Graph-Structured Operator Learning with Conformal Calibration for Airport Network Forecasting

Soufiane Momtaz[1,*]

[1]ENSET Mohammedia, Hassan II University of Casablanca, Morocco

Abstract

Graphs provide a natural representation for air traffic management (ATM) and airport systems, where traffic evolves over a network of coupled nodes and disruptions propagate structurally rather than independently. This paper casts daily airport traffic forecasting as a *graph-structured operator learning* problem: past traffic fields on an airport network are mapped to future fields across multiple horizons. This paper instantiates this perspective with Calibrated Probabilistic Graph Neural Operator (CP-GNO), a count-aware probabilistic graph neural operator that learns on a data-driven airport graph and outputs Negative Binomial predictive distributions for departures and arrivals. Total movements are derived by construction, ensuring exact accounting consistency. To make the forecasts decision-ready, the study applies split conformal calibration to obtain empirically reliable multi-horizon prediction intervals under both nominal and shifted regimes. Experiments on a publicly documented multi-year dataset of daily IFR movements show that graph-structured operator learning improves probabilistic accuracy relative to strong seasonal baselines and delivers its clearest robustness gains over a temporal-only baseline under disruption. The results of this paper position airport traffic forecasting as a concrete instance of trustworthy graph-based structural learning on dynamic infrastructure networks.

Keywords

graph-structured learning, neural operators, dynamic airport networks, air traffic management, probabilistic forecasting, conformal calibration

1. Introduction

Graphs can be considered as a native language for airport systems: airports are nodes, traffic interactions induce structured dependencies, and disruptions propagate across the network rather than remaining local. For forecasting reasons, this is very important because airport traffic is not only a temporal extrapolation problem but also a *structural* one. Ignoring the logic of the graph in this context can erase hub effects, regional co-movements, and non-local perturbations that dominate the operational behavior of the system; related evidence from spatio-temporal graph forecasting further shows that explicit network structure can improve prediction on coupled transportation systems.

Airport traffic forecasting supports decisions ranging from resource planning to demand–capacity balancing. Yet point accuracy alone is insufficient for air traffic management (ATM) facing decisions: planners also need uncertainty estimates that remain trustworthy when traffic regimes change. This reliability requirement connects probabilistic forecasting, proper scoring rules, and distribution-free calibration: overconfident predictions can produce brittle staffing plans, unstable slot-allocation policies, and misestimated demand–capacity buffers precisely when the network is under stress.

This paper frames daily airport traffic forecasting as *graph-structured operator learning with reliability constraints*. Our perspective is that the airport system can be treated as a traffic field evolving over a data-driven graph, and that the forecasting model should learn a field-to-field operator rather than a collection of disconnected local predictors. This is consistent with neural-operator work on learning mappings between function spaces, adapted here from regular scientific-computing domains to a discrete airport graph. Within this view, four design requirements become central: (i) encode airport interactions through an explicit graph structure; (ii) model the count nature and overdispersion of

1st Workshop on Graphs Across AI: From Structural Reasoning to Foundation Models, held in conjunction with the IEEE World Congress on Computational Intelligence (WCCI 2026), June 21–26, 2026, Maastricht, The Netherlands

*Corresponding author.

soufiane.momtaz@gmail.com (S. Momtaz)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-08

CEUR-WS.org/Vol-4319/paper3.pdf

movements probabilistically; (iii) preserve operational accounting identities exactly; and (iv) calibrate uncertainty so that empirical coverage remains reliable across horizons and under shift.

The study instantiates this perspective with CP-GNO (Calibrated Probabilistic Graph Neural Operator), which maps past daily traffic fields on an airport graph to future fields and outputs Negative Binomial predictive distributions for departures and arrivals. Total movements are derived by construction to satisfy the exact accounting constraint. We then apply split conformal calibration to produce empirically reliable multi-horizon prediction intervals, including under a disruption regime. The resulting system combines graph-structured learning, operator-based forecasting, and distribution-free uncertainty calibration in a single workflow. The term operator learning is used deliberately: the model learns a network-wide mapping from past traffic fields to future traffic fields over a structured domain, rather than a collection of airport-wise predictors with post-hoc coupling.

Contributions

- **Graph-structured operator view of airport forecasting:** we cast multi-horizon airport traffic prediction as learning a field-to-field operator over a data-driven airport graph, making structural dependencies first-class objects of the model rather than side information.
- **Count-aware graph neural operator with exact accounting:** CP-GNO outputs Negative Binomial predictive distributions for departures and arrivals and derives totals by construction, ensuring $y^{\mathrm{tot}} = y^{\mathrm{dep}} + y^{\mathrm{arr}}$ exactly.
- **Trustworthy graph forecasting under regime shift:** we combine the graph-structured predictor with split conformal calibration, yielding multi-horizon prediction intervals with empirically controlled coverage under both nominal and shifted conditions.

2. Related Work

We position this work at the intersection of four research streams: dynamic airport networks, graph-structured forecasting, operator learning on structured domains, and calibrated probabilistic prediction.

Airport networks as dynamic graphs. Airport operations are inherently networked: congestion, disruptions, and delays propagate through coupled flight schedules and downstream capacity constraints rather than remaining local phenomena [1, 2]. This motivates representing the airport system as a dynamic graph, where airports are interacting nodes and forecasting must exploit relational dependencies instead of modeling each airport as an isolated time series.

Graph-structured learning for networked forecasting. GNNs provide the methodological foundation for learning over relational structure, from graph convolutions to attention-based message passing [3, 4]. In transportation forecasting, key models include diffusion convolutional recurrent networks [5], spatio-temporal graph convolutional networks [6], Graph WaveNet [7], MTGNN [8], and AGCRN [9]. More recent work has introduced graph ODEs, decoupled dynamic spatial–temporal modeling, strong spatial–temporal identity baselines, and propagation-delay-aware transformers [10, 11, 12, 13]. At the same time, recent empirical and survey studies stress the importance of strong baselines, reproducible benchmarks, and careful attribution of gains to graph structure [14, 15, 16]. Our setting differs from standard traffic benchmarks because the target is overdispersed airport-count forecasting with explicit departure/arrival accounting and calibrated multi-horizon uncertainty.

Operator learning on structured domains. Neural operator methods learn mappings between function spaces and are designed for field-to-field evolution problems. Representative approaches include Fourier Neural Operators [17], DeepONet [18], and general neural-operator formulations [19]. Although our domain is a discrete airport graph rather than a regular PDE mesh, the same operator view is natural: past traffic fields over the airport network are mapped to future traffic fields over the same structured system.

Probabilistic forecasting and trustworthy uncertainty. Neural forecasting has advanced through probabilistic recurrent models, interpretable deep architectures, transformer-based models, diffusion-based forecasters, and recent pretrained time-series foundation models [20, 21, 22, 23, 24, 25]. Such models are commonly evaluated with strictly proper scoring rules [26], including CRPS and its ensemble

Table 1
Dataset at a glance (daily airport IFR movements).

Item	Value
Time span	2017-01-01 to 2025-11-30
Calendar years	9 (2017–2025)
Days	3,256
Airports (ICAO)	336
States/countries	43
Records (airport-days)	993,444
IFR movements (NM)	129,302,060 (64,627,799 dep. + 64,674,261 arr.)
Avg. network volume	39,712 movements/day; 305.1 airports/day
Regime shift	2020: 20,265 movements/day (vs. 45,666/day in 2017–2019)
Accounting identity	$y_t^{\mathrm{tot}} = y_t^{\mathrm{dep}} + y_t^{\mathrm{arr}}$ (0 violations)
Second source (APT IFR)	30.14% of airport-days; 136 airports with $\geq$1 observation
Experimental subset	2017–2021, $\geq$95% coverage: 209 airports; 378,303 airport-days

decompositions [27], but predictive accuracy alone does not ensure calibrated uncertainty [28]. Conformal prediction provides finite-sample, distribution-free coverage guarantees [29, 30]; conformalized quantile regression extends this principle to regression [31], and adaptive conformal methods address distribution shift [32]. Building on this line, we combine a graph-structured, count-aware predictor with split conformal calibration [33] to produce trustworthy multi-horizon airport-traffic forecasts.

3. Graph-Structured Forecasting Problem and Dataset

3.1. Problem Setup

Let $\mathcal{A} = \{1, \ldots, N\}$ be the set of airports (graph nodes) and t index days. For each airport $a \in \mathcal{A}$ we observe daily counts $x_t(a) = [y_t^{\mathrm{dep}}(a), y_t^{\mathrm{arr}}(a)] \in \mathbb{R}_+^2$. Total movements are defined as $y_t^{\mathrm{tot}}(a) = y_t^{\mathrm{dep}}(a) + y_t^{\mathrm{arr}}(a)$. Given a lookback window of length L and forecast horizon H, the goal is to learn

$$p_\theta(X_{t+1:t+H} \mid X_{t-L+1:t}), \tag{1}$$

where $X_t \in \mathbb{R}_+^{N\times 2}$ stacks $x_t(a)$ across airports.

3.2. Dataset

We use a publicly documented dataset of daily airport IFR traffic counts indexed by day and airport [34]. Table 1 summarizes its scale, diversity, and key statistics. The dataset spans **2017-01-01 to 2025-11-30** (**3,256** days; **9** calendar years) and contains **993,444** airport-day records covering **336** ICAO airports across **43** States/countries.

In this paper, we focus on the Network Manager (NM) series, which provides daily **departures** and **arrivals** and satisfies an exact accounting identity with **0 violations** over the full dataset: $y_t^{\mathrm{tot}}(a) = y_t^{\mathrm{dep}}(a) + y_t^{\mathrm{arr}}(a)$. Across 2017–2025, the NM series totals **129,302,060** IFR movements (**64,627,799** departures and **64,674,261** arrivals), with an average network volume of ≈**39,712 movements/day** and ≈**305.1 airports/day**.

The dataset exhibits strong regime diversity and non-stationarity: average daily network traffic drops from ≈**45,666 movements/day** in 2017–2019 to ≈**20,265/day** in 2020, and recovers to ≈**42,625/day** in 2022–2024. It also presents heavy-tailed spatial heterogeneity: the median airport averages **34.49** movements/day, while the top-10 airports account for **23.39%** of all movements (top-50: **64.04%**).

In addition to NM, the dataset includes an Airport Operator (APT) IFR series available for **30.14%** of airport-days across **136** airports; we use it only for descriptive statistics in this version and reserve multi-source learning for future work. For controlled network-scale evaluation, we restrict experiments

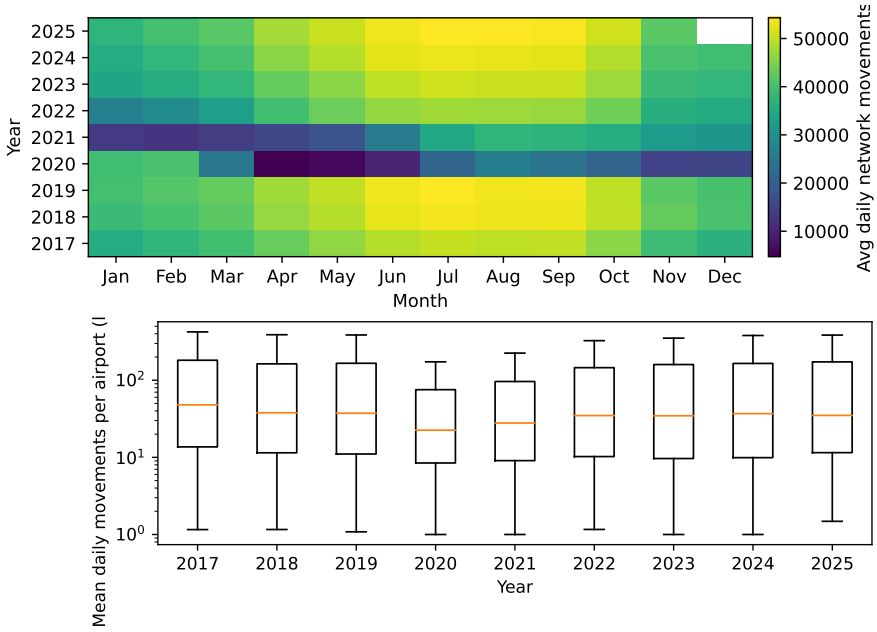


Figure 1: Dataset overview (Network Manager totals). The heatmap highlights seasonality and the 2020 disruption, while the boxplots show heavy-tailed airport heterogeneity and across-year regime changes.

to 2017–2021 and retain airports with at least **95%** daily coverage, resulting in $N{=}209$ airports (**378,303** airport-days).

4. Method: Graph-Structured Operator Learning with Conformal Calibration

4.1. Graph Construction (Structure Induction from Data)

To remain dataset-only, we build a weighted airport graph $G = (V, E, W)$ from the training split only by computing pairwise similarity between airports from co-movements in de-seasonalized total movements and retaining the top-K neighbors for each node. The resulting graph acts as a structural scaffold for the forecasting model: it encodes which airports should exchange information even when explicit route or rotation data are unavailable. Retained edge weights are normalized with a temperature-scaled softmax, and this fixed graph is kept unchanged throughout training and evaluation so that the reported gains do not rely on learned topology updates.

4.2. Graph Neural Operator Backbone

CP-GNO learns an operator that maps past traffic fields on the graph to future fields. At each day t, a graph operator mixes spatial information via weighted message passing, producing latent embeddings $z_t(a)$ that already encode relational context. A lightweight temporal convolutional network (TCN) then processes the lookback sequence $z_{t-L+1:t}(a)$ to produce a representation $h_t(a)$, from which horizon-specific heads output distribution parameters for departures and arrivals. This factorization lets the model treat graph structure as part of the computation itself, rather than as a post-hoc feature engineering step. The Local baseline uses the same temporal backbone and probabilistic heads, with the graph mixing removed.

4.3. Count-Aware Probabilistic Output and Exact Accounting

Daily movements are non-negative counts and typically overdispersed. We model departures and arrivals with a Negative Binomial (NB) likelihood. For each horizon h and airport a, CP-GNO outputs $(\mu^{(\cdot)}_{t+h}(a), r^{(\cdot)}_{t+h}(a))$ with $\mu, r > 0$, and $y^{(\cdot)}_{t+h}(a) \sim \mathrm{NB}(\mu^{(\cdot)}_{t+h}(a), r^{(\cdot)}_{t+h}(a))$ for $(\cdot) \in \{\mathrm{dep}, \mathrm{arr}\}$. To enforce operational consistency, totals are derived by construction:

$$\hat{y}^{\mathrm{tot}}_{t+h}(a) = \hat{y}^{\mathrm{dep}}_{t+h}(a) + \hat{y}^{\mathrm{arr}}_{t+h}(a). \tag{2}$$

4.4. Training Objective

We train by minimizing the NB negative log-likelihood over airports and horizons. Let $\ell_{\mathrm{NB}}(y; \mu, r) = -\log p_{\mathrm{NB}}(y \mid \mu, r)$. Then

$$\begin{aligned}\mathcal{L}(\theta) = \sum_{h=1}^{H} \sum_{a \in \mathcal{A}} \Big(& \ell_{\mathrm{NB}}\Big(y^{\mathrm{dep}}_{t+h}(a); \mu^{\mathrm{dep}}_{t+h}(a), r^{\mathrm{dep}}_{t+h}(a)\Big) \\ & + \ell_{\mathrm{NB}}\big(y^{\mathrm{arr}}_{t+h}(a); \mu^{\mathrm{arr}}_{t+h}(a), r^{\mathrm{arr}}_{t+h}(a)\big) \Big).\end{aligned} \tag{3}$$

In practice, this loss is averaged over all sliding training windows in the training split.

4.5. Split Conformal Calibration (Multi-Horizon)

Probabilistic models can be miscalibrated. We calibrate prediction intervals on a held-out calibration split (2020H1) and evaluate empirical coverage on Test-ID and Test-OOD. For miscoverage level α, we form central prediction intervals from NB quantiles and apply split conformal adjustment using nonconformity scores $s_i = \max\{L_i - y_i,\ y_i - U_i,\ 0\}$. The calibrated interval is $[\max(0, L - q_\alpha),\ U + q_\alpha]$, where q_α is the $(1-\alpha)$ empirical quantile of $\{s_i\}$. In the reported experiments, conformal quantiles are estimated separately for departures and arrivals, for each forecast horizon and nominal coverage level, pooling calibration residuals across airports on the 2020H1 split.

Table 2
Experimental protocol and temporal splits (filtered airport subset, $N{=}209$). "Airport-days" are the nominal maximum $N \times$ days; the actual total is slightly smaller due to missing days (overall 378,303 airport-days) while maintaining $\geq 95\%$ coverage per airport.

Split	Period	Days	Nominal airport-days
Train	2017–2019	1,095	228,855
Cal (for conformal)	2020H1	182	38,038
Test-OOD	2020H2	184	38,456
Test-ID	2021	365	76,285

5. Experiments

5.1. Protocol and Splits

We evaluate both nominal conditions (Test-ID, 2021) and a disruption regime (Test-OOD, 2020H2), using a recent calibration window (2020H1). Training uses pre-disruption data (2017–2019) on the filtered airport subset ($N{=}209$); graph construction is performed on the training split only. The calibration split deserves emphasis: 2020H1 already contains the COVID-19 shock. We choose it because it is temporally prior to Test-OOD and Test-ID while reflecting a recent disrupted regime; the experiment therefore tests whether early-disruption residuals calibrate later disruption and recovery periods, not universal robustness to arbitrary shifts. A pre-COVID calibration window would assess a stricter transport setting and is left as an explicit sensitivity experiment.

5.2. Baselines

We compare CP-GNO against (i) a seasonal naive baseline (SN) using same-week lags (7 days; 14 days for $h = 14$), and (ii) a temporal-only operator (Local) that removes graph coupling but keeps the same temporal features and probabilistic head. The external STGNN methods cited above are the relevant state-of-the-art comparison set; we do not report newly adapted numbers in this camera-ready version without the full retraining protocol, because fair adaptation from road-sensor point forecasting to daily airport-count NB likelihoods and conformal intervals is non-trivial. The Local model is therefore a controlled graph-ablation baseline rather than an exhaustive SOTA benchmark. For the same-week seasonal baseline, horizons 1, 3, and 7 all reduce to a 7-day lag predictor relative to the target day; as a result, aggregate scores can coincide across these horizons.

5.3. Metrics

We report mean absolute error (MAE) and negative log-likelihood (NLL). For uncertainty, we measure empirical coverage of central 90% and 95% prediction intervals. All scores are averaged over airports and dates, then averaged across departures/arrivals. Because total movements are derived exactly from departures and arrivals, we do not report a separate total-series benchmark; coherence is guaranteed algebraically.

6. Results

6.1. Main Results

We evaluate daily forecasts for departures and arrivals at horizons $h \in \{1, 3, 7, 14\}$ days.

Table 3 reports point and probabilistic accuracy on the disruption regime (Test-OOD, 2020H2). CP-GNO achieves the lowest NLL across all horizons under regime shift, indicating better probabilistic forecasting quality when the system is non-stationary.

Table 3
Point and probabilistic accuracy on Test-OOD (2020H2). Each cell reports MAE↓ / NLL↓, averaged over departures and arrivals. Boldface marks the best value for each metric within each horizon.

Method	h=1	h=3	h=7	h=14
SN	5.91 / 4.18	5.91 / 4.18	**5.91** / 4.18	**8.19** / 5.04
Local	5.32 / 3.29	6.12 / 3.49	6.33 / 3.66	8.82 / 4.36
CP-GNO	**4.93** / **3.23**	**5.82** / **3.48**	6.11 / **3.62**	8.56 / **4.33**

Table 4
Point and probabilistic accuracy on Test-ID (2021). Each cell reports MAE↓ / NLL↓, averaged over departures and arrivals. Boldface marks the best value for each metric within each horizon.

Method	h=1	h=3	h=7	h=14
SN	5.48 / 3.97	**5.48** / 3.97	**5.48** / 3.97	**7.14** / 4.47
Local	5.42 / 3.23	6.04 / **3.34**	5.77 / **3.36**	7.54 / **3.66**
CP-GNO	**5.20** / **3.22**	5.82 / **3.34**	5.86 / 3.38	7.71 / 3.69

Table 5
Regime-shift robustness measured by the OOD–ID score gap. Each cell reports ΔMAE / ΔNLL where Δ = Test-OOD − Test-ID (positive indicates degradation under disruption). Boldface marks the smallest degradation for each metric within each horizon.

Method	h=1	h=3	h=7	h=14
SN	+0.43 / +0.21	+0.43 / +0.21	+0.43 / +**0.21**	+1.05 / +**0.57**
Local	-0.10 / +0.06	+0.08 / +0.15	+0.56 / +0.30	+1.28 / +0.70
CP-GNO	**-0.27** / +**0.01**	+**0.00** / +**0.14**	+**0.25** / +0.24	+**0.85** / +0.64

A key observation is that the largest gains occur at short horizons under disruption: for h=1, CP-GNO reduces MAE from 5.91 (seasonal naive) to 4.93 while also reducing NLL from 4.18 to 3.23. At longer horizons, all methods degrade as expected, yet CP-GNO consistently improves NLL over both SN and the temporal-only operator (Local).

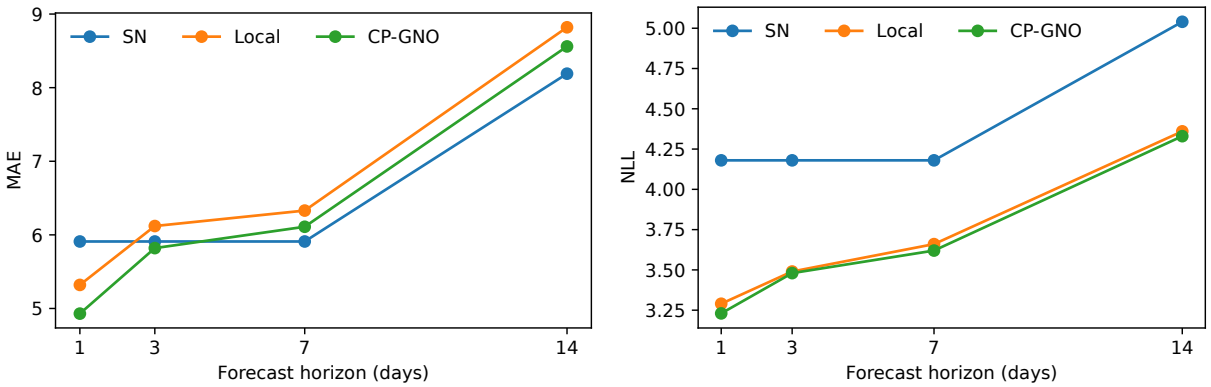
Table 4 reports the same metrics under nominal conditions (Test-ID, 2021). While the margin in MAE is smaller in the nominal year, CP-GNO remains competitive and substantially improves NLL compared to the seasonal baseline.

Robustness to regime shift. Absolute scores on Test-OOD and Test-ID are informative, but operational robustness is clarified by directly comparing how each method degrades under disruption. Table 5 reports the OOD–ID gap (Test-OOD minus Test-ID) for both MAE and NLL. Across horizons, CP-GNO exhibits a smaller NLL degradation than the temporal-only operator (Local), consistent with the hypothesis that spatial coupling becomes more informative when the network is non-stationary.

To visualize horizon effects on the disruption regime, Fig. 2 plots MAE and NLL versus horizon on Test-OOD. Both error and NLL increase with the horizon, but CP-GNO shows consistently better likelihood-based behavior across horizons.

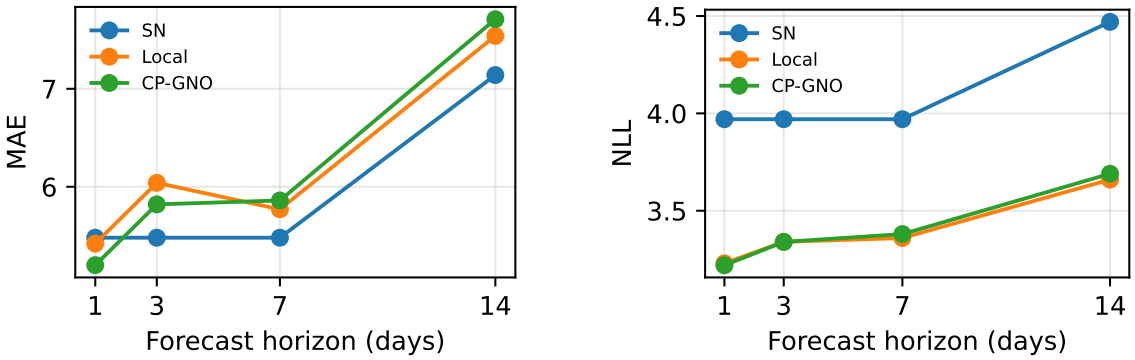
We also visualize the nominal regime in Fig. 3. While point-error differences are small on Test-ID, CP-GNO maintains strong likelihood-based performance, aligning with the observation that the main gains in nominal years are often probabilistic rather than purely point-accurate.

Relative gains across regimes. Tables 3–4 report absolute metrics, but relative gains clarify when improvements are operationally meaningful. Table 6 reports percent improvements of CP-GNO over baselines. On Test-OOD, CP-GNO improves NLL by 13–23% over seasonal naive across horizons and also improves MAE by 3–7% over the temporal-only operator, highlighting that the graph structure becomes most valuable when the network is perturbed. On Test-ID, the main benefit remains probabilistic quality (NLL) versus seasonal naive, while gains versus the strong temporal-only baseline are smaller and can vanish at longer horizons.



(a) MAE vs. horizon (Test-OOD 2020H2). (b) NLL vs. horizon (Test-OOD 2020H2).

Figure 2: Accuracy and probabilistic quality across horizons on the disruption regime (Test-OOD).



(a) MAE vs. horizon (Test-ID 2021). (b) NLL vs. horizon (Test-ID 2021).

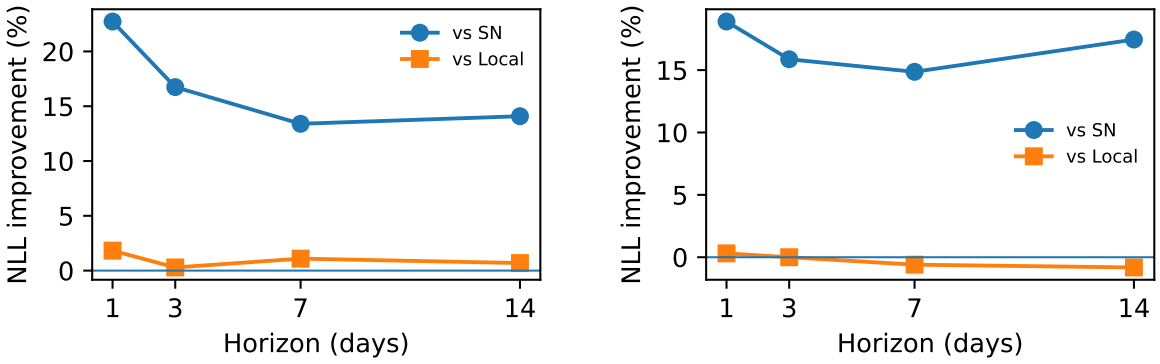
Figure 3: Accuracy and probabilistic quality across horizons on the nominal regime (Test-ID).

Table 6
Relative improvement (%) of CP-GNO over baselines across horizons (positive is better; negative indicates worse). Improvements are computed as (baseline − CP-GNO)/baseline × 100 for MAE and NLL.

(a) Test-OOD (2020H2)				
Metric	h=1	h=3	h=7	h=14
ΔMAE vs SN	+16.6	+1.5	-3.4	-4.5
ΔNLL vs SN	+22.7	+16.7	+13.4	+14.1
ΔMAE vs Local	+7.3	+4.9	+3.5	+2.9
ΔNLL vs Local	+1.8	+0.3	+1.1	+0.7
(b) Test-ID (2021)				
Metric	h=1	h=3	h=7	h=14
ΔMAE vs SN	+5.1	-6.2	-6.9	-8.0
ΔNLL vs SN	+18.9	+15.9	+14.9	+17.4
ΔMAE vs Local	+4.1	+3.6	-1.6	-2.3
ΔNLL vs Local	+0.3	+0.0	-0.6	-0.8

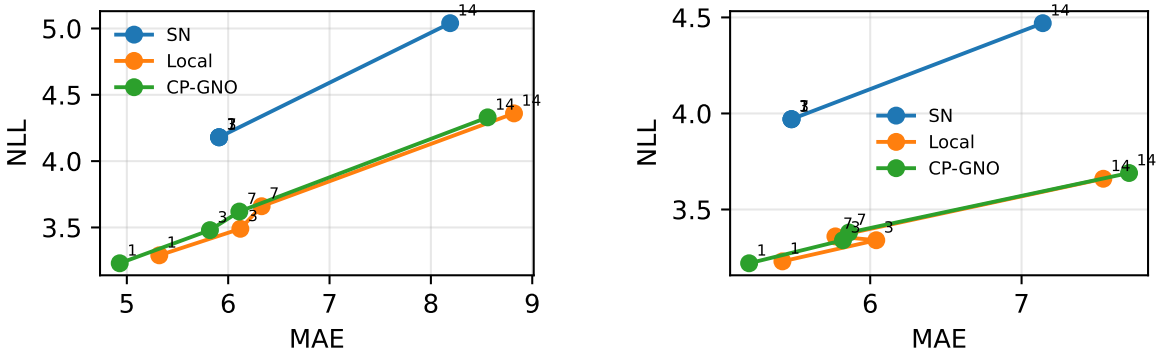
Fig. 4 visualizes the NLL improvements in both regimes.

MAE–NLL trade-offs. Because operational decisions may trade point error against uncertainty quality, Fig. 5 plots MAE versus NLL across horizons for each method. On Test-OOD, CP-GNO shifts the Pareto curve toward lower NLL for comparable MAE, while on Test-ID the curves cluster more tightly, indicating a smaller headroom for gains under nominal conditions.



(a) Test-OOD (2020H2). (b) Test-ID (2021).

Figure 4: Relative NLL improvement of CP-GNO over baselines across horizons. Positive values indicate lower (better) NLL for CP-GNO.



(a) Test-OOD (2020H2). (b) Test-ID (2021).

Figure 5: MAE–NLL trade-offs across horizons (labels indicate the horizon in days). Lower is better on both axes.

6.2. Ablations and Robustness

Graph operator ablation. Comparing CP-GNO to the temporal-only operator (Local) isolates the value of graph coupling while controlling for the probabilistic head and temporal features. Under regime shift (Table 3), CP-GNO improves both MAE and NLL over Local at all horizons, suggesting that spatial coupling becomes more informative when the network experiences disruptions.

Uncertainty calibration across regimes. We calibrate CP-GNO prediction intervals using split conformal prediction on a recent calibration window (2020H1) and report empirical coverage on both Test-OOD (2020H2) and Test-ID (2021). Uncalibrated intervals increasingly under-cover as the horizon grows (on Test-OOD, Cov@90% drops to 0.787 at h=14), whereas conformal calibration restores coverage close to target levels across horizons.

Fig. 6 visualizes the same trend: uncalibrated curves drift below nominal levels as the horizon increases, while calibrated curves remain stable near the targets 0.90 and 0.95 in both regimes.

Calibration lift and conservativeness. While Fig. 6 shows absolute coverage, the *coverage lift* $\Delta\text{Cov} = \text{Cov}_{\text{cal}} - \text{Cov}_{\text{unc}}$ highlights how calibration effects grow with the horizon. Fig. 7 shows that ΔCov is small at short horizons but becomes substantial at h=14 in both regimes, consistent with increasing miscalibration as uncertainty compounds. Table 8 summarizes the reliability trade-off: conformal calibration eliminates under-coverage risk (invalid intervals) across horizons, at the cost of moderate over-coverage (conservative intervals), especially on Test-ID.

Table 7
Empirical coverage before and after split conformal calibration. Each cell reports Cov@90% / Cov@95% (target is 0.90/0.95).

Method	h=1	h=3	h=7	h=14
(a) Test-OOD (2020H2)				
CP-GNO	0.900/0.936	0.875/0.913	0.858/0.897	0.787/0.834
CP-GNO+Cal	0.927/0.966	0.922/0.967	0.927/0.968	0.921/0.972
(b) Test-ID (2021)				
CP-GNO	0.909/0.943	0.896/0.932	0.891/0.926	0.844/0.888
CP-GNO+Cal	0.934/0.970	0.938/0.978	0.949/0.982	0.954/0.988

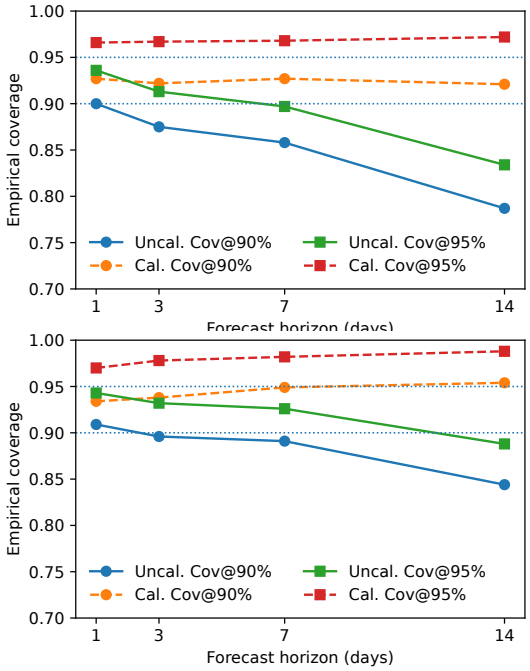


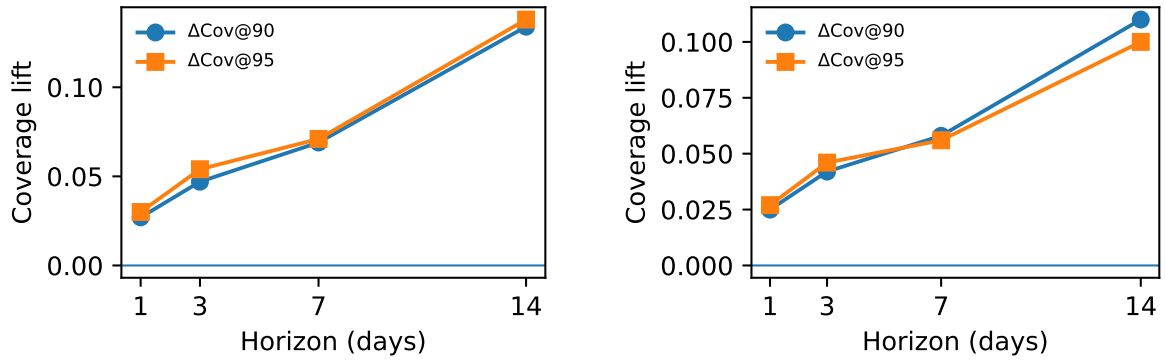
Figure 6: Empirical coverage vs. horizon before and after split conformal calibration (calibration split: 2020H1). Top: Test-OOD (2020H2). Bottom: Test-ID (2021). Dotted lines denote target coverages 0.90 and 0.95.

Table 8
Calibration trade-off across horizons. Under= $\max(0, \tau - \text{Cov})$ (risk) and Over= $\max(0, \text{Cov} - \tau)$ (conservativeness). Values are mean/max across horizons.

		Uncal. Under		Cal. Over	
Regime	τ	mean	max	mean	max
Test-OOD	0.90	0.045	0.113	0.024	0.027
Test-OOD	0.95	0.055	0.116	0.018	0.022
Test-ID	0.90	0.017	0.056	0.044	0.054
Test-ID	0.95	0.028	0.062	0.030	0.038

7. Discussion

The empirical picture is consistent with the intended repositioning of the paper: the main value of CP-GNO is not merely that it is another forecaster, but that it uses graph structure as an actionable inductive bias for a dynamic airport network. CP-GNO is most beneficial under disruption, where network coupling becomes more informative. On Test-OOD (2020H2), CP-GNO improves both point accuracy and probabilistic accuracy relative to the seasonal baseline, and consistently lowers NLL versus



(a) Test-OOD (2020H2).

(b) Test-ID (2021).

Figure 7: Coverage lift induced by split conformal calibration, $\Delta\text{Cov} = \text{Cov}_{\text{cal}} - \text{Cov}_{\text{unc}}$, for 90% and 95% target intervals across horizons.

the temporal-only operator (Table 3), indicating better distributional forecasts under non-stationarity. Under nominal conditions (Test-ID, 2021), CP-GNO remains competitive (Table 4), suggesting that graph coupling provides incremental gains beyond strong temporal modeling.

A central outcome is reliability. Without calibration, prediction intervals increasingly under-cover at longer horizons (Table 7), which would underestimate risk in operational planning. Split conformal calibration restores near-nominal coverage in both regimes (Fig. 6), with mild conservativeness at short horizons. This makes the method a form of *trustworthy graph forecasting*: the graph supplies structural coupling, the probabilistic head captures dispersion, and conformal calibration repairs the operational reliability of intervals.

Our reliability analysis focuses on empirical coverage rather than explicit interval-width metrics, so future work should complement coverage with sharpness diagnostics.

Limitations and scope of guarantees.

The dataset-only design is intentional but limits causal interpretation: graph edges should be understood as predictive co-movement dependencies rather than physical routes, rotations, or delay-causation links. Since meteorology, ATFM regulations, runway capacity, airline schedules, and event indicators are not included, abrupt capacity or policy shocks can only be captured through historical traffic patterns. The comparison scope is also limited, as the evidence demonstrates gains over a seasonal operational baseline and a controlled graph ablation, but not dominance over the full spatio-temporal forecasting SOTA without reproduced external STGNN baselines. Likewise, split conformal prediction provides marginal finite-sample coverage under exchangeability, not conditional guarantees for every airport, season, airport class, or disruption type. The use of COVID-affected 2020H1 for calibration may partly preserve residual-regime continuity with 2020H2, so stricter robustness tests should include pre-COVID calibration, rolling recalibration, and adaptive or group-conditional conformal variants. Future work should therefore incorporate exogenous operational drivers, report interval sharpness, reproduce external STGNN baselines, and assess heterogeneous reliability across airport groups.

Operationally, CP-GNO supports risk-aware ATM decisions by combining count-aware probabilistic forecasts with calibrated intervals, enabling staffing, gate/stand allocation, and demand–capacity mitigation based on upper-quantile demand. Its coherent construction of departures, arrivals, and totals allows consistent aggregation from airport-level to regional and network-wide forecasts without post-hoc reconciliation. Two modeling choices are central to reliability: the Negative Binomial head captures overdispersed airport counts and evaluates uncertainty beyond MAE, while the data-driven airport graph provides a structural prior based on observed co-movements. The robustness gaps suggest that graph coupling helps preserve uncertainty quality under disruptions, making the graph a core computational mechanism rather than a mere representation. In deployment, calibration should remain adaptive, with empirical coverage monitored over recent windows and conformal scores refreshed

when seasonal, policy, or ATFM-related drift is detected.

8. Conclusion

This paper presented airport traffic forecasting as a graph-structured operator learning problem and instantiated this view with CP-GNO, a calibrated probabilistic graph neural operator for multi-horizon prediction on airport networks. CP-GNO combines (i) operator learning on a data-driven airport graph, (ii) count-aware Negative Binomial predictive distributions, (iii) exact accounting constraints by construction, and (iv) split conformal calibration for empirically reliable uncertainty. Experiments show improved probabilistic accuracy and substantially better calibrated prediction intervals under regime shift, while remaining competitive under nominal conditions. More broadly, the results show that airport forecasting is a meaningful application domain for trustworthy graph-based AI on dynamic infrastructure networks.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT 5.4 for language editing then reviewed and edited the paper as needed and takes full responsibility for the publication's content.

References

[1] N. Pyrgiotis, K. M. Malone, and A. Odoni, "Modelling delay propagation within an airport network," *Transportation Research Part C: Emerging Technologies*, vol. 27, pp. 60–75, 2013, doi: 10.1016/j.trc.2011.05.017.

[2] P. Fleurquin, J. J. Ramasco, and V. M. Eguíluz, "Systemic delay propagation in the US airport network," *Scientific Reports*, vol. 3, Art. no. 1159, 2013, doi: 10.1038/srep01159.

[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[4] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[5] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[6] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 3634–3640, doi: 10.24963/ijcai.2018/505.

[7] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph WaveNet for deep spatial-temporal graph modeling," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, pp. 1907–1913, doi: 10.24963/ijcai.2019/264.

[8] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2020, pp. 753–763, doi: 10.1145/3394486.3403118.

[9] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17804–17815.

[10] Z. Fang, Q. Long, G. Song, and K. Xie, "Spatial-temporal graph ODE networks for traffic flow forecasting," in *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2021, pp. 364–373, doi: 10.1145/3447548.3467430.

[11] Z. Shao, Z. Zhang, W. Wei, F. Wang, Y. Xu, X. Cao, and C. S. Jensen, "Decoupled dynamic spatial-temporal graph neural network for traffic forecasting," *Proceedings of the VLDB Endowment*, vol. 15, no. 11, pp. 2733–2746, 2022, doi: 10.14778/3551793.3551827.

[12] Z. Shao, Z. Zhang, F. Wang, W. Wei, and Y. Xu, "Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting," in *Proc. ACM International Conference on Information and Knowledge Management (CIKM)*, 2022, pp. 4454–4458, doi: 10.1145/3511808.3557702.

[13] J. Jiang, C. Han, W. X. Zhao, and J. Wang, "PDFormer: Propagation delay-aware dynamic long-range Transformer for traffic flow prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, pp. 4365–4373, 2023, doi: 10.1609/aaai.v37i4.25556.

[14] X. Liu, Y. Liang, C. Huang, H. Hu, Y. Cao, B. Hooi, and R. Zimmermann, "Do we really need graph neural networks for traffic forecasting?" arXiv:2301.12603, 2023.

[15] Z. Al Sahili and M. Awad, "Spatio-temporal graph neural networks: A survey," arXiv:2301.10569, 2023.

[16] F. Corradini, F. Gerosa, M. Gori, C. Lucheroni, M. Piangerelli, and M. Zannotti, "A systematic literature review of spatio-temporal graph neural network models for time series forecasting and classification," *Neural Networks*, vol. 195, Art. no. 108269, 2026, doi: 10.1016/j.neunet.2025.108269.

[17] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, "Fourier Neural Operator for parametric partial differential equations," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
[18] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis, "Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 218–229, 2021, doi: 10.1038/s42256-021-00302-5.
[19] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar, "Neural Operator: Learning maps between function spaces with applications to PDEs," *Journal of Machine Learning Research*, vol. 24, no. 89, pp. 1–97, 2023.
[20] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020, doi: 10.1016/j.ijforecast.2019.07.001.
[21] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.
[22] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.
[23] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting," in *Proc. International Conference on Machine Learning (ICML)*, PMLR, vol. 139, 2021, pp. 8857–8868.
[24] A. F. Ansari *et al.*, "Chronos: Learning the language of time series," *Transactions on Machine Learning Research*, 2024.
[25] K. Rasul *et al.*, "Lag-Llama: Towards foundation models for probabilistic time series forecasting," arXiv:2310.08278, 2023.
[26] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007, doi: 10.1198/016214506000001437.
[27] H. Hersbach, "Decomposition of the continuous ranked probability score for ensemble prediction systems," *Weather and Forecasting*, vol. 15, no. 5, pp. 559–570, 2000.
[28] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. International Conference on Machine Learning (ICML)*, PMLR, vol. 70, 2017, pp. 1321–1330.
[29] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005.
[30] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, pp. 371–421, 2008.
[31] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[32] I. Gibbs and E. J. Candès, "Adaptive conformal inference under distribution shift," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 1660–1672.
[33] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," arXiv:2107.07511, 2021.
[34] EUROCONTROL Aviation Intelligence Unit, "Airport Traffic Dataset," *Aviation Intelligence Portal*, 2026. [Online]. Available: https://ansperformance.eu/reference/dataset/airport-traffic/. Accessed: Apr. 30, 2026.