Airport Delay Prediction with Temporal Fusion Transformers

Ke Liu
liuke126@berkeley.edu
University of California Berkeley

Kaijing Ding
University of California Berkeley

Xi Cheng
University of Illinois Chicago

Guanhao Xu
Oak Ridge National Laboratory

Xin Hu
University of Michigan—Ann Arbor

Tong Liu
University of Illinois
Urbana-Champaign

Siyuan Feng
Hong Kong University of Science and Technology

Binze Cai
Georgia Institue of Technology

Jianan Chen
University of British Columbia

Hui Lin
Northwestern University

Jilin Song
University of Toronto

Chen Zhu
Tsinghua University

Abstract

Since flight delay hurts passengers, airlines, and airports, its prediction becomes crucial for the decision-making of all stakeholders in the aviation industry and thus has been attempted by various previous research. However, previous delay predictions are often categorical and at a highly aggregated level. To improve that, this study proposes to apply the novel Temporal Fusion Transformer model and predict numerical airport arrival delays at quarter hour level for U.S. top 30 airports. Inputs to our model include airport demand and capacity forecasts, historic airport operation efficiency information, airport wind and visibility conditions, as well as en-route weather and traffic conditions. The results show that our model achieves satisfactory performance measured by small prediction errors on the test set. In addition, the interpretability analysis of the model outputs identifies the important input factors for delay prediction.

CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

Airport Delay Prediction, Temporal Fusion Transformer (TFT), Weather Impact on Aviation

ACM Reference Format:
Ke Liu, Kaijing Ding, Xi Cheng, Guanhao Xu, Xin Hu, Tong Liu, Siyuan Feng, Binze Cai, Jianan Chen, Hui Lin, Jilin Song, and Chen Zhu. 2024. Airport Delay Prediction with Temporal Fusion Transformers. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (IWCTS '24).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The aviation industry, despite experiencing a significant downturn in air traffic demand during the global pandemic, is now on a path to recovery. Projections indicate an annual growth in traffic of 1.5% to 3.8% over the next two decades [6]. Such growth, while promising, forecasts a burgeoning gap between demand and capacity at airports, potentially exacerbating flight delays — an outcome that significantly impacts passenger satisfaction, airline operating costs [14], and environmental sustainability [12].

To mitigate these challenges, it is increasingly vital for aviation authorities to develop robust mechanisms for predicting flight delays and to establish more efficient traffic management initiatives (TMIs). The literature is replete with studies aimed at forecasting flight delays using a variety of methodologies ([10, 11, 13]). One of the seminal works in this field by ([7]) employed deep learning techniques, specifically the Long Short Term Memory (LSTM) model, to analyze day-to-day sequences of departure and arrival delays at a single airport, successfully predicting delay classes based on predefined thresholds. Another noteworthy study by ([15]) utilized the LSTM model to predict aggregated daily delays for 123 U.S. airports, incorporating Monte Carlo Dropout techniques to refine parameter variance estimates. However, most existing studies in flight delay prediction focus on binary outcomes (delayed or not) or on categorizing delays into broad classes. Furthermore, these predictions often pertain to highly aggregated levels, such as daily delay forecasts.

Building on recent advancements, a cutting-edge study ([16]) demonstrated the application of transformers to predict airport delays at the quarter-hour level across several regional airports. This has paved the way for the adoption of more sophisticated models capable of tackling aviation challenges with greater precision.

In this paper, we propose a novel approach to predict the specific numerical values of airport arrival delays at a more granular level—specifically, every quarter-hour over a strategic horizon of

IWCTS '24, October 29th, 2024, Atlanta, GA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

up to four hours. We focus on the top 30 U.S. airports, incorporating variables such as airport demand and capacity forecasts, historical operational efficiency, and local weather conditions. To achieve this, we will deploy the Temporal Fusion Transformer (TFT), an attention-based deep neural network model designed for multi-horizon forecasting. The TFT model has demonstrated superior performance over other forecasting techniques like DeepAR, ARIMA, and traditional LSTM Seq2Seq across various datasets ([1]), validating its effectiveness for time-series data analysis.

The paper is structured as follows: Section 2 provides an overview of the datasets utilized and the preprocessing of input variables. Section 3 delves into the modeling details, while Section 4 presents and discusses the results. We conclude our study in Section 5 with a summary of our findings and insights.

2 Data

For this study, we focused on the top 30 busiest airports in the U.S. during the year 2016, spanning from January 1st to December 31st. To conduct a comprehensive analysis, we collected and integrated three key datasets from the Federal Aviation Administration's (FAA) Aviation System Performance Metrics (ASPM) and the Integrated Surface Database (ISD). Prior to integration, each dataset underwent rigorous cleaning, filtering, and time zone normalization to UTC to ensure data consistency and accuracy. The resulting master database is structured on a quarter-hourly basis for each airport, amounting to a total of 1,054,080 data points (30 airports × 366 days × 24 hours per day × 4 quarters per hour).

The datasets utilized include:

1. **FAA ASPM Flight Level Data:** This dataset provides detailed records for each flight, including flight plans, scheduled and actual times, and Estimated Departure Clearance Time (EDCT) for flights arriving at 77 major U.S. airports. It serves as a crucial source for analyzing flight-specific delay patterns and operational efficiency.
2. **FAA ASPM Airport Quarter-Hour Data:** Contains comprehensive data on operational conditions at 15-minute intervals. This dataset includes information on airport capacity, runway configurations, and terminal weather conditions, essential for understanding the operational dynamics that influence flight delays at the airport level.
3. **Global Hourly – Integrated Surface Database (ISD):** Comprises hourly weather records from 2,330 surface stations across the U.S., providing extensive meteorological data crucial for correlating weather conditions with flight delays. The details of the stations and their data coverage are documented in **Appendix 1(c)**.

This robust integration of flight, airport operational, and weather data provides a solid foundation for developing predictive models aimed at forecasting airport delays with high precision. By harnessing detailed historical data, our approach seeks to uncover nuanced relationships between airport operations and delay occurrences, enabling more accurate and timely predictions that could significantly enhance traffic management and operational planning at major U.S. airports.

This section outlines the input variables for airport-level delay prediction and describes their processing from raw datasets. The primary data sources are from the FAA's ASPM airport quarter-hour dataset. Inputs include:

Airport ID: Captures systematic variations in traffic management efficiency and congestion at different airports. Time Index, Month, Local Hour, and Day of the Week: Helps capture general traffic volume and flight operation patterns. Scheduled Flight Departure and Arrival Counts: Set 2-6 months in advance of the actual flight day. Airport Capacity Forecasts for Departures and Arrivals: Predictive data on expected airport throughput. Observed Airport Arrival and Departure Throughput: Actual counts of flights handled by the airport. Observed Airport Demand: The number of flights scheduled to arrive and depart. On-Time Percentage for Arrivals and Departures: Statistics on the punctuality of flights. Average Arrival/Departure Delays: Prediction variable, smoothed by moving average over adjacent three quarter hours to mitigate extreme values. Additionally, to assess congestion effects, we calculate cumulative queuing delays for arrivals and departures using a deterministic queuing model. This model operates with quarter-hour intervals, counting demand based on scheduled 'wheels-on' times and throughput based on actual 'wheels-on' times. If a flight lands earlier than scheduled, it is only counted in the earlier interval. We then compute the cumulative actual arrivals and demands for each quarter-hour, where the area between the arrival and demand curves indicates total queuing delays for that day.

For operational variables, we include:

- Enroute Traffic Density: This considers the traffic managed by terminal air traffic controllers, who are responsible for aircraft entering and exiting the airport and ensuring safe separation over the busy surrounding airspace. The U.S. airspace is divided into 0.25° cells, extending from 25°N to 50°N latitude and 66°W to 125°W longitude. Using distance-based interpolation along the great circle route, we track en-route locations at quarter-hour intervals, aggregating traffic density across the grid. An example of this grid system is shown in **Appendix 1(a)(b)**.
- Convective Weather Factors: Thunderstorms, crosswinds, tailwinds, ceiling, and visibility at airports are considered. Thunderstorms, in particular, can drastically reduce airport capacity. We resample weather data hourly and apply 2D grid interpolation to assign weights to grid cells for convective weather, creating hourly weather matrices for the airspace grid (see [12] for detailed feature engineering), as displayed in **Appendix 1(d)**.

This detailed input processing approach allows for a nuanced understanding of factors influencing airport delays, ensuring our predictive models are both comprehensive and precise.

3 Modeling

3.1 Temporal Fusion Transformer

Temporal Fusion Transformer (TFT) is an attention-based architecture which combines high-performance multi-horizon forecasting with interpretable insights into temporal dynamics, first proposed by [8]. It integrates the mechanisms of several other neural architectures we learned in class, for instance LSTM layers and the attention heads used in Transformers. The major components of TFT include gating mechanisms, variable selection networks, static

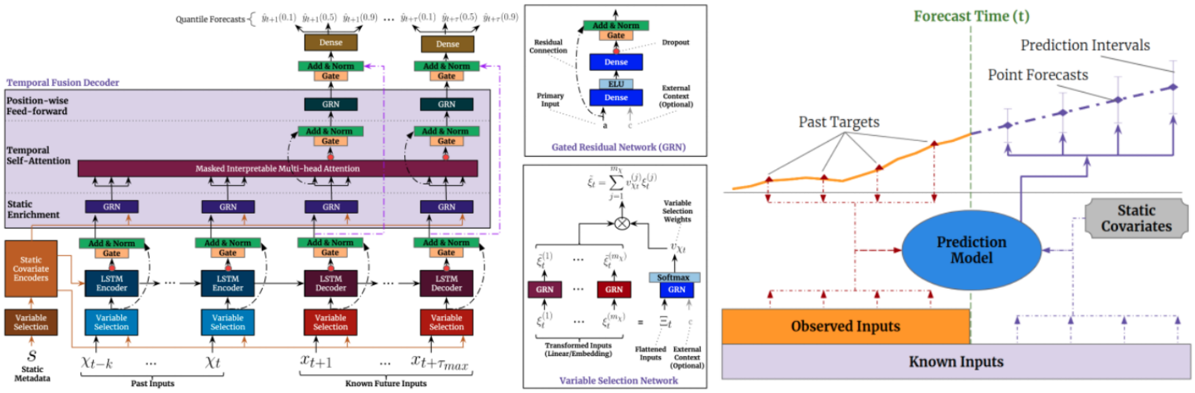


Figure 1: Left TFT Architecture ([8]; Right:Illustration of multi-horizon forecasting([8])

covariate encoders, temporal processing, and prediction intervals via quantile forecasts. Details can be found in [8] and the general architecture is shown in Fig. 2.

Compared with the classic transformer model, it has several advantages and novelties, which makes it a good fit for our work. First, the input features to TFT can be of three types: i) temporal data with known inputs into the future, e.g., the future airport demands; ii) temporal data known only up to the present, e.g., the historical airport delays; and iii) exogenous categorical/static variables, also known as time-invariant features, e.g. airport IDs. Second, it supports multi-step predictions. This characteristic facilitate our prediction of airport arrival delays for different quarter hours in the future. Furthermore, TFT also outputs prediction intervals, by using the quantile loss function. Therefore, TFT can provide range estimates rather than a single point estimate, which will offer air traffic controllers more information when they are making decisions. In addition, TFT has good interpretability, which can help identify the key contributions to airport delays. Together with its advantages in high performance and available open source implementations, all these strengths makes it a preferrable model for us to implement delay prediction.

3.2 Problem Formulation

The goal of this study is to predict arrival delays for the U.S. Top 30 airports over a strategic time horizon. In a generic form, it can be formulated as a multi-variate, multi-step, time series forecasting problem. Here we determined the time lag variable, or the look-back time of what has happened as 2 hours, or 8 time steps (since our data is in quarter hours); and the maximum look-ahead time as 4 hours, or 16 time steps, which is determined by the data updating frequency as well as prediction needs. Fig. 2 further illustrates the relationship of inputs and outputs in a given time horizon. The inputs and the outputs of our TFT model are prepared from the sources specified in Section 2. Table 1 lists the detailed inputs to TFT for delay prediction.

With these inputs, we use the TemporalFusionTransformer model given by PyTorch Forecasting package to train our flight delay forecast model. The model is trained and validated on the first eleven and a half months and the last 15 days of the year are set aside as test set. The hyperparameters are finely tuned, including but not limited to size of the hidden layers, dropout rate, attention head size, and learning rate; and the parameters are learned.

Table 1: Input Variables Description

Variable Types	Variables
Static Covariates	Airport ID, Month, Local hour, Day of the week
Past-Observed Inputs	Actual arrival and departure counts; Reported number of aircraft intending to arrive and depart; Average arrival delays (based on flight plan) and departure delays; Arrival and departure on-time percentage statistics; Cumulative arrival and departure queuing delays
Apriori-Known Inputs	Scheduled quarter-hourly arrival and departure demand; Quarter-hourly airport arrival and departure capacity; Airport visibility and ceiling conditions; Airport headwind and tailwind conditions; Enroute convective weather; Enroute traffic densities

4 Results

4.1 Delay Prediction Results

Fig. 2 shows the model performance for the 30 airports on the testing dataset of the last 15 days of 2016, measured by mean absolute error(MAE). The performance of the model varies among different airports, with MAE ranging from 5 minutes to 12 minutes. In general, airports with higher delays have a higher MAE.

To understand the temporal differences of the prediction performance, Fig. 3 show the prediction results of arrival delays at SFO for selected testing days. The comparison of actual (blue line) with predicted (orange line) values demonstrates that the TFT model can capture most upward and downward trends of delays.

However, the model still does not make predictions that match the ground truth perfectly. Possible causes are also examined. First, although we have moved average the delay data, it still has a huge variation within adjacent quarter hours. This unsmoothness characteristic of the target variable makes it difficult to predict. Second,

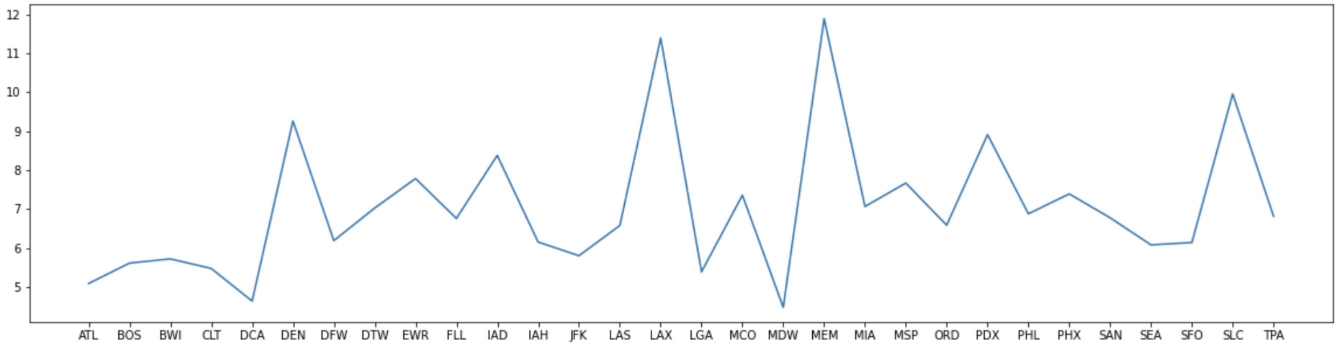


Figure 2: Model Performance for 30 Airports Measured by MAE(min)

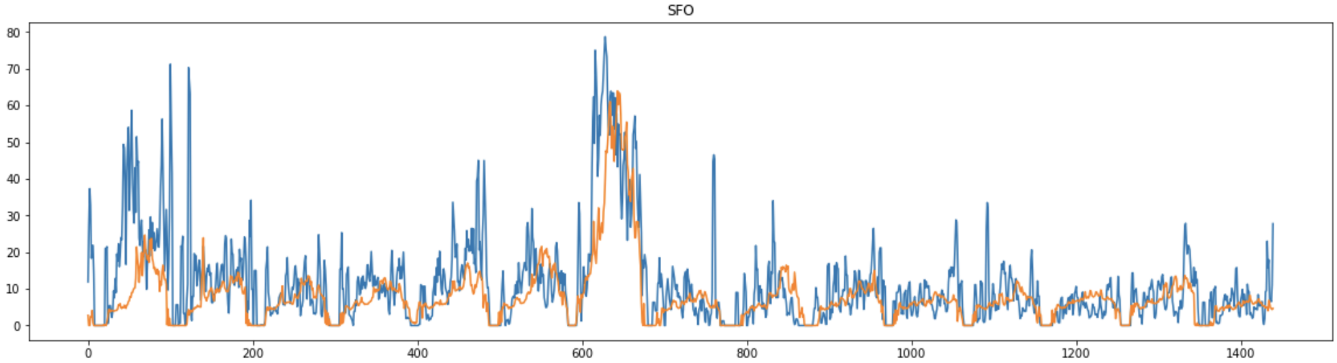


Figure 3: Model prediction of arrival delays(min) at SFO for selected testing days

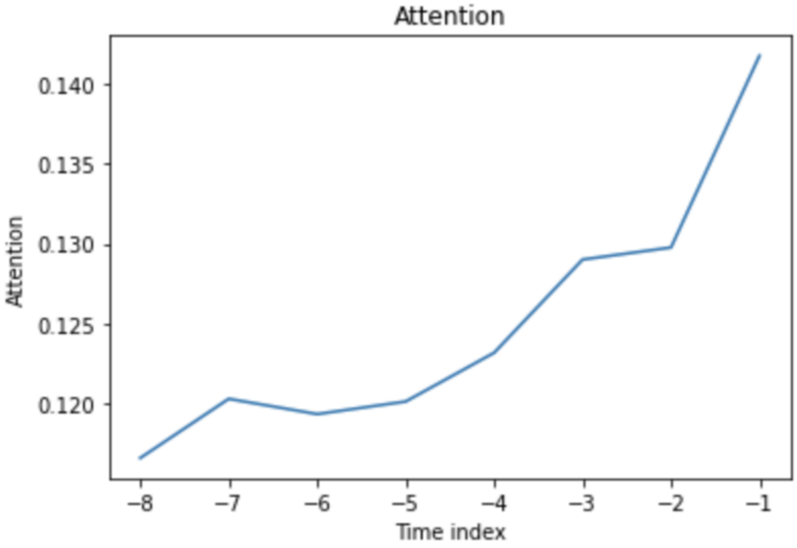


Figure 4: Attention score by time index

this study applies the model to all U.S. core 30 airports while the different airports may be influenced by the inputs differently. For example, SFO may be more impacted by bad weathers than LAX is, which problem, however, cannot be handled by the setting of this model. Third, there are other factors influencing flight delays that are not captured by the dataset we collect. For example, we do not

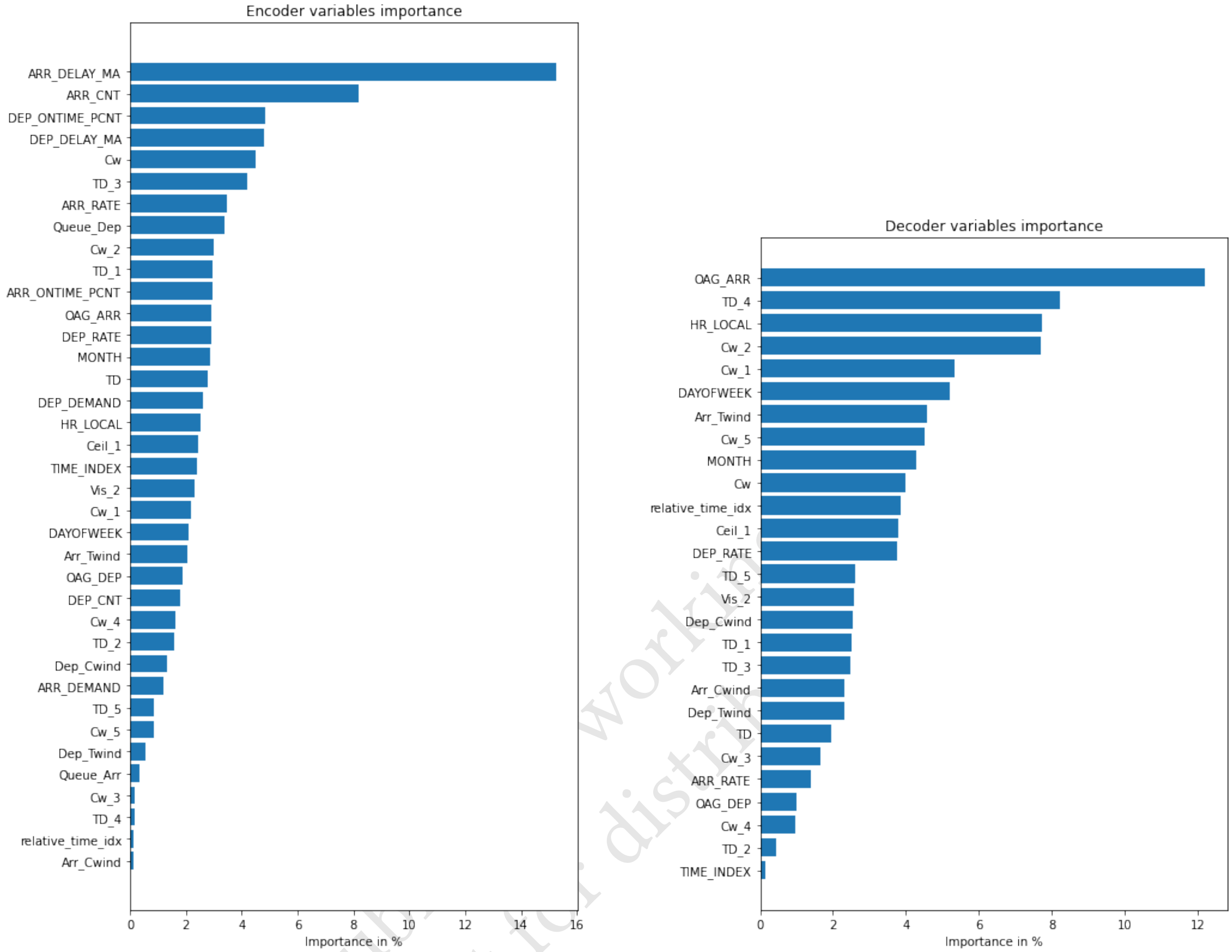


Figure 5: (Left) Encoder; (Right) Decoder

have access to traffic management data like Ground Delay Program implementation or Airspace Flow Program implementation, which can better capture the traffic impact of other airports and enroute airspace on the study airport. Last but not least, the training and test dataset may not be of the same distribution. The second half of December (test set) is likely to have more severe weather days than the rest of the year. Due to the imbalanced nature of the data, the model would perform better (in terms of error minimization) on the non-severe weather days.

4.2 Model Interpretation

TFT has an advantage in interpretability of time dynamics. For example, Fig. 4 shows the attention score of each time index. A higher attention score means a bigger contribution from that time step to the prediction outputs. Since the look-back time of our model is set to 2 hours, or 8 time steps, the time indices are from -8 to -1. As expected, the closer time steps receive higher scores and thus make a bigger contribution.

Furthermore, the TFT model is also interpretable in terms of the importance of each input variable for the response variables. Based on Fig. 5, the relative importance of the input variables for delay prediction can be identified. For the encoder inputs, historic arrival delays and arrival counts are top 2 factors. For the decoder input variables, scheduled arrival demands, enroute traffic density, local hour, enroute convective weather are influential factors.

5 Conclusion

In this study, we applied Temporal Fusion Transformer models to the prediction of average flight arrival delay of U.S. core 30 airports at quarter hour interval with the maximum look-ahead time as 4 hours(or 16 time steps). The inputs of model comes from a wide range of data and are processed to involve the information of airport operation conditions (capacity, schedule flights, demand), terminal and nearby enroute congestion level, airport weather conditions and nearby enroute thunderstorm. The TFT model handle well with these heterogeneous inputs. The performance of the model, though varies among airports, is acceptable in that it captures most upward and downward trends of delays. Future research could focus on enhancing the processing of convective weather information, incorporating enroute wind conditions, and, most importantly, integrating Traffic Management Initiatives (TMIs), such as miles-in-trail [9] and the ground delay program. Furthermore, future work could explore the integration of platooning [4], electrification [2, 5], and queueing systems [3] in aviation applications.

References

[1] Sercan O. Arik and Tomas Pfister. 2021. Interpretable Deep Learning for Time Series Forecasting. https://ai.googleblog.com/2021/12/interpretable-deep-learning-for-time.html

[2] Xi Cheng and Jane Lin. 2024. Is electric truck a viable alternative to diesel truck in long-haul operation? *Transportation Research Part D: Transport and Environment* 129 (2024), 104119.

[3] Xi Cheng, Theodoros Mamalis, Subhonmesh Bose, and Lav R Varshney. 2024. On Carsharing Platforms With Electric Vehicles as Energy Service Providers. *IEEE Transactions on Intelligent Transportation Systems* (2024).

[4] Xi Cheng, Yu Marco Nie, and Jane Lin. 2024. An autonomous modular public transit service. *Transportation Research Part C: Emerging Technologies* (2024), 104746.

[5] Xi Cheng, Hui Shen, Yantao Huang, Yi-Ling Cheng, and Jane Lin. 2024. Using Mobile Charging Drones to Mitigate Battery Disruptions of Electric Vehicles on Highways. In *2024 Forum for Innovative Sustainable Transportation Systems (FISTS)*. IEEE, 1–6.

[6] IATA. 2019. 20 Year Passenger Forecast. https://www.iata.org/en/publications/store/20-year-passenger-forecast/

[7] Young Jin Kim, Sun Choi, Simon Briceno, and Dimitri Mavris. 2016. A deep learning approach to flight delay prediction. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. IEEE, 1–6.

[8] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. 2019. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. https://doi.org/10.48550/ARXIV.1912.09363

[9] Ke Liu and Mark Hansen. 2021. Miles-in-Trail Restrictions and Aviation System Performance: Chicago O'Hare Case Study. In *Fourteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2021)*.

[10] Ke Liu and Mark Hansen. 2024. Excess Delay from GDP: Measurement and Causal Analysis. *arXiv preprint arXiv:2405.11211* (2024).

[11] Ke Liu, Fan Hu, Hui Lin, Xi Cheng, Jianan Chen, Jilin Song, Siyuan Feng, Gaofeng Su, and Chen Zhu. 2024. Deep Reinforcement Learning for Real-Time Ground Delay Program Revision and Corresponding Flight Delay Assignments. *arXiv preprint arXiv:2405.08298* (2024).

[12] Ke Liu, Zhe Zheng, Bo Zou, and Mark Hansen. 2023. Airborne flight time: A comparative analysis between the US and China. *Journal of Air Transport Management* 107 (2023), 102341.

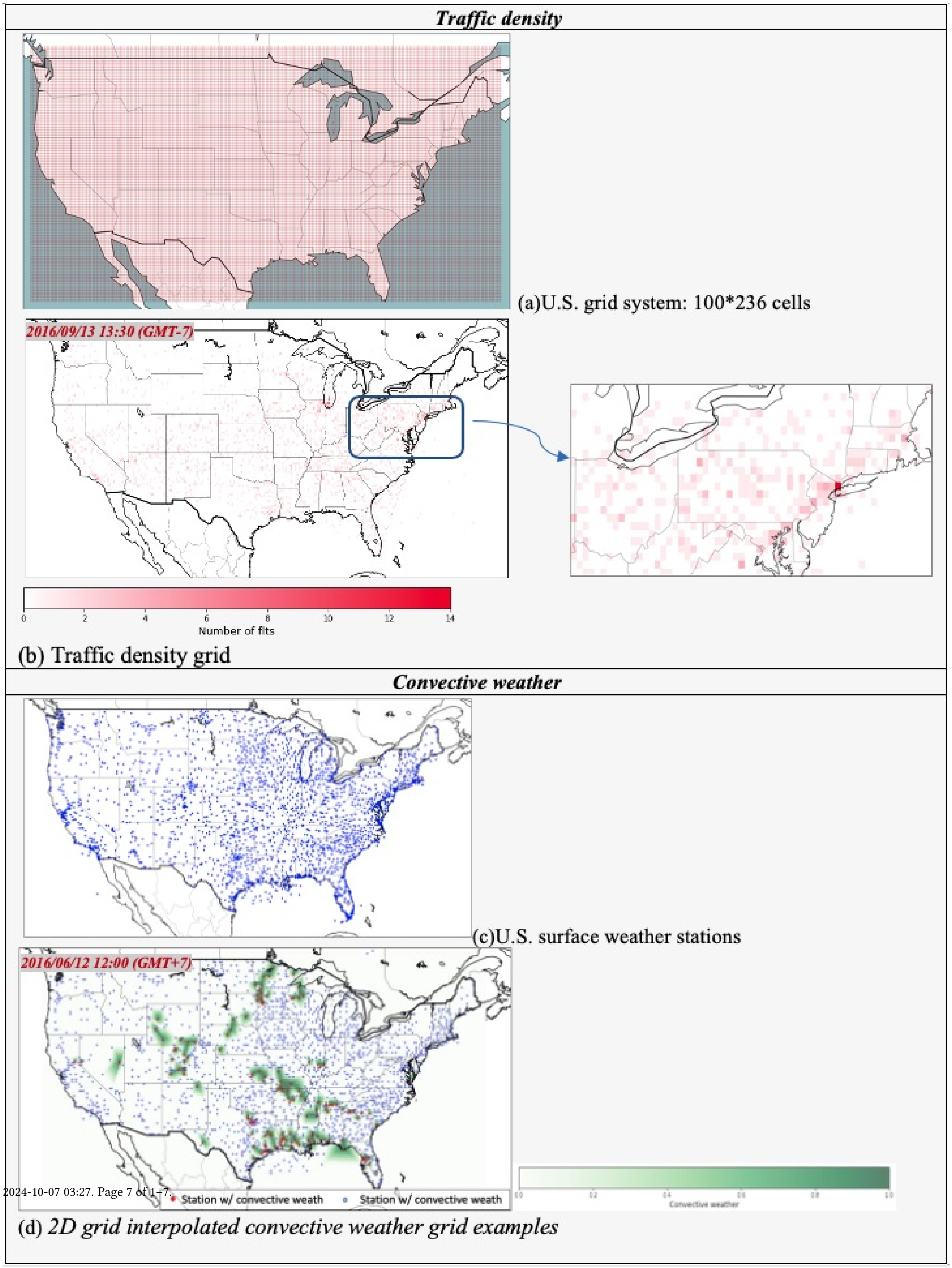
[13] Alice Sternberg, Jorge Soares, Diego Carvalho, and Eduardo. Ogasawara. 2017. A Review on Flight Delay Prediction. https://arxiv.org/abs/1703.06118

[14] Gaofeng Su, Xi Cheng, Siyuan Feng, Ke Liu, Jilin Song, Jianan Chen, Chen Zhu, and Hui Lin. 2024. Flight Path Optimization with Optimal Control Method. *arXiv preprint arXiv:2405.08306* (2024).

[15] Thomas Vandal, Max Livingston, Camen Piho, and Sam Zimmerman. 2018. Prediction and Uncertainty Quantification of Daily Airport Flight Delays. In *Proceedings of The 4th International Conference on Predictive Applications and APIs*. PMLR, 82:45–51.

[16] Liya Wang, Alex Tien, and Jason Chou. 2021. Multi-Airport Delay Prediction with Transformers. https://arxiv.org/abs/1912.09363

Appendix 1



Traffic density

(a)U.S. grid system: 100*236 cells

(b) Traffic density grid

Convective weather

(c)U.S. surface weather stations

(d) *2D grid interpolated convective weather grid examples*