c = |y_c - \hat{y}_c|.$$

For nominal miscoverage α and n validation cells, the split-conformal half-width is an empirical order statistic:

$$q_\alpha = r_{((n+1)(1-\alpha))},$$

where $r_{(k)}$ is the k -th smallest value in the set of validation residuals $\{r_c: c \in \mathcal{V}\}$. The resulting airport-hour interval is

$$C_\alpha(i, t) = [\hat{y}_{i,t} - q_\alpha, \hat{y}_{i,t} + q_\alpha].$$

Under exchangeability between the validation and test cells, the interval has marginal finite-sample coverage. Because the test period contains November-December holiday operations after September-October validation, this guarantee is not asserted for the shifted test period. Instead, the empirical test coverage is reported as a stress test of the calibration layer.

3.3 Flight-level buffer policy

Flight f has origin o_f , destination d_f , scheduled block time b_f^{sched} , and route-median actual elapsed time m_{o_f, d_f} . The published arrival buffer is as follows:

$$\beta_f^{\text{obs}} = \max\{0, b_f^{\text{sched}} - m_{o_f, d_f}\}.$$

BTS signed arrival delay D_f^{obs} is already measured relative to the published schedule. A nonnegative measure would discard the early arrival information that a counterfactual needs, and subtracting a proposed buffer from it would double-count the observed buffer. Therefore, the policy audit uses a signed latent delay-exposure proxy:

$$Z_f = D_f^{\text{obs}} + \beta_f^{\text{obs}}.$$

In the cost audit this exposure is floored at zero, $\max(Z_f, 0)$, so unused scheduled slack is never credited as negative delay; this clipping leaves the headline saving unchanged at 16.0%. This quantity approximates the delay exposure that a scheduled-arrival buffer must absorb relative to the route-typical block-time baseline. For each test flight, the destination-hour STPN forecast provided the conditional mean delay. The latent exposure mean is approximated as

$$\mu_f^Z = \hat{y}_{d_f, t_f} + \beta_f^{\text{obs}},$$

where t_f is the scheduled destination-arrival hour. Split-conformal residuals provide an empirical scale for diagnostics. For the closed-form newsvendor rule, the Gaussian-equivalent scale implied by the calibrated 80% interval is as follows:

$$\sigma_f = \frac{q_{0.20}}{1.2816}.$$

Therefore, the conformal layer is interpreted as calibration evidence and a transparent scale input, not as a distribution-free guarantee under the shifted test period. Let the latent exposure approximation be

$$X_f \sim N(\mu_f^Z, \sigma_f^2).$$

With candidate buffer β idle-slack cost c_{idle} primary-delay cost c_{delay} downstream propagation cost c_{prop} and k_f same-tail downstream flights within 24 h, the expected cost is:

$$E[C_f(\beta)] = c_{\text{idle}}E[(\beta - X_f)_+] + (c_{\text{delay}} + k_f c_{\text{prop}})E[(X_f - \beta)_+].$$

The marginal benefit of an extra buffer minute is the avoided unabsorbed delay cost, whereas the marginal cost is the unused scheduled slack. Therefore, the service level used by the buffer rule is

$$\tau_f = \frac{c_{\text{delay}} + k_f c_{\text{prop}}}{c_{\text{idle}} + c_{\text{delay}} + k_f c_{\text{prop}}}.$$

For the Gaussian exposure approximation, the closed-form recommendation is as follows

$$\beta_f^* = \max\{0, \mu_f^Z + \sigma_f \Phi^{-1}(\tau_f)\}.$$

The default assumptions were 50 EUR/min for idle slack, 100 EUR/min for primary delay, and 80 EUR/min for downstream propagation. They are literature-based delay cost values rather than carrier-specific accounting estimates. Sensitivity is evaluated over idle-slack costs {30,50,75,100}, primary-delay costs {75,100,150}, and downstream propagation costs {50,80,120}, using ranges commonly reported in airline delay-cost studies (Cook and Tanner 2015; Malavolti-Grimal and Lesgourgues 2023; Rosenow et al. 2020; Scheelhaase et al. 2023). The Gaussian closed form is a transparent, pragmatic approximation; replacing it with the empirical heavy-tailed exposure-residual quantile leaves the headline essentially unchanged (17.1% versus 16.0% proxy saving), so the result is not an artifact of the Gaussian assumption.

4. Data Collection

Flight data were obtained from the Bureau of Transportation Statistics Reporting Carrier On-Time Performance database, which contains scheduled and actual gate times, elapsed times, delay minutes, carrier identifiers, tail numbers, cancellation and diversion flags, and delay-cause fields (Bureau of Transportation Statistics 2024). Weather data were obtained from the Iowa Environmental Mesonet ASOS/AWOS/METAR archive, which provides hourly airport observations such as temperature, wind, visibility, pressure, and precipitation (Iowa Environmental Mesonet 2024). The working scope contains the FAA Core 30 plus high-traffic secondary continental U.S. hubs for 49 airports in total. Honolulu was excluded because it is geographically isolated from the continental propagation networks.

Preprocessing localizes BTS local time fields to airport time zones, handles daylight-saving exceptions, merges origin and destination weather by scheduled hour, removes canceled and diverted records, excludes post-event variables from predictors, and splits the panel chronologically. Months 1-8 were used for training, months 9-10 for validation and conformal calibration, and months 11-12 for test evaluation. Weather variables are observed airport-hour state variables in this retrospective decision audit; a prospective tactical deployment would replace them with weather forecasts, whereas strategic schedule design would use climatological risk features. Route-level baseline elapsed times and airport graphs were treated as pre-test statistics to avoid future leakage in the policy audit. Table 1 summarizes the dataset after filtering and mapping.

Table 1. Dataset scope and temporal split.

Metric	Value
Airports in scope	49
Flight records in 49-airport scope	4,001,199
Non-canceled, non-diverted modeled flights	3,937,321
Distinct aircraft tail numbers	6,014
Reporting carriers	16
Canceled / diverted records excluded	1.36% / 0.24%
Weather coverage at origin and destination	99.5% / 99.5%
Train / validation / test modeled flights	2,614,408 / 669,479 / 653,434
Mapped test flights with STPN destination-hour forecast	650,091

5. Results and Discussion

5.1 Numerical Results

Tables 2 and 3 separate the two prediction targets to avoid overstating the cross-target comparisons. XGBoost and LightGBM are flight-level diagnostic baselines that predict individual arrival delay minutes. LSTM, GCN, and STPN are airport-hour models that predict the destination-hour mean arrival delay and feed the decision layer. Therefore, forecast accuracy is treated as a validity check for the decision experiment, not as the final evaluation criterion.

Table 2. Flight-level diagnostic baselines for validation and test splits

Model	Split	MAE (min)	RMSE	AUC	F1
XGBoost	validation	13.46	35.70	0.775	0.460
XGBoost	test	15.11	40.19	0.785	0.514
LightGBM	validation	8.62	36.92	0.743	0.429
LightGBM	test	10.85	42.47	0.764	0.489

Table 3. Airport-hour forecasting models used in the decision layer

Model	Split	MAE (min)	RMSE	AUC	F1
LSTM	validation	6.86	17.85	0.707	0.254
LSTM	test	8.53	19.37	0.743	0.318
GCN	validation	7.04	17.95	0.715	0.252
GCN	test	8.99	19.15	0.758	0.507
STPN	validation	6.75	18.08	0.716	0.205
STPN	test	8.44	19.16	0.764	0.400

Within the airport-hour cohort, the STPN had the lowest validation and test MAE. The improvement over LSTM is modest: 1.6% on validation and 1.1% on testing. A paired bootstrap over common airport-hour cells confirms that the test-MAE gap over LSTM is statistically significant but operationally small (-0.08 min, 95% CI [-0.11, -0.05]), while the gap over GCN is larger (-0.56 min); therefore, the contribution rests on the decision pipeline rather than forecast dominance. Other metrics are mixed; GCN has a slightly lower test RMSE and higher test F1, while STPN has the highest test AUC among the airport-hour models. The forecast is competitive enough to support the decision audit, but the paper does not rely on a claim of universal point forecast dominance.

Tables 4 and 5 extend the numerical evaluation from prediction accuracy to calibration and buffer-policy values.

Table 4 reports the split-conformal intervals. STPN had the narrowest 80% interval, with an 8.19-minute half-width and 16.38-minute full width, but the test coverage was 0.749 rather than the nominal 0.800. The 95% interval also fell short of nominal coverage, with 0.926 empirical coverage. This result is consistent with the validation-to-test shift: calibration residuals come from September-October, while test flights include Thanksgiving and December holiday peaks. Undercoverage is a useful deployment finding: static calibration should be stress-tested under temporal shifts rather than assumed reliable. A rolling (online) conformal variant that re-estimates the residual quantile on a trailing 1,000-cell window restores test coverage to 0.797, close to the 0.800 nominal level, versus 0.749 for the static interval, showing that the undercoverage is a correctable consequence of distribution shift rather than a failure of the conformal layer.

Table 4. Split-conformal intervals for the test split

Model	q0.20	80% width	80% coverage	q0.05	95% width	95% coverage
LSTM	8.47	16.94	0.752	22.51	45.02	0.917
GCN	8.90	17.80	0.729	23.05	46.11	0.933
STPN	8.19	16.38	0.749	22.88	45.75	0.926

The main empirical result is the decision value. At default costs, the policy recommends a median scheduled-arrival buffer of 22.9 min compared with 9.0 min in the observed schedule. The retrospective proxy cost saving under the

signed latent delay-exposure accounting is 16.0%. The model-implied expected saving is larger because the shifted test period is heavier-tailed than the validation-calibrated forecast; therefore, the realized out-of-sample number is the primary policy result. This 16.0% cost reduction, rather than the 1.1% MAE gain alone, is the strongest evidence that calibrated graph forecasts can support schedule-buffer planning (Table 5).

Table 5. Policy results at default literature-based costs ($c_{idle} = 50$, $c_{delay} = 100$, and $c_{prop} = 80$ EUR/min)

Metric	Value
Mapped test flights evaluated	650,091
Median observed schedule buffer	9.0 min
Median recommended buffer (β^*)	22.9 min
Median buffer change	+13.4 min
Model-implied expected fleet-cost saving	+61.3%
Retrospective proxy cost saving (signed exposure)	+16.0%
Sensitivity cells with positive realized saving	33 of 36
Sensitivity range / median saving	-2.6% to +30.2% / +13.7%

To test whether the network forecast drives savings rather than the newsvendor mechanism alone, the same audit is applied under five buffer policies (Table 6). All conditional model policies clustered near 16%: STPN 16.0%, LSTM 16.5%, and GCN 16.4%. An unconditional historical newsvendor that ignores the network state reaches only 12.9% and requires a far larger median buffer (34.6 versus approximately 23 min). The observed schedule was the 0% reference. Two conclusions follow: the decision pipeline adds value over historical buffering, and the savings are robust to the choice of forecaster. Thus, the contribution is the integrated forecast-to-buffer mapping rather than the STPN forecast dominance.

Table 6. Decision baseline comparison at default literature-based costs. All policies are audited on the same signed latent delay-exposure proxy

Buffer policy	Median buffer (min)	Proxy cost saving (%)
Observed schedule (reference)	9.0	0.0
Historical route-hour newsvendor	34.6	12.9
LSTM-policy	22.4	16.5
GCN-policy	26.0	16.4
STPN-policy (default)	22.9	16.0

5.2 Graphical Results

Figure 2 shows a one-week excerpt from Atlanta from the November-December test period. The forecast tracked the low-to-moderate delay regime and flagged an elevated risk through the interval band. Several holiday-period spikes exceed the 80% interval, matching the aggregate undercoverage in Table 4 and motivating the rolling or shift-aware calibration in deployment.

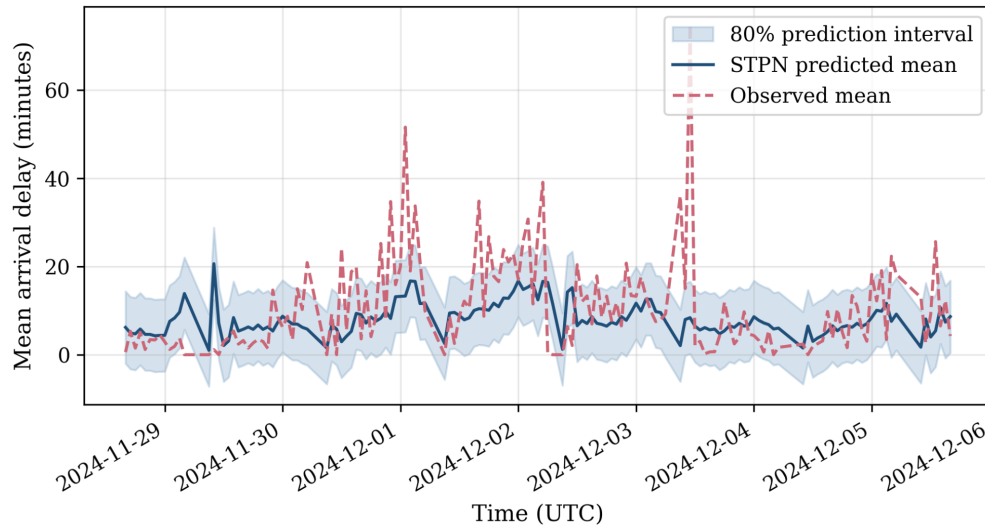


Figure 2. One-week STPN forecast excerpt for Atlanta, Georgia. The solid line is the predicted mean, the shaded region is the 80% conformal interval, and the dashed line is the observed airport hour mean delay.

Figure 3 compares the observed and recommended buffers on the mapped test flights. The shift is a cost-driven consequence of the default policy, not an arbitrary increase in slack: a median test flight has approximately three same-tail downstream flights within the 24-hour rotation window; therefore, the newsvendor rule values avoided propagated delay more than idle scheduled slack.

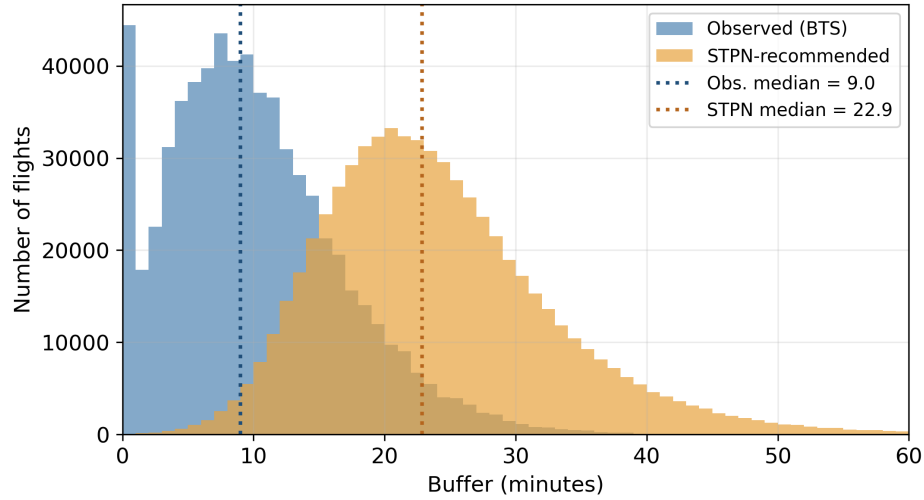


Figure 3. Observed and recommended buffer distributions for the 650,091-flight test set. The corrected policy shifts the median from 9.0 to 22.9 min under the default literature-based cost ratio

Figure 4 shows that the default result is not an isolated-parameter choice. Savings were positive in 33 of the 36 cost cells. The three negative cells occur when the idle slack is the most expensive, the propagation cost is the lowest, and the policy has little reason to protect downstream rotations. Savings increase with downstream propagation costs because rotation-aware buffers become more valuable when downstream delays are costly.

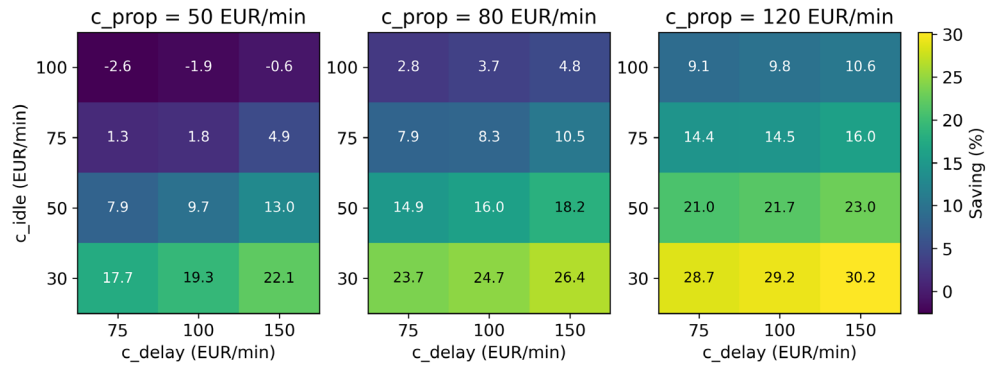


Figure 4. Retrospective proxy cost savings across cost settings under signed latent delay-exposure accounting. Rows vary c_{idle} , columns vary c_{delay} and panels vary c_{prop}

5.3 Proposed Improvements

The policy is a flight-level planning signal that can feed a full timetable optimizer, not a complete timetable reconstruction. Three improvements would move it closer to deployment: rolling or weighted conformal calibration during holiday or weather-shifted periods (Tibshirani et al. 2019), airline-specific cost coefficients when carrier accounting data are available, and a rotation-level optimization layer that enforces crew duty, maintenance, passenger connections, and minimum-turn constraints. These upgrades preserve the data interface because the forecaster, calibrator, and optimizer can be improved while retaining the flight-level recommended buffer as input.

5.4 Validation

Table 7 summarizes the checks used to reduce the leakage and audit errors. The most important correction is the latent exposure definition $Z_f = D_f^{obs} + \beta_f^{obs}$. Without this correction, a proposed buffer would be subtracted from delay minutes that were already measured relative to the published schedule, which would double the benefit of the observed schedule buffer.

Table 7. Validation and audit controls

Risk	Control	Audit result
Future leakage	Chronological split by month	Months 1-8 train, 9-10 validation, 11-12 test
Post-event leakage	Exclude realized operations variables	Actual departure, taxi, wheels-off/on, and actual elapsed time excluded
Buffer double-counting	Use signed latent exposure	Counterfactual proxy audit; early arrivals no longer inflate exposure
Output-to-feature mismatch	Independent matching of STPN cells to test flights	Of 653,434 test flights, 650,091 matched an STPN destination-hour forecast and were all evaluated; unmatched flights were excluded before the audit, so no flights were dropped at the audit stage.
Temporal distribution shift	Report empirical coverage on test period	Holiday-shift undercoverage reported directly
Baseline leakage	Route medians from training months only	Typical route elapsed times use months 1-8 only

6. Conclusion

This study links calibrated airport network delay forecasting to flight-level airline schedule-buffer decisions. Using a 2024 BTS-METAR panel for 49 continental U.S. hubs, the study benchmarks spatiotemporal graph forecasting under chronological splits, measures residual uncertainty with split-conformal calibration, and evaluates the schedule-buffer policy induced by the forecast. Together, these results address the objectives in Section 1.1 by constructing the public data panel, validating the forecasting layer, assessing the calibration under temporal shift, and testing a costed buffer decision. The central contribution is a corrected decision audit that measures the realized latent delay exposure, avoiding double-counting of the buffer already embedded in reported arrival-delay minutes.

The results show why decision evaluation is important. The STPN achieved the lowest airport-hour MAE among the recurrent and graph baselines, but the prediction gains were modest and not uniform across all metrics. Static conformal calibration produces sharp intervals but falls short of nominal coverage during the November-December holiday shift, indicating that deployment should use rolling or shift-aware calibration. When the calibrated forecast is converted into a rotation-aware buffer quantile, the resulting policy reduces the retrospective proxy delay-exposure cost by 16.0% across 650,091 mapped test flights and remains positive in 33 of 36 cost scenarios.

The practical implication is that schedule buffers should not be assigned only based on route averages or unconditional historical delay distributions. They should respond to conditional network risks and downstream rotation exposure of the aircraft. The methodological implication is broader: delay-forecasting models should be judged not only by prediction error but also by the quality of the operational decisions they enable. Future work should embed flight-level recommendations in a full timetable optimizer and replace retrospective METAR observations with planning-horizon weather forecasts or climatological risk features. All model and training parameters are reported in Appendix A. The code and processed audit files are available from the author and will be released in a public repository upon acceptance, subject to source data licensing and redistribution constraints.

Appendix A. Reproducibility Details

All models used seed 42 and the chronological split (train months 1-8, validation/calibration 9-10, test 11-12). The key settings are listed in Table 8.

Table 8. Model, calibration, and policy settings for reproducibility

Component	Setting
Common	Seed 42; chronological split; 49 airports; hourly airport-hour panel; missing predictors imputed with training-set medians
Airport graphs	Geographic: Gaussian kernel (bandwidth 1500 km), 8 nearest neighbors; scheduled-flow: row-normalized training-period flight counts
STPN	12-hour lookback; 4 temporal-attention heads; hidden size 64; attention dropout 0.1; Adam, learning rate 1e-3; 30 epochs; batch 32; L1 loss on the mean head (log-sigma head clamped to [-3,3], diagnostic only)
LSTM	2 layers; hidden size 64; dropout 0.2; Adam, learning rate 1e-3; 30 epochs; batch 256
GCN	2 graph-convolution layers; hidden size 64; Adam, learning rate 1e-3; 30 epochs
Classification metrics	AUC and F1 for the ≥ 15 -minute late class; F1 at the validation-optimal probability threshold
Conformal calibration	Split conformal with $(n+1)/n$ finite-sample correction; absolute-residual quantiles at $\alpha=0.20$ and 0.05 on all September-October validation airport-hour cells
Buffer policy	Rotation window 24 h for downstream count n_r ; default costs $c_{idle}=50$, $c_{delay}=100$, $c_{prop}=80$ EUR/min; Gaussian-equivalent scale from the calibrated 80% interval
Robustness	Rolling/online conformal (trailing 1,000-cell window); empirical heavy-tailed exposure-residual quantile buffer; paired bootstrap (2,000 resamples) on airport-hour MAE

References

- Abdelghany, A., Abdelghany, K., and Guzhva, V.S., Schedule-level optimization of flight block times for improved airline schedule planning: A data-driven approach, *Journal of Air Transport Management*, vol. 115, 102535, 2024.
- Angelopoulos, A.N., and Bates, S., A gentle introduction to conformal prediction and distribution-free uncertainty quantification, *arXiv preprint arXiv:2107.07511*, 2021.
- Bao, J., Yang, Z., and Zeng, W., Graph to sequence learning with attention mechanism for network-wide multi-step-ahead flight delay prediction, *Transportation Research Part C: Emerging Technologies*, vol. 130, 103323, 2021.
- Brueckner, J.K., Czerny, A.I., and Gaggero, A.A., Airline mitigation of propagated delays via schedule buffers: Theory and empirics, *Transportation Research Part E: Logistics and Transportation Review*, vol. 150, 102333, 2021.
- Bureau of Transportation Statistics, Reporting Carrier On-Time Performance database, U.S. Department of Transportation, 2024. Available: https://www.transtats.bts.gov/Tables.asp?DB_ID=120, Accessed: May 2025.
- Cai, K., Li, Y., Fang, Y.-P., and Zhu, Y., A deep learning approach for flight delay prediction through time-evolving graphs, *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11397-11407, 2022.
- Choi, S., Kim, Y.J., Briceno, S., and Mavris, D., Prediction of weather-induced airline delays based on machine learning algorithms, *IEEE/AIAA Digital Avionics Systems Conference (DASC)*, 2016.
- Cook, A., and Tanner, G., European airline delay cost reference values, *EUROCONTROL Performance Review Unit*, 2015.

- Du, W.-B., Zhang, M.-Y., Zhang, Y., Cao, X.-B., and Zhang, J., Delay causality network in air transport systems, *Transportation Research Part E: Logistics and Transportation Review*, vol. 118, pp. 466-476, 2018.
- Iowa Environmental Mesonet, ASOS/AWOS/METAR data download, Iowa State University, 2024. Available: <https://mesonet.agron.iastate.edu/request/download.phtml>, Accessed: May 2025.
- Li, C., Mao, J., Li, L., Wu, J., Zhang, L., Zhu, J., and Pan, Z., Flight delay propagation modeling: Data, methods, and future opportunities, *Transportation Research Part E: Logistics and Transportation Review*, vol. 185, 103525, 2024b.
- Li, C., Qi, X., Yang, Y., Zeng, Z., Zhang, L., and Mao, J., FAST-CA: Fusion-based adaptive spatial-temporal learning with coupled attention for airport network delay propagation prediction, *Information Fusion*, vol. 107, 102326, 2024a.
- Malavolti-Grimal, E., and Lesgourgues, A., Social cost of airline delays: Assessment by the use of revenue management data, *Transportation Research Part A: Policy and Practice*, vol. 170, 103613, 2023.
- Pineda-Jaramillo, J., Munoz, C., Mesa-Arango, R., Gonzalez-Calderon, C., and Lange, A., Integrating multiple data sources for improved flight delay prediction using explainable machine learning, *Research in Transportation Business & Management*, vol. 56, 101161, 2024.
- Pyrgiotis, N., Malone, K.M., and Odoni, A.R., Modelling delay propagation within an airport network, *Transportation Research Part C: Emerging Technologies*, vol. 27, pp. 60-75, 2013.
- Rosenow, J., Michling, P., Schultz, M., and Schoenberger, J., Evaluation of strategies to reduce the cost impacts of flight delays on total network costs, *Aerospace*, vol. 7, no. 11, 165, 2020.
- Scheelhaase, J., Braun, M., Maertens, S., and Grimme, W., Costs for passengers and airlines due to the significant delays and other irregularities at European airports in the 2022 summer season, *Transportation Research Procedia*, vol. 75, pp. 96-105, 2023.
- Tibshirani, R.J., Foygel Barber, R., Candès, E.J., and Ramdas, A., Conformal prediction under covariate shift, *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- Vovk, V., Gammerman, A., and Shafer, G., *Algorithmic Learning in a Random World*, Springer, 2005.
- Wandelt, S., Chen, X., and Sun, X., Flight delay prediction: A dissecting review of recent studies using machine learning, *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 4, pp. 4283-4297, 2025.
- Wu, C.-L., Inherent delays and operational reliability of airline schedules, *Journal of Air Transport Management*, vol. 11, pp. 273-282, 2005.
- Wu, Y., Yang, H., Lin, Y., and Liu, H., Spatiotemporal propagation learning for network-wide flight delay prediction, *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 1, pp. 386-400, 2024.
- Zheng, H., Wang, Z., Zheng, C., Wang, Y., Fan, X., Cong, W., and Hu, M., A graph multi-attention network for predicting airport delays, *Transportation Research Part E: Logistics and Transportation Review*, vol. 181, 103375, 2024.

Biography

Mohamed Abdelmagid received the M.Sc. degree in Engineering Systems and Management from Khalifa University, Abu Dhabi, United Arab Emirates, in 2021. He is currently a Ph.D. candidate in the Department of Management Science and Engineering at Khalifa University. His research focuses on optimization and machine learning for decision-making under uncertainty, with applications in aviation operations, energy systems, sustainable fuel procurement, photovoltaic forecasting, intelligent transportation systems, and smart mobility. His methodological interests include stochastic and risk-aware optimization, mixed-integer programming, uncertainty quantification, conformal prediction, probabilistic forecasting, graph neural networks, trustworthy machine learning, and decision-focused machine learning.