

A Lightweight Transformer Framework for Multi-Scale, Spatiotemporal Air Traffic Density Forecasting in Regional Airport Safety Research

Enya Zheng¹, Eva Zheng¹, Barnas Monteith²

¹Campolindo High School, ²Center for Science Engagement

Correspondence: enyalzh@gmail.com; barnas@engagescience.org

Abstract

Large-scale global air traffic models impose computational demands that can limit their practical deployment in resource-constrained, region-specific aviation safety research. This paper presents a lightweight, reproducible Transformer pipeline for localized, multi-scale air traffic density forecasting. Using five years of publicly available weekly (Monday) ADS-B snapshot data for the San Francisco Bay Area (254 observed Mondays, 2017-2022; 3,051,163 position records), we use K-Means to group hourly air traffic conditions into eight states. We then train three compact encoder-only Transformers (0.8M parameters each), one per operational scale: same-day tactical (12-hour), weekly (7-day), and strategic (two-month) forecasting. The learned advantage is concentrated at the tactical scale, where same-day context is available; at the two-month horizon, the Transformer does not outperform hourly climatology, the strongest non-learned baseline. The compact model achieves 60.4% tactical accuracy, compared with 57.8% for climatology. This advantage is centered on atypical hours, when the observed traffic state differs from the state normally seen at that hour. Such deviations occur in 42% of test hours. The model correctly predicts the observed state in 23% of the atypical hours, while climatology, which only predicts the typical state, cannot predict any of them. The entire pipeline runs in about an hour on a 2019 laptop CPU, and all code and derived data are released to support reproduction. We argue that such compact, transparent, locality-first models offer a practical and underexplored approach to applied aviation safety research, particularly in regions where large infrastructure investments are infeasible.

Keywords: aviation safety, air traffic forecasting, Transformer architecture, spatiotemporal modeling, edge deployment, small model, regional airports, density prediction, K-Means clustering, multi-scale prediction

1. Introduction

Air traffic management has long relied on real-time surveillance systems that alert controllers to hazards as they emerge in the operational environment. The operational limitations of this reactive approach are illustrated by the 2017 Air Canada Flight 759 near-miss at San Francisco International Airport (SFO). During a busy nighttime period, an Airbus A320 on approach misidentified a taxiway as a runway and descended to about 60 feet above the first of four aircraft awaiting departure on the taxiway (National Transportation Safety Board, 2018). This incident also underscores the importance of identifying potentially challenging traffic conditions before they develop into immediate operational hazards. The past decade of ongoing growth in commercial and general aviation demand has further highlighted the need for predictive models for traffic management (International Air Transport Association [IATA], 2025a; IATA, 2025b; Belz et al., 2025). Such models could provide earlier indications of elevated traffic conditions and support proactive safety analysis.

A substantial literature on air traffic forecasting has emerged, ranging from traditional time-series models (Xu et al., 2019) through Long Short-Term Memory (LSTM) recurrent networks (Kim et al., 2016; Gui et al., 2020) to Graph Neural Networks (GNNs) that capture relationships within airport taxiway and approach corridor networks (Wang et al., 2026; Xu et al., 2021; Yin et al., 2025). More recently, Transformer architectures (Vaswani et al., 2017) have been adapted for trajectory-level forecasting (Luo et al., 2024; Yoon & Lee, 2025; Guo et al., 2024) and passenger flow prediction at major hub airports (Du et al., 2026).

As deep learning has grown more powerful, researchers have increasingly favored training models on large, aggregated multi-airport datasets to achieve generalizable performance across diverse environments. At the same time, emerging AI applications in air traffic management illustrate the value of designing computational tools around specific operational contexts and human decision making. For example, CHATATC uses historical traffic-management data to support, rather than replace, human traffic managers in strategic decision making (Abdulkhak et al., 2024). However, there is a research gap at the intersection of lightweight modeling and facility-specific analysis; few studies explore small, computationally efficient models designed specifically for a single regional facility or airspace. This represents an important research opportunity because facility-specific aviation safety applications may benefit from models tailored to the operational and spatial patterns of local air traffic. A localized approach can be designed to capture these facility-specific patterns while maintaining low computational requirements. This paper introduces a framework that deliberately embraces locality and compactness as design principles. Our contributions are as follows:

- **A Reproducible Data Preprocessing Workflow:** We describe and validate an end-to-end pipeline that condenses multi-gigabyte ADS-B state-vector archives into a compact table of 5,973 hourly observations. Five-minute per-aircraft downsampling and a weekly-pattern imputation scheme (missing hours filled from the same hour on preceding observed Mondays) applied to model input only - never to training targets or to reported

metrics - preserve sequence continuity while keeping every scored value a real observation, and the complete workflow executes on a commodity laptop CPU.

- **A Cluster-Sequence Formulation of Density Forecasting:** Building upon this data foundation, we present a methodology that transforms high-variance, continuous flight trajectory data into an eight-state discrete traffic state representation via K-Means clustering. This discretization allows a compact encoder-only Transformer to perform classification rather than regression and to emit every forecast horizon in a single non-autoregressive forward pass.
- **Multi-Scale Empirical Validation:** We provide validation results across tactical (12-hour), weekly (7-day), and strategic (two-month) horizons, evaluated against a baseline hierarchy that culminates in an hourly-climatology reference far stronger than the persistence baselines common in this literature. Accompanying ablation studies isolate the architectural components that carry the learned signal and delineate the horizons at which learning improves on climatology, qualifying rather than simply confirming the proposed Transformer's advantages.
- **A Framework for Localized Edge Deployment:** We articulate a practical use-case framework for applying these compact, localized models across diverse regional aviation safety contexts, including mountainous terrain corridors, individual airport facilities, and wildlife-sensitive approach paths.

2. Related Work

2.1 Statistical and Early Sequential Baselines in Air Traffic Forecasting

Traditional time-series approaches for air traffic demand forecasting have historically relied on Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) models. These models assume linear relationships and require data stationarity (Skowronek, 2025). To address this constraint, researchers employ differencing techniques to resolve non-stationary trends before model fitting (Xu et al., 2019). While ARIMA captures non-seasonal trends and SARIMA accounts for predictable seasonal cycles, both models struggle with the complex, non-linear dynamics of daily airport operations. As a result, their accuracy drops during sudden external disruptions or irregular events, such as extreme weather delays, network problems, or unique public events.

The adoption of LSTM networks represented a significant advance for modelling sequential aviation data (Hochreiter & Schmidhuber, 1997). Exploration by Kim et al. (2016) applied LSTM based recurrent neural networks to air traffic delay prediction by modeling temporal sequences of airport departure and arrival delays, demonstrating that deeper sequential architectures achieved higher accuracy than previous prediction methods. Gui et al. (2020) developed a generalized flight delay prediction task utilizing automatic dependent surveillance-broadcast messages integrated with weather and schedule data. While they found LSTMs capable of handling the sequential data, the architecture suffered from overfitting on

limited datasets, leading them to propose a Random Forest-based model that overcame this issue to achieve a 90.2% binary classification accuracy.

2.2 Spatiotemporal Network Architectures and Attentional Mechanisms

Despite their sequential improvements, LSTMs process sequences strictly in order, which limits their ability to capture long-range temporal dependencies (such as weekly cyclicity in traffic patterns) and constrains parallelization during training. In addition, these sequential approaches often treat airports as isolated entities, failing to capture the systemic propagation of delays across interconnected hubs. GNNs have emerged as an effective approach for modeling airport networks, where airports and their flight connections can be represented as a graph (Xu et al., 2021; Yin et al., 2025). Moving toward system-wide optimization, Wang et al. (2026) applied multi-agent GNN architectures to air traffic management decision support with promising results, using dynamic spatiotemporal graph modeling to resolve complex spatial conflicts and temporal scheduling within terminal maneuvering areas. However, GNNs require explicit definition of spatial topology as a precondition for training, limiting their applicability to well-mapped facilities and making rapid adaptation to new locations costly.

The Transformer architecture introduced by Vaswani et al. (2017) catalyzed rapid adoption across domains. In aviation, Luo et al. (2024) applied an improved Transformer to long-term 4D trajectory prediction, introducing a trajectory stabilization module and a one-step inference strategy to mitigate the autoregressive error compounding inherent to LSTM-based methods. To prevent compounding errors and overcome the limitations of binary encoding, Guo et al. (2024) developed FlightBERT++. This non-autoregressive framework utilizes a horizon-aware context generator and a differential prompted decoder to achieve efficient, multi-horizon trajectory predictions by leveraging sequence stationarity. Shifting away from standard sequence structures, Yoon and Lee (2025) introduced an inverted Transformer design that embeds each variable series into distinct variate tokens rather than traditional temporal tokens for aircraft trajectory prediction. This approach allows the attention mechanism to capture deep multivariate correlations within the trajectory sequence. In a comprehensive review of network delay prediction frameworks, Sun et al. (2025) traced the evolution of methodologies from classical statistical baselines to emerging deep learning models, while proposing a conceptual Spatial-Temporal Causal Large Language Model framework for high-precision delay propagation modeling.

2.3 Localized Airspace Forecasting and Downstream Operational Safety

Moving from macro network delays to localized hub infrastructure, Du et al. (2026) developed a Deformable Temporal–Spectral Transformer for airport passenger flow prediction at Beijing Capital International Airport, pairing a multiscale deformable partitioning module with a frequency-domain attention mechanism to capture the volatile and periodic trends across multiple prediction horizons. Addressing specific airspace domains, the ASCENT model applied lightweight Transformer architectures specifically to non-towered general aviation terminal

airspace, combining domain-aware 3D coordinate normalization with query-based decoder to generate multi-modal trajectory hypotheses with low-latency suitability for edge deployment (Prutsch et al., 2026). However, the focus of this literature has been on trajectory-level or passenger-level forecasting at large hub airports with extensive infrastructure. There remains a significant gap for aggregate density forecasting at the regional and sub-regional level using deliberately compact architectures.

This gap is particularly acute when considering downstream operational safety. Runway incursion research has a long history in aviation safety science, with the Federal Aviation Administration (FAA) tracking foundational, systemic initiatives to mitigate surface risks since 1991 (FAA, n.d.-a). More recent data-driven approaches include spatiotemporal GCN-based prediction of surface movement conflicts by modeling the dynamic spatial relationships between ground objects (Zhang et al., 2022), deep learning computer vision frameworks like the intelligent Network System for Runway Safety for real-time runway surveillance and intrusion tracking (Zhang et al., 2026), and language AI systems that utilize rule-enhanced named entity recognition to parse pilot-ATC voice transcripts for proactive ground collision risk assessment (Pang et al., 2026). While these works confirm a clear operational appetite for anticipatory, AI-driven approaches to surface safety, they remain heavily reliant on real-time, terminal-proximate data streams.

3. A Framework for Compact, Localized Models in Aviation Safety Research

The dominant paradigm in applied machine learning rewards scale: more parameters, more data, and broader generalization. While this scaling approach is effective for general-domain tasks, the heavy reliance of current state-of-the-art systems on resource-intensive, terminal-proximate data streams creates a critical barrier for localized deployment. Aviation safety research, particularly at the regional and facility-specific level, presents structural and operational characteristics that favor a different strategy. In this paper, we present a compact, geographically localized approach representing a high-value, highly accessible, and underexplored avenue for safety research. This approach yields five primary operational and methodological advantages:

3.1 Computational Portability and Edge Deployment

Many of the incidents in the FAA's safety databases occur at smaller regional airports and general aviation facilities, demonstrating that aviation safety concerns extend beyond major commercial hubs (FAA, n.d.-b, n.d.-c). However, computationally intensive deep learning systems may be less practical for localized applications with limited resources. The proposed model addresses this with a compact architecture of approximately 0.8M parameters that can be trained end-to-end on a laptop CPU without any GPU. In our implementation, training was completed in under an hour without requiring distributed computing infrastructure. These modest computational requirements lower the technical and resource barriers to predictive traffic modeling across a wider range of aviation settings.

3.2 Capture of Regional Operational Rhythms

Every airport has a distinctive spatiotemporal signature. The unique combination of runway geometry, localized prevailing wind patterns, airline scheduling agreements, noise abatement procedures, and surrounding geography can produce locally specific traffic patterns. A global model trained on aggregated multi-airport data may prioritize broader patterns at the expense of operational variations that are specific to individual locations.

By contrast, a localized approach is designed to capture these location-specific patterns and how they change over time. In our analysis, prediction error varies by time of day: it peaks at 55% during the overnight-to-morning transition window (06:00–09:00) and falls to 24% during stable mid-morning operations (see Section 4.2.3). Identifying such local variation in prediction accuracy is particularly important for aviation safety research. From this perspective, geographic specificity is a deliberate feature of the framework rather than a limitation, allowing the model to reflect local operational conditions rather than patterns averaged across multiple airports.

3.3 Interpretability via Discrete State Representations

Safety-critical aviation applications require predictive outputs that are actionable for human operators. In an active operational environment, an abstract numerical output provides minimal intuitive value to an air traffic controller. As illustrated in Section 4, our approach produces categorical outputs that are more directly interpretable for operational purposes than continuous numerical values. For example, rather than returning an abstract traffic density value such as 0.73, our model identifies the traffic condition as “Cluster 7: Peak Traffic–Terminal Area,” corresponding to roughly 91 aircraft simultaneously aloft, concentrated around terminal areas (Table 1).

The K-Means discretization at the core of our approach represents continuous traffic data as eight interpretable states, each corresponding to a distinct traffic regime. This representation produces categorical predictions that are easier for air traffic controllers, safety researchers, and airport planners to interpret and potentially use in operational decision making. The focus on interpretability is also consistent with evolving regulatory guidance, such as the European Union Aviation Safety Agency’s AI Roadmap 2.0 (2023), which emphasizes explainability and transparency in the development and deployment of AI systems in aviation.

3.4 Transferability as a Reproducible Research Methodology

While the model weights are specific to a given airspace, the underlying engineering pipeline is independent of airport-specific infrastructure. This transferability relies on standardized data formats and location-agnostic logic, which operate across three clear layers:

- The Ingestion Layer: The pipeline utilizes standard, publicly broadcast ADS-B telemetry rather than proprietary, facility-specific ground radar feeds. This allows a consistent data collection framework across locations.

- The Mathematical Processing Layer: The hourly traffic features (record counts, mean altitudes, and spatial centroids) and the K-Means clustering that groups them into traffic states use simple geometric operations. Neither step requires predefined knowledge of an airport's physical layout or runway topology.
- The Sequence Modeling Layer: The core predictive engine is a compact Transformer that learns how traffic conditions change over time and uses these patterns to make predictions. While the model needs to be retrained using data for each target location, the same model architecture can be used at different locations.

Together, these three layers provide a reproducible and standard framework for comparative aviation safety research. The same pipeline can be adapted to alternative airports, high-density transit corridors, or global Flight Information Regions with minimal computational overhead. Safety researchers can retrain the model on local historical data, enabling systematic comparisons of traffic patterns across regions.

3.5 Environmental Overlay Extensibility

A recurring challenge in aviation safety modeling is the integration of heterogeneous data streams operating at different temporal and spatial resolutions. Bird migration data, weather observations, and Notices to Air Missions (NOTAMs), for example, differ in frequency and spatial coverage from high-frequency Automatic Dependent Surveillance-Broadcast (ADS-B) position reports. Jointly incorporating these sources into a single predictive model requires additional data alignment and increases model complexity.

The proposed framework addresses this challenge through architectural separation. The core model is trained only on local traffic data, while additional environmental or operational factors can be considered separately downstream. This design allows such factors to be added, modified, or removed independently.

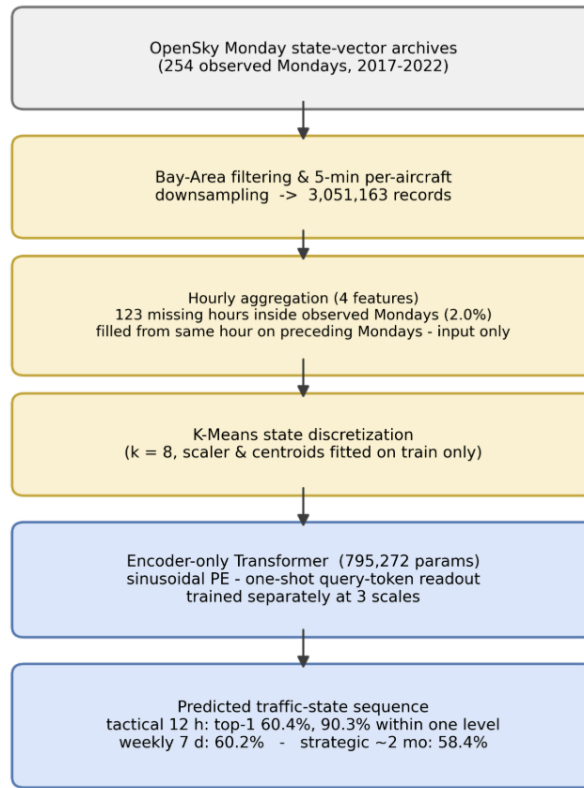
- The forecasting path illustrated in Figure 1 (Section 4) uses traffic data only. Hourly ADS-B data are first classified into eight air traffic states using K-Means ($k = 8$). A compact encoder-only Transformer (0.8M parameters) then predicts how these traffic states change over time. No stage of this forecasting path requires environmental input.
- A downstream layer could incorporate an environmental or operational factor at the spatial and temporal resolution appropriate to that data source. For example, historical bird activity could be aggregated by location and time and then considered alongside the predicted traffic state. We do not implement or evaluate such a layer in this study. Section 5.4 discusses its potential application as an extension of the proposed framework.

This separation keeps the core model lightweight and focused on traffic density while allowing the broader framework to accommodate additional safety-relevant information without retraining the forecasting model. The tradeoff is that these downstream factors do not influence

the traffic forecast itself. They provide additional context for interpreting predicted traffic density in subsequent safety analyses.

4. Applying the Framework: Traffic Density Prediction in the San Francisco Bay Area Terminal Region

We now instantiate the framework of Section 3 in a complete case study: multi-scale traffic-density forecasting for the San Francisco Bay Area terminal region (Figure 1). Section 4.1 describes the data pipeline and models; Section 4.2 reports results across three prediction horizons.



Entire pipeline reproduces in ~64 min on a 2019 laptop CPU (no GPU).

Figure 1. The air traffic density forecasting pipeline, instantiated for the San Francisco Bay Area terminal region

4.1 Methodology

4.1.1 Data Source

The dataset, *Weekly 24 Hours of State Vector Data 2017-2022*, comprises Automatic Dependent Surveillance-Broadcast (ADS-B) state vectors from the OpenSky Network's public

historical archive (Schäfer et al., 2014; The OpenSky Network, n.d.). Two structural characteristics of this archive shape the study design. First, the public release samples one day per week: hourly archives are published for Mondays only. The five-year extract spans 254 observed Mondays from June 5, 2017 to June 27, 2022; a further 11 Mondays in this interval are absent from the archive entirely. Second, even observed Mondays contain gaps: 123 hour-archives across 14 dates are absent from the source repository itself, leaving 5,973 of 6,096 possible Monday hours (98.0%) observed. Rather than synthesizing a continuous calendar timeline from this weekly sample, we treat the observed Mondays as consecutive time points in the sequence (Section 4.1.2).

Raw archives were filtered to the San Francisco Bay Area bounding box (latitude 37.0-38.5 N, longitude 123.0-121.5 W) and downsampled to one record per aircraft per five-minute window, yielding 3,051,163 position records. This region is dominated by SFO traffic but also encompasses Oakland International (OAK) and San Jose Mineta International (SJC). The forecast target is therefore regional airspace density rather than single-facility traffic. Under this sampling, one record represents one aircraft present in the airspace during a five-minute interval, with at most 12 records per aircraft per hour. The hourly record count provides a measure of traffic volume that is less affected by differences in individual aircraft reporting rates. Dividing the hourly count by 12 approximates the mean number of aircraft simultaneously aloft. Section 4.1.2 describes how the records are aggregated into hourly airspace statistics.

4.1.2 Hourly Aggregation and Missing-Hour Handling

Filtered records are aggregated into hourly statistics: record count (traffic volume), mean barometric altitude over airborne records, and the latitude and longitude of the spatial traffic centroid. These four features characterize each of the 5,973 observed hours.

The 123 missing Monday hours (2.0%) are handled with a strict separation between model input and model evaluation. Because input sequences must be temporally contiguous, each missing hour is filled with the modal traffic state observed at the same hour of day on the preceding four observed Mondays (weekly-pattern imputation). Where no preceding observation exists, the training-set modal state for that hour is used. Imputed values serve as input context only: they are masked out of the training loss and excluded from all reported metrics, so no reported number is computed on imputed data. K-Means fitting (Section 4.1.3) likewise uses observed hours only.

Critically, no continuous calendar timeline is synthesized. All sequence modeling operates on the Monday sequence directly, so the 6/7 of calendar time that the archive does not observe is never fabricated, and every training target and every evaluation point is a real observation.

4.1.3 Traffic State Discretization via K-Means Clustering

Rather than predicting continuous traffic density values through regression, we group the air traffic conditions into eight operationally meaningful states using K-Means clustering (MacQueen, 1967) over the four hourly features defined in Section 4.1.2. Each feature is standardized prior to clustering. To prevent information leakage, both the feature scaler and the K-Means model are fitted exclusively on the training partition (the first 70% of Mondays; Section 4.1.6). Hours in the validation and test periods are assigned to the nearest training centroid.

The number of clusters was selected via Silhouette analysis (Rousseeuw, 1987) and the Davies-Bouldin index (Davies & Bouldin, 1979) on the training partition, sweeping k in $\{2, 4, 8, 12\}$. Silhouette scores decline from 0.39 ($k = 2$) to 0.29, 0.23, and 0.22, respectively, while the Davies-Bouldin index is nearly flat (1.19-1.20) across the four values. The modest silhouette scores are consistent with the continuous nature of traffic density, where clear boundaries between traffic conditions may not naturally exist. We select $k = 8$ because it is the finest granularity at which each state keeps a distinct operational meaning (Table 1). With $k = 4$, the two busiest states (High Traffic and Peak Traffic–Terminal Area) merge into one, hiding exactly the distinction that matters most for congestion warning. With $k = 12$, several of the additional clusters are near-duplicates with almost identical traffic volumes, adding states without adding meaning. We return to the consequences of boundary ambiguity when analyzing error structure (Section 4.2.1). Clusters are indexed 0-7 in ascending order of mean traffic volume, giving the state space an ordinal interpretation (Table 1).

Table 1. K-Means Traffic State Characterization ($k=8$)

Cluster	Label	Hours (share)	Avg Records*/hr	≈Concurrent Aircraft	Avg Altitude (m)	Modal Hour (UTC)
0	Minimal Traffic – High Altitude	221 (3.7%)	33	3	5,054	11
1	Minimal Traffic – Cruise Overflight	221 (3.7%)	33	3	6,510	10
2	Very Low Traffic	217 (3.6%)	58	5	3,845	12
3	Low Traffic	893 (15.0%)	141	12	3,146	8
4	Moderate Traffic	1,459 (24.4%)	366	31	2,482	5
5	Moderate-High Traffic	559 (9.4%)	412	34	2,204	22
6	High Traffic	1,293 (21.6%)	710	59	2,159	2
7	Peak Traffic – Terminal Area	1,110 (18.6%)	1,096	91	1,822	18

* One record = one aircraft per five-minute window (Section 4.1.1); dividing by 12 approximates mean simultaneous aircraft.

4.1.4 Transformer Architecture

The forecasting model is an encoder-only Transformer with a linear classification head over the eight cluster IDs (795,272 trainable parameters). Input tokens (vocabulary size 8) are embedded into a 128-dimensional space with sinusoidal positional encoding (Vaswani et al., 2017). The model contains four encoder layers, each with eight attention heads, a feed-forward width of 512, GELU activation (Hendrycks & Gimpel, 2016), and dropout of 0.1. Forecasts are produced in a single forward pass. A shared learnable query token is appended at each of the H future positions, and the classification head reads out an eight-class probability distribution for each position. This non-autoregressive design gives every future hour its own positional identity while allowing all future positions to be predicted simultaneously. Three models sharing this architecture are trained independently, one per operational scale: the tactical model receives the two preceding Mondays plus the current Monday’s first 12 observed hours (60 tokens) and predicts the remaining 12 hours; the weekly model receives the four preceding Mondays (96 tokens) and predicts the next Monday (24 tokens); the strategic model receives the eight preceding Mondays (192 tokens) and predicts the next eight Mondays (192 tokens), corresponding to a lead time of approximately two months. Input windows are designed so that every input window includes complete daily cycles. The tactical model additionally receives same-day context unavailable to the longer-horizon models, a distinction that proves decisive in Section 4.2.1. Key architectural decisions are summarized in Table 2.

Table 2. Model Architecture Summary

Parameter	Value	Rationale
Model Type	Encoder-only Transformer + linear classification head	Long-range dependency capture
Vocabulary Size	8 cluster IDs	Traffic state tokens
Latent Dimension (d_model)	128	Compact representation
Attention Heads	8	Multi-pattern attention
Encoder Layers	4	Best validation accuracy per ablation
Feed-Forward Dim	512	4× expansion ratio
Dropout	0.1	Regularization
Activation	GELU	Smoother gradients vs. ReLU
Layer Normalization	Pre-LN (Xiong et al., 2020)	Training stability
Loss Function	Cross-Entropy + Label Smoothing ($\epsilon=0.1$)	Regularization
Total Parameters	795,272 (~0.8M)	Edge-deployable

Parameter	Value	Rationale
Wall-clock time (full pipeline)	64 min	2019 Intel laptop CPU (8 cores, 16 GB RAM), no GPU
Positional Encoding	Sinusoidal	Critical component (−19.5 points when removed)
Decoding	One-shot query-token readout (non-autoregressive)	All horizons in a single forward pass

4.1.5 Computational Footprint

A design goal of this framework is to make the complete research cycle executable on commodity hardware. The five-minute per-aircraft downsampling reduces the multi-gigabyte raw archives to a 519 MB filtered dataset; hourly aggregation reduces this further to a table of 5,973 rows that can be readily processed in memory. The full pipeline includes feature aggregation, clustering with model-selection sweep, training of all three forecasting models, five baselines, and nine ablation configurations across three random seeds. The pipeline completes in approximately 64 minutes on a 2019 Intel laptop CPU (8 cores, 16 GB RAM), without a GPU. All scripts, derived data, and result artifacts are released in the accompanying repository to support reproduction of the study.

4.1.6 Training Configuration and Evaluation

The 254 Mondays are split temporally: the first 177 (70%) for training, the next 38 (15%) for validation, and the final 39 (15%) for testing. All model selection, including early stopping (patience 10) and every ablation comparison, uses the validation set only. The test set is scored once per final model. Training uses AdamW (Loshchilov & Hutter, 2019) with a learning rate of 5×10^{-4} , weight decay of 0.01, cosine annealing over a maximum of 60 epochs, batch size of 32, and label smoothing of 0.1 (Szegedy et al., 2016). All reported results are the means and standard deviations over three random seeds.

The primary metric is per-hour token accuracy over observed target hours. Imputed hours are never scored. Because the eight states are ordinal in traffic volume, we additionally report within-one-level accuracy (a predicted state within one traffic level of the true state), macro-averaged precision, recall and F1 (Section 4.2.1), as well as the 8x8 confusion matrix (Table 4). Model comparisons use McNemar’s test (McNemar, 1947) on paired per-hour outcomes.

We evaluate the model against a baseline hierarchy of increasing strength: (i) chance (12.5% for eight classes); (ii) persistence, which carries the last input token forward; (iii) seasonal-naïve, which copies the most recent input Monday hour-by-hour (Hyndman & Athanasopoulos, 2021); (iv) hourly climatology, which predicts the training-set modal state for each hour of day without any learning; and (v) a two-layer LSTM with 128 hidden units (Hochreiter & Schmidhuber, 1997), trained under the same conditions as the Transformer. For the weekly and strategic tasks, carrying a single token forward may not be meaningful.

Persistence at these scales is therefore defined as repeating the most recent input Monday and consequently coincides with the seasonal-naive baseline, producing identical entries in Table 3. A month-conditioned climatology was also evaluated and found substantially weaker than the pooled version (51-53% versus 58-60% across tasks). With only 11-18 training Mondays per calendar month, the monthly estimates are based on limited data. We therefore use pooled hourly climatology as the strongest non-learned baseline. A learned model needs to outperform the baseline, not merely persistence.

4.2. Results

4.2.1 Multi-Scale Validation

Three findings emerge from Table 3. First, hourly climatology is a strong baseline across all three forecasting horizons (57.8-59.8%), outperforming the persistence baselines and seasonal-naive baseline. Its performance also remains stable as the forecasting horizon increases. Second, the advantage of the learned models (the Transformer and LSTM) appears at the tactical scale. The Transformer achieves 60.4% accuracy, 2.6 points above hourly climatology (57.8%). At the weekly scale, the Transformer’s advantage narrows to 0.4 points (60.2% vs. 59.8%, within seed noise). At the strategic scale, neither learned model meaningfully improves on climatology. These results suggest that learning provides the greatest benefit for within-day prediction, where same-day traffic information is available. Third, the Transformer and LSTM perform similarly at the tactical scale (60.4% vs. 60.2%) with no statistically significant difference (McNemar’s test on 453 paired tactical predictions, seed 0: the Transformer is correct where the LSTM errs on 27 hours and the reverse on 25, chi-squared = 0.02). Our choice of the Transformer as the framework backbone is therefore based on architectural considerations rather than a demonstrated accuracy advantage: it can attach external information such as weather forecasts or flight schedules to each individual hour, it handles the gaps between observed Mondays directly, and it produces the entire forecast in one step instead of hour by hour (Section 4.1.4).

Table 3. Multi-Scale Prediction Performance

Method	Tactical (12 h)	Weekly (7 d)	Strategic (~2 mo)
Chance	12.5%	12.5%	12.5%
Persistence	3.8%	57.3%	52.0%
Seasonal-naive	56.1%	57.3%	52.0%
Hourly climatology	57.8%	59.8%	59.6%
LSTM (2-layer)	60.2% $\pm$ 0.6	59.6% $\pm$ 1.3	59.7% $\pm$ 0.2
Transformer (ours)	60.4% $\pm$ 1.4	60.2% $\pm$ 0.6	58.4% $\pm$ 1.1

Note: Tactical persistence falls below chance because the last input token (11:00 UTC, an overnight low-traffic hour) is carried into busy daytime target hours. Within-one-level accuracy on the tactical task is 90.3% for the Transformer and 89.8% for hourly climatology (same 453 scored hours as Table 4).

Tactical top-1 accuracy of 60.4% does not fully characterize the model’s performance. For seed 0, the confusion matrix (Table 4) shows a within-one-traffic-level accuracy of 90.3%. Moreover, 75.4% of misclassifications occur between adjacent states, consistent with the boundary ambiguity quantified in Section 4.1.3 (silhouette 0.23).

The model’s contribution becomes clearer when performance is separated according to whether observed traffic follows or deviates from the hourly climatological pattern. Climatology correctly predicts 57.8% of tactical test hours (Table 3). On these hours, the Transformer achieves 87.8% accuracy. Among the remaining 42.2% of hours, the Transformer correctly identifies the state in 23% of cases when the observed state differs from the climatological prediction. Climatology cannot correctly predict these deviations. This ability to capture some departures from typical hourly patterns helps explain the Transformer’s tactical advantage over climatology.

Two limitations should be noted. First, the confusion matrix (Table 4) indicates class imbalance in model performance. Macro-averaged precision, recall and F1 are 0.277, 0.303 and 0.282 respectively, far below the overall accuracy of 60.4%. Second, Table 4 shows that the Transformer predicts only four of the eight states (3, 4, 6, and 7). States 0, 1, 2, and 5 are never predicted and together account for 6.8% of tactical test hours (20.4% of all observed hours). Reliable identification of these less frequent states may therefore require class rebalancing or a coarser state representation.

Table 4. Confusion matrix, tactical 12-hour horizon (seed 0, $n = 453$ scored hours). Rows are true states, columns predicted.

True \ Pred	0	1	2	3	4	5	6	7
0	0	0	0	2	0	0	0	0
1	0	0	0	4	0	0	0	0
2	0	0	0	10	0	0	0	0
3	0	0	0	38	11	0	1	0
4	0	0	0	15	24	0	7	11
5	0	0	0	2	2	0	10	1
6	0	0	0	0	16	0	44	69
7	0	0	0	0	0	0	18	168

4.2.2 Ablation Study

The ablation results in Table 5 clarify which components contribute most to tactical performance. First, the results show that historical Mondays alone achieve 59.5% accuracy, while same-day morning context alone achieves 56.9%. Combining both history Mondays and same-day morning context yields the highest accuracy at 60.4%, suggesting that each provides useful information for tactical prediction. The largest performance decline occurs when

positional encoding is removed, reducing accuracy by 19.5 percentage points to 40.9% and demonstrating the importance of temporal position. By comparison, the remaining architectural choices have relatively small effects. Using a single attention head reduces accuracy by 2.1 points. Changing the number of encoder layers from four to two or eight reduces accuracy by 0.8 and 1.5 points, respectively. Replacing GELU with ReLU produces no change in mean accuracy, while removing label smoothing results in a slight increase of 0.2 points. Overall, these results suggest that input information and temporal position have a greater influence on performance than the other architectural choices. Several of these differences fall within seed-to-seed variation, so at 175 training samples (177 training Mondays minus the two context-only Mondays at the sequence start) they should be read as null results rather than evidence for any particular architectural choice.

Table 5. Ablation Results (Tactical 12-Hour Horizon)

Configuration	Accuracy (mean $\pm$ std, 3 seeds)	Δ vs full
Full model (history + same-day morning)	60.4% $\pm$ 1.4	—
Input: same-day morning only	56.9% $\pm$ 2.0	−3.5
Input: history Mondays only	59.5% $\pm$ 0.4	−0.9
No positional encoding	40.9% $\pm$ 0.2	−19.5
Single attention head	58.4% $\pm$ 0.4	−2.1
2 encoder layers	59.6% $\pm$ 1.0	−0.8
8 encoder layers	58.9% $\pm$ 0.5	−1.5
ReLU instead of GELU	60.4% $\pm$ 0.3	0.0
No label smoothing	60.6% $\pm$ 0.7	+0.2

4.2.3 Temporal Error Analysis

Tactical-task errors vary across time. Across the forecast window, accuracy is lowest during the overnight-to-morning transition (44.7% at 14:00 UTC, approximately 06:00 local) and highest in the stable mid-morning operating period (76.3% at 18:00 UTC, approximately 10:00 local). This pattern suggests that transitions between traffic states are more difficult to predict than relatively stable periods. Seasonally, winter test Mondays have the lowest accuracy (47.7% in December-February), while summer Mondays have the highest (83.3% in June-August, although the summer estimate is based on limited observations). These seasonal differences are consistent with weather-driven disruption of typical traffic patterns and motivate the weather-conditioned extension discussed in Section 6.

5. Additional Use Cases for Small Localized Transformers in Aviation Safety

The San Francisco Bay Area terminal region application demonstrates the general methodology. We now describe four additional use case categories. The first three apply the

forecasting pipeline directly; the fourth extends it with the downstream overlay mechanism of Section 3.5.

5.1 Facility-Specific Airport Applications

The current study region includes traffic associated with San Francisco International Airport, San Jose Mineta International Airport, and Oakland International Airport. The present model treats this area as a regional airspace rather than modeling each airport separately. A natural extension is to apply the same pipeline to individual facilities using tighter, facility-centered geographic boundaries. Historical ADS-B data from the OpenSky Network can be filtered to each facility's bounding box. The model can then be retrained on the resulting local traffic patterns without changing the underlying architecture. This facility-specific implementation could reveal patterns that are less visible when traffic across the broader San Francisco Bay Area is modeled together. These patterns may reflect local operating conditions, such as runway use, approach patterns, and recurring airline schedules.

5.2 Mountainous Terrain Corridors and Complex Airspace

Airports in mountainous terrain (e.g., Aspen-Pitkin County Airport, Jackson Hole Airport, and facilities in remote areas of Alaska) operate under spatiotemporal constraints that differ from those at sea-level hub airports. Terrain can restrict approach and departure paths, while rapidly changing weather can further limit operational flexibility. These conditions make the combination of local traffic patterns, terrain, and environmental factors particularly relevant to safety analysis. The proposed framework could be adapted to these airports using localized ADS-B data. A facility-specific model could predict periods of elevated traffic and changes in traffic patterns before they occur. Such forecasts could then be considered alongside terrain, weather, and other local factors to support safety analysis in complex operating environments.

5.3 General Aviation and Non-Towered Airports

Non-towered airports represent a distinct setting for aviation safety because traffic separation relies heavily on pilot awareness, communication, and established operating procedures rather than local air traffic control service. These airports lack dedicated air traffic control resources and sophisticated surveillance infrastructure. A lightweight, localized forecasting model could provide an additional source of information by identifying periods of elevated traffic before they occur. Such forecasts could potentially be delivered through a web interface or mobile application to supplement the existing pilot communication and situational-awareness practices. Recent work by Prutsch et al. (2026) indicates the feasibility of lightweight Transformer-based trajectory prediction in non-towered terminal airspace, providing support for further exploration of computationally efficient predictive models in this setting.

5.4 Environmental Risk Overlays: Wildlife Strikes and Beyond

Wildlife activity provides one example of how external safety information could be incorporated without retraining the traffic forecasting model. The FAA's Wildlife Strike Database

records over 17,000 wildlife strikes annually (FAA, n.d.-c), with birds accounting for the majority of reported events. Under the density-overlay architecture proposed in Section 3.5, historical wildlife activity could be analyzed separately and then considered alongside the Transformer’s predicted traffic states. Periods in which elevated traffic coincides with greater wildlife activity could then be identified for further safety assessment. For example, historical strike records could be aggregated by season and time of day to characterize when wildlife encounters are more frequently reported. These patterns could then be compared with predicted High or Peak Traffic states (6–7). The resulting overlap would not itself represent a prediction of wildlife-strike risk, but it could identify periods that warrant additional attention. Such information could potentially support wildlife-management planning at airports where wildlife activity is an important operational concern.

The same approach could extend beyond wildlife. Volcanic ash advisories, space weather alerts, significant meteorological information, or near-real-time runway condition reports could be considered alongside predicted traffic states at the spatial and temporal resolution appropriate to each source. These additional layers would provide contextual information for subsequent safety analysis without requiring the core traffic forecasting model to be retrained.

6. Limitations and Future Directions

Several important limitations bound the current work. First, the study covers a single airport region; the transferability claims of Section 3.4 remain to be demonstrated empirically. Second, the public OpenSky archive samples Mondays only. The models are trained and evaluated exclusively on observed Monday hours, and generalization to other days of the week is untested. The framework models the Monday-to-Monday changes in air traffic rather than a continuous seven-day week. Third, continuous air traffic conditions do not always fall neatly into eight distinct states. The relatively low silhouette score (0.23 at $k = 8$) indicates that the clusters are not clearly separated. As a result, two hours with similar traffic conditions may be assigned to adjacent states. A prediction of state 6 instead of state 7, for example, is counted as incorrect even though the two states may represent similar traffic conditions. The high within-one-level accuracy is consistent with this pattern, but further analysis is needed to determine how much prediction error is attributable to these cluster boundaries. Fourth, the test period (late September 2021 to June 2022) falls in the post-pandemic traffic recovery, representing a potential distribution shift relative to much of the training period. The reported performance should be interpreted with this distribution shift in mind. Fifth, the tactical model has only 175 training samples. Architecture comparisons, including the observed Transformer and LSTM performance, should therefore be interpreted as findings at this data scale rather than general conclusions about the architectures. Sixth, the K-Means traffic state representation is interpretable and computationally efficient but discards fine-grained spatial information. Future work could incorporate spatial attention or geographic embeddings to retain more spatial detail while preserving the interpretability of discrete traffic states. Seventh, we do not empirically compare the proposed framework with a large multi-airport model. The case for localization

therefore rests on computational accessibility, resource efficiency, and interpretability rather than a demonstrated accuracy advantage over large-scale alternatives. Moreover, although predicted traffic states may provide useful context for downstream safety analysis, their relationship with specific safety outcomes is not evaluated in this study.

Several of these limitations point to concrete next steps. Adding external information (e.g., hourly weather, published airline schedules, and calendar features such as holidays) is the clearest path to improving on climatology at longer horizons. The concentration of errors in winter and at transitions between traffic states (Section 4.2.3) also suggests that weather might be the most important factor missing from the model. Full-week data sources such as OpenSky research access would remove the Monday-only restriction. Training the model at a second airport would directly test whether the framework transfers to new locations. Finally, predicting probability distributions over traffic states or continuous traffic volumes, rather than a single state, could provide a more fine-grained representation of traffic dynamics and uncertainty.

7. Conclusion

This paper has presented a lightweight Transformer framework for air traffic forecasting and demonstrates the potential value of geographically localized models for regional aviation safety research. Against a strong hourly climatology baseline, the compact model (0.8M parameters) shows its clearest advantage at the tactical scale, where same-day context is available, achieving 60.4% accuracy compared with 57.8% for climatology.

The core methodological contributions include a Monday-sequence study design that avoids constructing a synthetic continuous timeline, a strict input/evaluation separation for missing data, a baseline hierarchy that includes a strong non-learned reference, and an end-to-end pipeline that runs on a laptop in about an hour. Collectively, these features demonstrate that rigorous multi-scale spatiotemporal forecasting research can be conducted with modest computational resources. More broadly, the framework treats geographic localization as a deliberate design choice. Its focus on local traffic patterns, interpretable state representation, and modular incorporation of external factors demonstrates the potential of localized approaches for aviation safety research.

The use cases described in Section 5 illustrate how the framework could be extended to other aviation settings, including individual airport facilities, mountainous terrain corridors, non-towered general aviation fields, and areas with elevated wildlife activity. In each case, the same basic pipeline applies: collect localized ADS-B data, group hourly air traffic conditions into operationally meaningful states, and train a compact Transformer to forecast those states. This combination of low computational requirements, interpretability, and adaptability highlights the potential value of compact, localized models for applied aviation safety research.

Acknowledgments

The authors thank the OpenSky Network community for maintaining open-access ADS-B historical data. Flight trajectory data used in this study was obtained from the OpenSky Network (Schäfer et al., 2014) and is used in accordance with the OpenSky Network's General Terms of Use and Data License Agreement for non-profit research purposes. This project received support from the Center for Science Engagement. All experiments reported in this paper were conducted end-to-end on commodity hardware. The code, derived data, and result artifacts are available at <https://github.com/eithnezh/sfo-traffic-density-forecast> and are archived at <https://doi.org/10.5281/zenodo.22741638>, enabling reproduction of reported numbers. The authors used large language models (ChatGPT-5.5, OpenAI; Claude Sonnet 5, Anthropic) as writing assistants during manuscript revision, for copy-editing, rewording, and checking the internal consistency of the text. All research design, data processing, experiments, code, and results are the authors' own work. The authors reviewed and edited all AI-assisted text and take full responsibility for the content of the paper.

References

- Abdulkhak, S., Hubbard, W., Gopalakrishnan, K., & Li, M. Z. (2024). CHATATC: Large language model-driven conversational agents for supporting strategic air traffic flow management. *arXiv preprint arXiv:2402.14850*.
- Belz, T., Gordon, A., Murphy, C., Reixa, N., & Wulz, A. (2025, April 2). *Air travel demand outlook 2025: Steady growth despite supply constraints*. Boston Consulting Group. <https://www.bcg.com/publications/2025/air-travel-demand-outlook>
- Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224–227.
- Du, W., Han, L., Xiong, Y., Zhang, L., Li, B., Lv, Y., & Guo, T. (2026). Airport Passenger Flow Forecasting via Deformable Temporal–Spectral Transformer Approach. *IEEE Transactions on Intelligent Transportation Systems*.
- European Union Aviation Safety Agency. (2023). EASA artificial intelligence roadmap 2.0: A human-centric approach to AI in aviation. <https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-roadmap-20>
- Federal Aviation Administration. (n.d.-a). A brief history of the FAA. *U.S. Department of Transportation*. https://www.faa.gov/about/history/brief_history
- Federal Aviation Administration. (n.d.-b). Runway incursions. *U.S. Department of Transportation*. https://www.faa.gov/airports/runway_safety/resources/runway_incursions

- Federal Aviation Administration. (n.d.-c). *Wildlife strike database*. U.S. Department of Transportation. <https://wildlife.faa.gov>
- Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, 69(1), 140-150.
- Guo, D., Zhang, Z., Yan, Z., Zhang, J., & Lin, Y. (2024, March). FlightBERT++: a non-autoregressive multi-horizon flight trajectory prediction framework. *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 38, No. 1, pp. 127-134).
- Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. <https://otexts.com/fpp3/>
- International Air Transport Association. (2025a, January 30). *Global air passenger demand reaches record high in 2024*. <https://www.iata.org/en/pressroom/2025-releases/2025-01-30-01/>
- International Air Transport Association. (2025b, February 21). *Air passenger demand forecasting: The future of global air travel*. <https://www.iata.org/en/publications/newsletters/iata-knowledge-hub/air-passenger-demand-forecasting-the-future-of-global-air-travel/>
- Kim, Y. J., Choi, S., Briceno, S., & Mavris, D. (2016). A deep learning approach to flight delay prediction. *AIAA/IEEE Digital Avionics Systems Conference*.
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations*.
- Luo, A., Luo, Y., Liu, H., Du, W., Wu, X., Chen, H., & Yang, H. (2024). An improved transformer-based model for long-term 4D trajectory prediction in civil aviation. *IET Intelligent Transport Systems*, 18(9), 1588-1598.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.

- National Transportation Safety Board. (2018). Taxiway overflight Air Canada Flight 759, Airbus A320-211 (NTSB/AIR-18/01). <https://www.nts.gov/investigations/AccidentReports/Reports/AIR1801.pdf>
- Pang, Y., Kendall, A. P., Porcayo, A., Barsotti, M., Jain, A., & Clarke, J. P. (2026). From voice to safety: language ai powered pilot-atc communication understanding for airport surface movement collision risk assessment. *Transportation Research Part C: Emerging Technologies*, 184, 105540.
- Prutsch, A., Schinagl, D., & Possegger, H. (2026). ASCENT: Transformer-Based Aircraft Trajectory Prediction in Non-Towered Terminal Airspace. *arXiv preprint arXiv:2603.16550*.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- Schäfer, M., Strohmeier, M., Lenders, V., Martinovic, I., & Wilhelm, M. (2014). Bringing up OpenSky: A large-scale ADS-B sensor network for research. *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks (IPSN)*, 83–94.
- Skowronek, M. (2025). *Enhancing profitability in the air transport industry through improved air passenger forecasting: a comparative analysis of ARIMA, Holt-Winters and LSTM time series forecasting techniques*. Mississippi State University.
- Sun, M., Tian, Y., Li, J., Wu, C. L., Peng, L., & Xu, S. (2025). A review of network delay prediction and advances in large language models for air traffic. *Artificial Intelligence Review*, 59(1), 36. <https://doi.org/10.1007/s10462-025-11400-w>
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.
- The OpenSky Network. (n.d.). *Weekly 24 hours of state vector data 2017–2022* [Data set]. <https://opensky-network.org/data/scientific>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, E., Xu, H., Yu, N., Liu, F., Ji, G., Xu, S., ... & Chen, Y. (2026). MSTAGNN-MARL: A Multi-Level Intelligent Decision Framework for Integrated Spatial-Temporal Conflict Resolution in High-Density Airspace. *Aerospace*, 13(2), 175.
- Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., ... & Liu, T. (2020, November). On layer normalization in the transformer architecture. In *International conference on machine learning* (pp. 10524-10533). PMLR.

- Xu, S., Chan, H.K., & Zhang, T. (2019). Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach. *Transportation Research Part E: Logistics and Transportation Review*.
- Xu, X., Li, J., Mao, R., & Shi, Z. (2021). A graph-based deep learning method for short-term airport surface traffic flow forecasting. *Transportation Research Part C*, 124, 102963.
- Yin, J., Liu, J., Zeng, W., Wei, W., Tian, W., & Zhang, Y. (2025). Spatiotemporal Dynamic Graph Convolutional Networks for Air Route Network Congestion Prediction. *Journal of Aerospace Information Systems*, 22(3), 177-188.
- Yoon, S., & Lee, K. (2025). Aircraft trajectory prediction with inverted transformer. *IEEE Access*, 13, 26318–26330.
- Zhang, T., Hu, S., Wang, C., & Zhong, Q. (2026). Real-time runway safety surveillance using deep learning based video analytics. *Discover Computing*, 29(1), 72.
- Zhang, X., Zhong, S., & Mahadevan, S. (2022). Airport surface movement prediction and safety assessment with spatial–temporal graph convolutional neural network. *Transportation research part C: emerging technologies*, 144, 103873.