Artificial Intelligence Review (2026) 59:36
https://doi.org/10.1007/s10462-025-11400-w

A review of network delay prediction and advances in large language models for air traffic

Mengyuan Sun[1,2] · **Yong Tian**[1] · **Jiangchen Li**[1,3] · **Cheng-Lung Wu**[2] · **Liqun Peng**[4] · **Shucai Xu**[5]

Received: 2 May 2025 / Accepted: 16 September 2025 / Published online: 16 December 2025
© The Author(s) 2025

Abstract

Traffic network delays seriously affect the air transportation system's safety, economy, and efficiency, and have always been a global concern. Flight delays usually propagate within airport networks, causing subsequent flights to be delayed. However, existing works lack in considering network causality, and the incorporation of emerging large language models (LLMs). Thus, this paper endeavours to examine the literature on network delay prediction that combines different background knowledge with journal paper publishing data. Particularly, the network delay prediction methods are categorized into four aspects: classic methods without explicit network topology modelling, traditional explicit network-based prediction methods, emerging deep learning methods, and the application of LLMs in transportation. Classic methods without explicit network topology modelling, including statistical analysis, operations research, traditional machine learning and causal inference without network structures, offer interpretable baselines but fail to capture the complexity and nonlinearity of air traffic systems. Traditional explicit network-based prediction methods often approach air traffic systems through frameworks such as complex networks and queuing theory, with an increasing focus on causal relationship analysis. However, these methods fall short in capturing the spatiotemporal dependencies of network delays, particularly in modelling spatiotemporal causality. In contrast, emerging deep learning methods have advanced significantly, enabling the construction of spatiotemporal causal networks and improving the accuracy of network delay prediction. In addition, some future trends are analyzed. It is concluded that graph neural networks with causality and emerging deep learning methods (e.g., spatiotemporal GCN) are identified as essential directions. Moreover, a conceptual AirTraffic LLM is suggested via a novel Spatial-Temporal Causal Large Language Model (STC-LLM) framework for high-precision flight delay prediction, which requires further experimental validation and real-world testing. Nevertheless, issues such as data privacy, model opacity, and high computational costs must be carefully addressed when applying LLMs. Finally, the findings are expected to enhance understanding of delay propagation among researchers, practitioners, and policymakers, while providing insights and guidance to airports, airlines, and air traffic control.

Extended author information available on the last page of the article

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Keywords Large language models · Flight delay · Network delay prediction · Propagated delay · Causality · Deep learning

1 Introduction

With economic globalization, the demand for air traffic has surged in recent decades. According to IATA (2025), global air passenger demand is projected to grow 10.4% year-on-year, exceeding pre-pandemic (2019) levels by 3.8%. Also, capacity is expected to increase by 8.7%, driving the load factor to a record 83.5% (IATA 2025). According to the aviation passenger rights service platform AirHelp, in 2018, global flight delays reached a record level, affecting over 10 million passengers, with nearly one-fifth of flights experiencing delays (at least 15 min) or cancellations. This led to airline losses approaching $300 billion.

The term "flight delay" first appeared in the 1929 Warsaw Convention, the first international treaty on air transport. Delays can be described as the difference between a flight's scheduled time and actual time (Wieland 1997). The U.S. Department of Transportation defines flights as on time if their arrival or departure occurs within 15 minutes of the schedule time, with delays exceeding 15 minutes classified as delays for public reporting of punctuality and delay statistics (FAA 1987; U.S. Department of Transportation 2024). EUROCONTROL measures flight delays as the difference between actual and scheduled arrival/departure times, using a more than 15-minute threshold for arrival delays, a common benchmark reflecting passenger experience (EUROCONTROL 2024, 2025). The Civil Aviation Administration of China (CAAC) has the corresponding regulations on flight delays (MTPBC 2016). Figure 1 shows that the CAAC defines regular and irregular flights in its "Civil Aviation Flight Punctuality Statistics Methodology (MTPBC 2016)." Especially, Article 3 of the "Flight Punctuality Management Regulations" by CAAC classifies delays into three categories (MTPBC 2016): (1) "flight delay" refers to a situationwhere the Actual In-Block Time (AIBT) exceeds the Scheduled Arrival Time (SAT) by more than 15 minutes. (2) "departure delay" is defined as a condition in which the Actual Off-Block Time (AOBT) exceeds the Scheduled Departure Time (SDT) by more than 15 minutes. and (3) "onboard delay" refers to a situation in which passengers wait inside the aircraft beyond the SDT, either after cabin doors are closed and before takeoff, or after landing and before doors are opened. These categories are depicted in the flight operation flowchart in Fig. 2. Based on the above analysis, it is indicated that a 15-minute threshold is commonly accepted by institutions. Delay propagation refers to the spread of delays from upstream flights to downstream flights. Flight delays are categorized as initial delays or propagated delays (PDs) based on their sequence of occurrence (Gopalakrishnan and Balakrishnan 2021). Initial delays involve issues with individual flights, such as ground operation issues or equipment malfunctions. PDs occur when shared resources such as aircraft, crew, passengers, and airport facilities connect an initially delayed flight to subsequent downstream flights (Kafle and Zou 2016). As aircraft typically operate multiple daily flights, interdependencies can cause delays to accumulate, creating a snowball effect that disrupts multiple airports or the entire network (Fleurquin et al. 2013). When initial and PDs interact at scale, they cause widespread disruptions known as network delays (Wang et al. 2022a). This study focuses on PDs within the context of these network-level disruptions.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

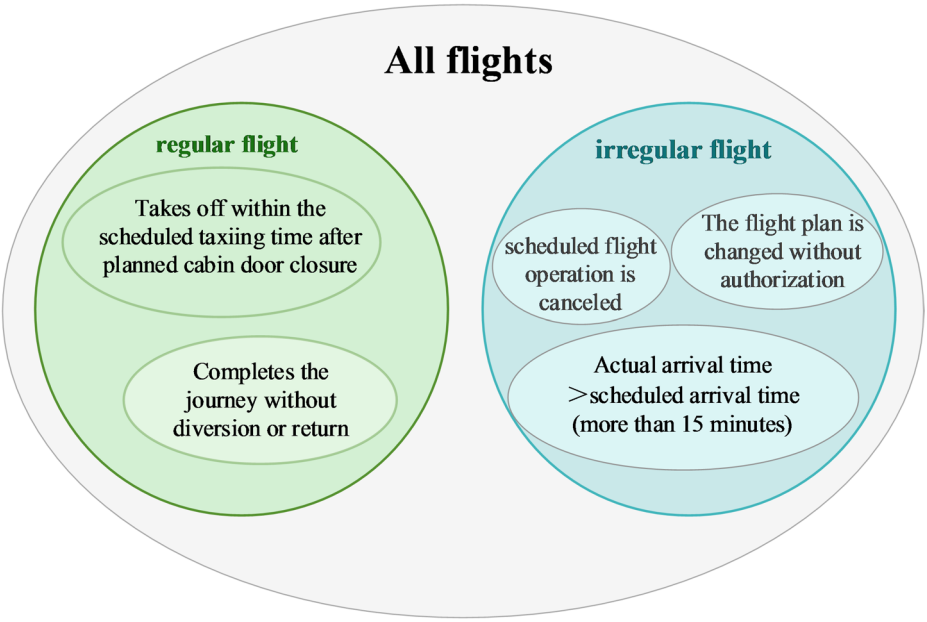

Fig. 1 The definition of CAAC's normal and irregular flights

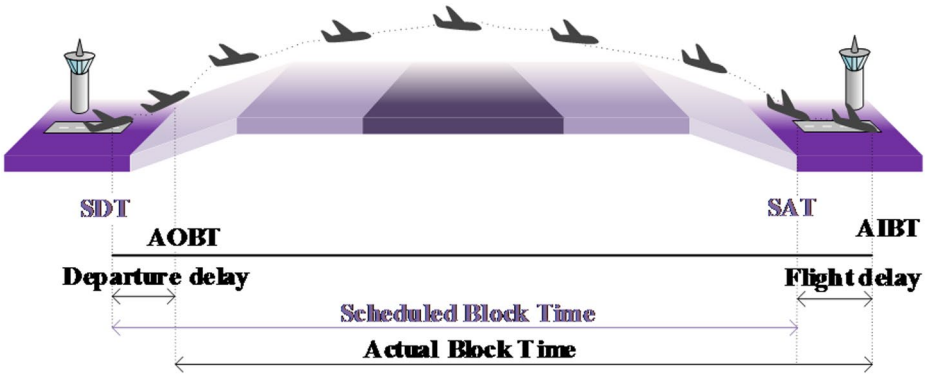

Fig. 2 The flow chart of flight operation

Flight delays exhibit spatiotemporal propagation, particularly when the same aircraft operates consecutive flight missions, leading to PDs in subsequent flights. To better understand the entire flight ecosystem, aviation authorities continuously collect and store vast amounts of operational data. Therefore, in the current context of data explosion, efficiently allocating limited resources to mitigate delays and minimize their economic impact has become a key priority for both researchers and industry practitioners (Wu and Maher 2018).

In contemporary research, scholars utilize diverse methods for conducting literature reviews in traffic management. Zhou et al. (2022b) conducted a comprehensive analysis of vehicle speed prediction research, categorizing it into macro traffic, micro vehicles, and

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

meso lane levels. They summarized evaluation metrics, public datasets, and open-source code. Chen et al. (2022) utilized the Web of Science database and CiteSpace software to analyze research on aircraft approach and landing safety. Their findings, conforming to Rice's logarithmic curve, highlight the increasing significance of this topic in aviation safety research. Bergantino et al. (2024) categorized research on transportation network resilience using real-world data, examining spatial patterns and the impact of resilience attributes. Most of these studies employed traditional literature analysis, bibliometric tools, and conventional classification methods to review ground transportation systems, but offered limited in-depth analysis of air traffic delays.

As research progresses, many scholars have gradually focused their attention on the field of flight delays. Wang et al. (2022a) analyzed the causes and propagation mechanisms of air traffic delays, proposed a data science-based classification of delay prediction methods, and conducted a comprehensive comparative evaluation. Li et al. (2024a) examined delay propagation models with a focus on data, methodology, and future directions, suggesting a comprehensive classification framework and outlining data sources and application scenarios for various data types. Wandelt et al. (2025) reviewed recent advances in flight delay prediction using a six-step comparative framework, evaluating progress in data collection, model selection, feature selection, evaluation methods, technology integration, and repeatability. While existing studies have advanced understanding of flight delays, research on applying large language models to this domain remains limited amid the rise of generative artificial intelligence.

Despite progress in flight delay prediction across civil aviation, transportation, and data science, critical knowledge gaps remain, requiring systematic investigation:

(1) The characteristics of various flight delay prediction methods at various developmental stages are not elaborated, particularly lacking a comprehensive review of airport network delay prediction methodologies from the perspective of propagated delays.
(2) Scholars have primarily focused on the effectiveness of flight delay prediction approaches, overlooking the actual operational context of flights in air traffic networks and the deeper investigation of delay propagation relationships, particularly causality.
(3) Despite the rise of the Internet, mobile devices, and cloud computing, few studies have comprehensively analyzed the application of generative artificial intelligence, particularly platforms like ChatGPT, in the transportation sector.

Therefore, this study approaches the network delay prediction issue from the perspective of propagated delays, proposing a systematic review on air traffic delay prediction methods, providing theoretical support for the safe, efficient, and intelligent operation of the civil aviation industry. The contributions of this article are as follows:

- First, for various delay prediction methods, the characteristics at different development stages are clarified, with a particular focus on research from the perspective of propagated delay prediction, highlighting representative techniques at each developmental phase.
- Second, considering the actual operation of flights within the air traffic network, this study analyzes the characteristics of classic methods without explicit network topology modelling, traditional explicit network-based methods, and emerging deep learning

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

methods in various development situations, with careful focus on network-level, causality, and graph neural networks.
- Third, this study explores the application of generative artificial intelligence in transportation and proposes a conceptual AirTraffic LLM, featuring a novel Spatial-Temporal Causal Large Language Model (STC-LLM) framework for high-precision flight delay prediction.
- Last, this study provides a forward-looking perspective to analyze the current research gaps and the future research directions. It is found that network causality and spatiotemporal graph neural networks are the main directions.

The remainder of this paper is as follows. Section 2 gives the methodologies of network delay prediction methods, covering classic methods without explicit network topology modelling, traditional explicit network-based methods, emerging deep learning methods, and large language model applications in transportation. Section 3 suggests a novel STC-LLM and discusses challenges of network delay prediction in air traffic, while presenting the future research directions.

2 Methodologies for network delay prediction

To acquire an accurate understanding of the operational dynamics within the civil aviation transport industry, literature from the WOS core database was selected as the data source. With 'topic' set as the search condition, the literature was searched using the string TS = (((propagated delay) OR (delay propagation) OR (delay*)) AND ((flight delay) OR (air transport) OR (air transportation) OR (flight) OR (air*)) AND ((machine learning) OR (deep learning))) AND PY = (1990–2025) for an advanced search spanning the years 1990 to 2025. This search yielded a total of 876 records, which were subsequently exported in plain textual format for analysis. The metadata processing followed four key steps: identification, screening, eligibility assessment, and inclusion. Duplicate records were removed during the identification and screening process. In the eligibility phase, studies were excluded based on criteria such as relevance to air traffic, source availability, and non-preprint status. Final inclusion was based on the data's relevance to propagated delay research. As a result, 110 articles were selected for methodological analysis, with an additional 35 used to examine challenges and future directions. To enable a deeper analysis of delay prediction, a statistical summary of relevant data sources was conducted, as presented in Table 1.

Table 1 summarizes flight delay prediction data sources, including data types, availability, and access links. The U.S. Bureau of Transportation Statistics (BTS) provides publicly available flight operation data widely used in delay studies. In contrast, flight operation data from the CAAC, though frequently employed, is not publicly accessible and is typically obtained through close institutional collaboration. Several websites provide access to aircraft trajectory and weather data. Trajectory data is typically limited to the current day or past week, with extended access requiring purchase. In contrast, weather data is generally more accessible and widely available. Understanding the sources and accessibility of these data can provide valuable guidance for future research in flight delay prediction, trajectory forecasting, and flight schedule optimization. Depending on specific research goals, current prediction methods for network delay prediction can be divided into three categories: classic

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 1 Statistics of commonly used data sources related to flight delay prediction

Data source	Type	Public availability	Link
Bureau of Transportation Statistics (BTS)	FO	√	http://www.transtats.bts.gov
Civil Aviation Administration of China (CAAC)	FO	NA	NA
AirNav Radar	TD	√	https://www.airnavradar.com
Flightradar24	TD	Partially	https://ansperformance.eu/acronym/coda/
VariFlight	FO, TD	Request	https://data.variflight.com
FlightAware	TD	Partially	https://www.flightaware.com/zh-CN/
Opensky	TD	Request	https://opensky-network.org/
Aviation System Performance Metrics (ASPM)	FO	Partially	https://aspm.faa.gov/apm/sys/AnalysisAP.asp
National Oceanic and Atmospheric Administration (NOAA)	WD	√	https://www.ncdc.noaa.gov/cdo-web/datatools/lcd
National Weather Service	WD	√	https://www.weather.gov/
OurAirports	Airports	√	https://ourairports.com/data/
Weather Underground	WD	√	https://www.wunderground.com/history
Aviation Weather Center	WD	√	https://aviationweather.gov/
OGIMET	RD	√	https://www.ogimet.com/index.phtml.en
Aircraft Type Designators	EPD	√	https://www.icao.int/publications/doc8643/pages/search.aspx

FO data: Flight operation data; ADS-B: Automatic Dependent Surveillance-Broadcast; TD: Trajectory data; WD: Weather data; Airports related information: airports, airport-frequencies, airport-comments, runways, navaids, countries, regions; Reports data (RD): METAR, SPECI and TAF reports; EPD: Engine performance data

methods without explicit network topology modelling, traditional explicit network-based prediction methods, and emerging deep learning methods (see Table 2).

2.1 Classic methods without explicit network topology modelling

Classic methods for network delay prediction without explicit network topology modelling include statistical analysis, operations research, traditional machine learning, and causal inference without network structures. Statistical analysis mainly involves techniques such as descriptive statistics, regression models, correlation analysis, parametric and non-parametric tests, econometric models, and multivariate analysis. Regarding regression models, both delay multipliers and recursive models help airlines understand the network delays and estimate the delay costs (Wang et al. 2003; Markovic et al. 2008). Some econometric models (Xiong and Hansen 2013) have also been used to evaluate flight operating efficiency. Furthermore, several studies have concentrated on statistical inference. Pathomsiri et al. (2008) employed non-parametric methods to evaluate delay efficiency indicators at U.S. airports. Xiong and Hansen (2013) developed an econometric model incorporating prior delays, destination airport and airline characteristics, aircraft size, and ticket prices to analyze the factors contributing to airline flight cancellations.

Operational research involves advanced analysis methods, including optimization, simulation, and queuing theory (Wu and Caves 2002). Specifically, analyses of airport capacity

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 2 Literature review of typical articles

Methods		Literature	Features of propagated delay prediction		Graph Neural Network		Other features
			Causality	General network features	Graph feature	Time feature	
Classic methods without explicit network topology modelling		Wieland (1997)	X	√, Queue model	X	X	-
		Wang et al. (2003)	X	√, Regression model	X	X	Delay multiplier
		Wu et al. (2012)	X	√, Aircraft routing	X	X	Combined with crew pairing
		Xiong and Hansen (2013)	X	√, Econometric model	X	X	Flight Cancellation Reasons
		Qin and Yu (2014)	X	√, k-means	X	X	Periodicity of flights
		Choi et al. (2016)	X	√, Decision trees	X	X	Weather, random trees, AdaBoost
Traditional explicit network-based prediction methods	Early network development stage	Ahmadi-Beygi et al. (2008)	X	√, Delay tree	X	X	Propagated delay
		Wu (2016)	X	√, Bayesian network	X	X	Markov chain model
		Baspinar and Koyuncu (2016)	X	√, Epidemic model	X	X	Based on flight/airport
		Kim et al. (2016)	X	X	X, RNN	X, LSTM	-
	Causality modelling in network delays	Zanin et al. (2017)	√, GCT	√, Complex network	X	X	-
		Zhang et al. (2019)	√, GCT, TE	√, Neural network	X	X	Propagation index
		Xiao et al. (2020)	√, TE	√, Complex network	X	X	High-dimensional information
		Gui et al. (2020)	X	X	X	X	LSTM, RF
		Zanin (2021)	√, GCT	√, Functional networks	X	X	Granger causal clustering

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 2 (continued)

Methods	Literature	Features of propagated delay prediction		Graph Neural Network		Other features
		Causality	General network features	Graph feature	Time feature	
Emerging deep learning methods	Guo et al. (2022)	√, CCM	√, Complex network	X	X	-
	Zeng et al. (2022)	√, PCMCI	X	X	X	Dynamic causality
	Jia et al. (2022)	√, Nonlinear GCT	√, Complex network	X	X	Low dimensionality of TE
	Wang et al. (2022b)	√, Causality	X	X	X, LSTM, Attention	-
	Cai et al. (2023)	X	X	√, TA-aDGCN	X, but Attention	Air traffic flow forecast
	Du et al. (2023)	X	X	√, GNN, Attention	X, Skip-LSTM	Hybrid DL model
	Yang et al. (2023)	√, Causality	√, Bayesian network	√, GCN-Graph-SAGE	√, GNN	Ground equipment service
	Li et al. (2024c)	√, TE	√, Full-connected network	X	X, causal biased random walk	Propagated delays in airports
	Bala Bisandu and Moulitsas (2024)	X	√, Deep operating network	X	X, deep learning model	Flight delays
	Sun et al. (2024)	√, TE	√, Recurrent neural network	√, GCN	√, GRU	Propagated delay prediction

often consider variances in departure and arrival delays under diverse meteorological conditions (Schaefer and Millner 2001; Hunter et al. 2007). In addition, some scholars used the queue model to develop delay propagation research. Wieland (1997) employed a queuing model to predict root-cause delays. Hansen (2002) developed a streamlined deterministic queuing model to question how delays spread to subsequent flights. Dunbar et al. (2012) introduced a method to accurately calculate and minimize delay propagation costs within the aircraft routing and crew pairing framework. There are also studies analyzing delays in scheduling stability through simulation (Dück et al. 2012) and reliability aspects (Wu 2005). Gopalakrishna and Balakrishnan (2021) developed a dynamic air traffic delay network model using operational data, applying Markov jump linear systems to characterize network behaviour.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Traditional machine learning algorithms mainly include logistic regression, hidden Markov models, perceptron, support vector machine (SVM), decision tree, gradient boosting decision tree (GBDT), extreme gradient boosting (XGBoost), k-means clustering, random forest (RF), and long short-term memory (LSTM) models (Woodburn and Ryerson 2014; Rebollo and Balakrishnan 2014; Choi et al. 2016).

Causal inference without network structure is primarily used to model causal relationships between variables. In 1969, Granger introduced the Granger Causality Test (GCT) using a bivariate model to define and formalize the concept of causality (Granger 1969). The classical GCT has been increasingly applied across diverse fields, including physiology (Kugiumtzis et al. 2017), ecology (Sugihara et al. 2012), sociology (Frank et al. 2018), physics (Ringbauer et al. 2016), and finance (Papana et al. 2017). In the context of air traffic, such methods enable the representation of delay propagation as causality between airport delay time series.

Particularly, LSTM models have proven effective for flight delay prediction due to their ability to capture long-term dependencies, adapt to data variability, maintain robustness against anomalies, and support end-to-end learning. Gui et al. (2020) compared LSTM and Random Forest models for flight delay prediction. They tested three LSTM architectures: a baseline model, a fully connected variant, and a dropout-enhanced version. Experimental results demonstrated strong time-delay correlations and showed that the dropout layer effectively mitigated overfitting (Fig. 3). Li et al. (2023) established a predictive model combining CNN for spatial analysis, LSTM for temporal dynamics, and RF for integrating these features with external data. Mamdouh et al. (2024) presented an Attention-based Bidirectional LSTM (ATT-BI-LSTM) ensemble model for flight delay prediction, utilizing attention-enhanced bidirectional LSTM to capture the spatial and temporal features of flight and weather data, with its effectiveness validated through experimental results.

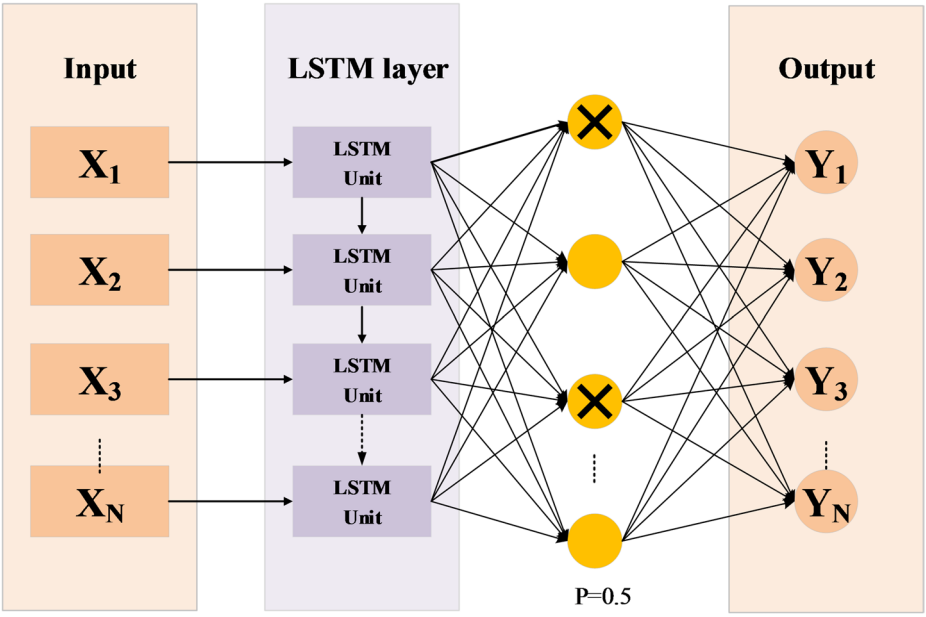

Fig. 3 LSTM with dropout layer (redrew) (Gui et al. 2020)

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Classic statistical models, while interpretable and capable of tracing delay propagation dynamics, fail to fully capture the nonlinear and stochastic nature of airport networks. For instance, statistical models are widely employed in delay propagation analysis, enabling researchers to assess multifactorial impacts on flight delays across entire air traffic networks. In addition, although econometric approaches are broader and effective in identifying network interrelationships for mitigation, these methods struggle to capture complex nonlinear dynamics, limiting their applicability in sophisticated scenarios. Next, for optimization and simulation methods, while adaptable for modelling diverse delay scenarios through specialized software, they require extensive flight parameters, operational rules, and hypothetical conditions (Hao et al. 2014). In general, these approaches demand significant domain expertise and technical proficiency. Although valuable for macroscopic delay analysis and mitigation strategy development, their high implementation complexity presents notable limitations. What's more, traditional machine learning methods have been applied to network delay prediction research through dataset analysis, demonstrating strong capabilities in identifying nonlinear patterns and predicting future network delays. While effective for handling large-scale data, these methods require substantial domain expertise for manual feature engineering and significant computational resources, resulting in high implementation costs. Amid the rapid growth of air traffic, understanding and applying inter-airport relationships in network delay prediction has become increasingly essential. Current research predominantly focuses on initial delays and traffic flow variations, with limited attention to PDs and explicit network causality in complex air traffic networks. Future work should prioritize network delay prediction combined with causality in realistic air traffic system environments, though this will demand extensive experimental validation to ensure accuracy.

2.2 Traditional explicit network-based prediction methods

Traditional explicit network-based prediction methods are primarily grounded in causal relationships between airports. Thus, considering the operational nuances of airport networks, network delay prediction methods are classified into early network development stage and causality modelling in network delays, respectively.

2.2.1 Early network development stage

In the early network development stage, researchers built large and complex network models to understand the characteristics of air traffic systems. AhmadBeygi et al. (2008) investigated delay propagation patterns in both hub-and-spoke and point-to-point networks through delay tree modelling. Khanmohammadi et al. (2016) introduced an innovative multi-level input layer neural network approach for predicting flight delays. In the same year, Baspinar and Koyuncu (2016) presented a novel delay propagation method inspired by the process of epidemic spread, using this concept to construct two distinct data-driven epidemic models. Wu et al. (2018) developed an enhanced delay propagation tree model with a Bayesian network (DPT-BN) to analyze multi-flight delay propagation and interdependencies. Sun et al. (2020) used complex network theory to study the dynamic spatial-temporal evolution of the new coronavirus in the air transportation network. Li and Jing (2021) established a

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

delay propagation network based on the Bayesian network and examined the impact of PDs between different types of airports.

Early network-based research primarily employed Bayesian networks, queuing theory, and complex network theory to model air traffic systems, yet neglected dynamic delay propagation processes, network-level delays, and spatiotemporal dependencies. In practice, delays rapidly spread between airports and airlines due to aircraft movement, creating a "snowball effect" that accelerates delay propagation across the airport network. Therefore, understanding these propagation mechanisms is critical for enhancing prediction accuracy and supporting operational decision-making in aviation.

2.2.2 Causality modelling in network delays

Given the high complexity of the current air transportation system, flight delay time series exhibit nonlinear characteristics. However, GCT demonstrates limited effectiveness when analyzing systems with numerous factors exhibiting nonlinear relationships. The transfer entropy (TE) method, originating from Shannon's (1948) information entropy theory, overcomes the linear methods' inability to analyze nonlinear signal characteristics. It quantifies system complexity through its derived formula. The definition of entropy *H (X)* is as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{1}$$

In Eq. (1), the sequence X comprises n states x_i (i = 1, 2, …, n), each with a probability $p(x)$. According to the principle of TE in information theory, incorporating the historical information of both source and target signals improves the accuracy of predicting future states compared to using only their historical information, thereby indicating a causal influence from the source to the target. The entire development trend of TE is shown in Fig. 4.

In 2000 year, Schreiber (2000) first proposed the TE method, which is defined by the following formula:

$$TE_{Y\to X} = \sum_{x_{i+1},x_i,y_j} p(x_{i+1}, x_i^{(q)}, y_j^{(t)})\log_2 \frac{p(x_{i+1}x_i^{(q)}, y_j^{(t)})}{p(x_{i+1}\,|\,x_i^{(q)})} \tag{2}$$

Equation (2) calculates the TE from the variable Y to variable X. Here, $p(A\,|\,B)$ denotes conditional probability, x_i is the value of X at time i, y_j is the value of Y at time j, and x_{i+1} is the value of X at the next time point. The dimensions q and t of x and y are given by $x_i^{(q)} = [x_i, x_{i-1}, ..., x_{i-q+1}]$ and $y_j^{(t)} = [y_j, y_{j-1}, ..., y_{j-t+1}]$ respectively.

Transfer entropy from Y to X measures the reduction in the uncertainty of X due to information from Y, quantifying how much information Y conveys to X. Thus, it serves as an effective indicator for assessing causality.

Building on Schreiber's work, Bauer (2007) enhanced TE method by considering time delays in information transfer between variables. To better describe the system's information

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

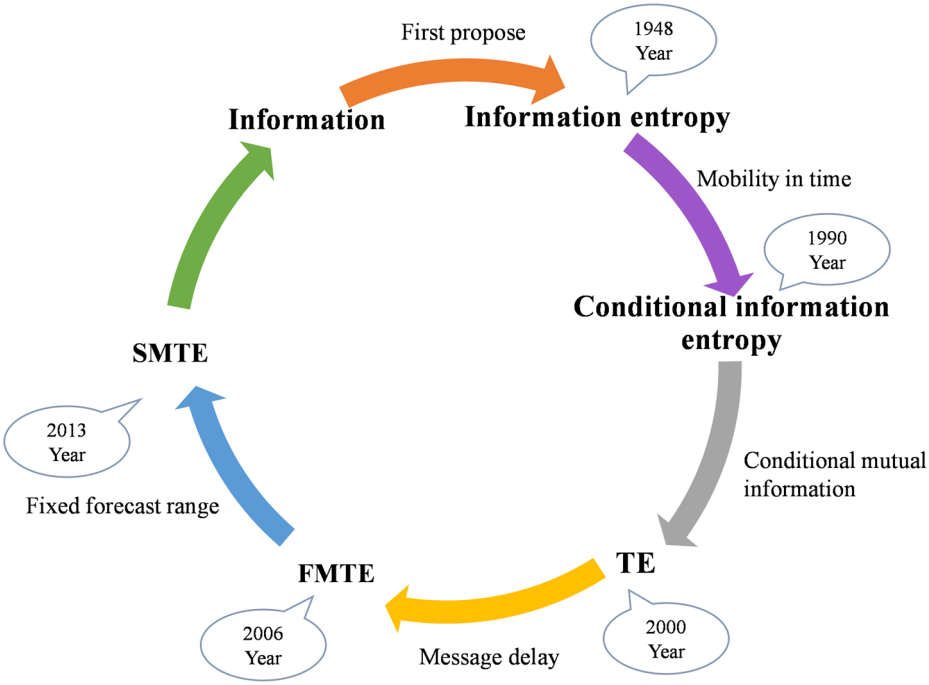

Fig. 4 The development of the TE method

interaction, he proposed the First Modified Transfer Entropy (FMTE) method, expressed as follows:

$$TE_{Y\to X} = \sum_{x_{i+s},x_i^{(q)},y_j^{(t)}} p(x_{i+s},x_i^{(q)},y_j^{(t)})\log_2 \frac{p(x_{i+s}\left|x_i^{(q)},y_j^{(t)}\right.)}{p(x_{i+s}\left|x_i^{(q)}\right.)} \tag{3}$$

In this context, the TE method assesses the influence of y_j on the information entropy of x_{i+s}, representing the information flow from y to the s-th time step of x. Here, s denotes the time delay in the information transmission from Y to X, known as the forecasting horizon. Adjusting s enables transfer entropy to accommodate variable time delays between system variables, enhancing the formula's realism. Additionally, adjusting s helps identify the maximum transfer entropy value, enhancing the significance of the calculation results.

However, Bauer's TE method requires adjusting the forecasting horizon s, causing the interval between x_i and x_{i+s} to vary with s. This inconsistency leads to variations in the referenced information entropy, rendering the process unreasonable. To address this issue, Shu and Zhao (2013) offered the Second Modified Transfer Entropy (SMTE) method.

$$TE_{Y\to X} = \sum_{x_{i+s},x_{i+s-1}^{(q)},y_j^{(t)}} p(x_{i+s},x_{i+s-1}^{(q)},y_j^{(t)})\log_2 \frac{p(x_{i+s}\left|x_{i+s-1}^{(q)},y_j^{(t)}\right.)}{p(x_{i+s}x_{i+s-1}^{(q)})} \tag{4}$$

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Equation (4) shows that by substituting x_i with x_{i+s}, the interval between x_{i+s-1} and x_{i+s} remains constant regardless of changes in s. This ensures algorithm stability by maintaining a consistent reference for transfer entropy values.

Recent advances in causal reasoning methods (Cramer et al. 2017; Feng et al. 2024) provide data-driven approaches to identify delay propagation mechanisms between airports. These techniques quantify delay causality by modelling the causal relationship between airport delay time series, revealing the underlying propagation pathways. Que et al. (2018) developed a chain aviation network model to analyze initial delay impacts. Their simulation revealed positive correlations between initial delays and three network-level metrics: number of affected airports, total delay duration, and average delay time. Zhang et al. (2019) introduced a propagation index to quantify delay impacts through causal time-series analysis. This method effectively characterized airport interdependence and flight delay relationships. Tang et al. (2021)systematically reviewed recent advances in complex network theory applications for flight delay analysis, with particular focus on: causality identification, propagation modelling, and optimal spreader detection in network delays. Zhou et al. (2022a) developed an inter-airport delay propagation network using causal analysis, which effectively characterized the dynamic propagation process and elucidated the delay propagation mechanism from a global perspective. Bombelli and Sallan (2023) constructed airline- and event-specific delay propagation networks, employing Granger causality analysis to identify causal patterns in hourly airport delays. These studies shared a common focus: elucidating intricate unit-time interactions in dynamic systems. By employing time-series analysis and causality testing, they identify interaction patterns to uncover latent mechanisms governing complex systems.

The main trends in causality research are shown in Table 3. The methodologies encompass GCT, causal network learning algorithm (CNLA) (Spirtes and Glymour 1991), structural causal models (SCMs) (Chickering 2013; Elwert 2014), transfer entropy (TE), and nonlinear state space (NLSS) (Sugihara et al. 2012; Guo et al. 2022).

Figure 5 summarizes several causal analysis methods. Specifically, Fig. 5(a) displays the multivariate GCT method, which assesses whether excluding the past of these time series X (black dashed box) from a model incorporating Y's history and other covariates (purple solid box) increases the prediction error of Y at time t (black node), thereby revealing the time-lagged causal relationships. Figure 5(b) introduces the convergent cross-mapping of the nonlinear state-space (NLSS) method, which uses the chaotic Lorenz system as an example and reconstructs the state space of variables through time-delay coordinate embedding (M_X, M_Y). If the points on M_X can be predicted by the nearest points on M_Y (orange ellipse), and the denser the points on the attractor, the better the prediction quality, hence the conclusion is drawn. Figure 5(c) is the TE method, which is derived from information entropy. When the TE of X to Y is greater than the TE of Y to X, X can be called the cause and Y is an effect, and a causality between the two variables can be established. Specially, the small triangle in the middle represents the transfer entropy ($T(X_t > Y_t, t)$), while the other three circles show the information entropy of X and Y at time $t-1$ and the information entropy of Y at time t. Fig 5(d) shows the SCM, which exploits the asymmetry between cause and effect (the mechanism independence principle) to detect the direction of cause and effect in Markov equivalence classes. In addition, the LINGAM method shown in Fig. 5(d) (Hoyer et al. 2008) (assuming a linear model with Gaussian noise), can identify causality because the model residuals in that direction (black fitted line) are independent of Y (top subplot),

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 3 Major trends in causality

Literature	Characteristics of network delays				
	Network causality				Other features of network delays
	Development stage	Methodology	Causality	Classification	Other features
Granger (1969)	Original model stage	Granger sequence	Yes	GCT	Two-variable model
Spirtes and Glymour (1991)		CNLA	Yes	PC algorithm	Conditional verification
Schreiber (2000)		TE series	Yes	TE	TE was proposed
Allan et al. (2001)	Original flight delay model	Granger sequence	Yes	GCT	Correlation between weather and delay
Neuberg (2003)	Basic stage	SCM	Yes	SCM	Bayesian causal network, graphical effect
Bauer et al. (2007)		TE series	Yes	TE	FMTE
Yang et al. (2010)			Yes	TE	SMTE
Sugihara et al. (2012)		NLSS	Yes	CCM	Causal network
Chickering (2013)		CNLA	Yes	Greedy search	Define a fractional function
Zanin et al. (2017)	Intermediate stage	Granger sequence	Yes	GCT	Complex network
Zhang et al. (2019)				GCT, TE	Propagation index, regression model
Xiao et al. (2020)		TE series	Yes	TE	Complex network, multi-airport
Pastorino et al. (2021)		Granger sequence	Yes	GCT	Network neuroscience
Zanin (2021)				GCT	Granger causal clustering
Guo et al. (2022)	Rapid development stage	NLSS	Yes	CCM	Two-stage analysis framework
Bombelli and Sallan (2023)		Granger sequence	Yes	GCT	Extreme weather, delayed networks
Sun et al. (2024)		TE series	Yes	TE	SMTE, GCN, GRU
Li et al. (2024a, 2024b, 2024c)		Other	Yes	Causality	Score-based temporal causal, LLM
Çelik et al. (2025)		Extended GCT	Yes	Bootstrap Fourier	Air transport-economic growth causality

Springer
Content courtesy of Springer Nature, terms of use apply. Rights reserved.

whereas (yellow line), this is not the case. Figure 5(e) is the CNLA, which demonstrates superior performance in handling high-dimensional data and identifying simultaneous relationship directions. Using the Box 1 model (PC algorithm for time series (Spirtes et al. 2001)) as a case study, the process begins with a graph connecting all variable pairs with unconditional ($p=0$) correlations (assuming stationarity). Subsequently, the PC method performs iterative conditional independence tests as the number of conditioning variables p increases. Unlike GCT, it avoids conditioning on the entire past, thereby reducing estimation dimensionality.

Generally, scholars have explored delay propagation mechanisms and PDs using causality analysis, offering new insights into network delay prediction. While early studies relied on the GCT, alternative causality methods have since emerged. As shown in Table 3, causality research originated earlier at the theoretical model stage, but its application to air traffic delays only began in the early 21 st century, remaining at a foundational level for some time. Few studies initially focused on flight delays, but research gradually advanced to an intermediate stage, with increasing attention to air traffic systems. In recent years, causality analysis has developed rapidly, and future research on PDs is expected to integrate causality more extensively. Given the multi-layered interconnectivity of air traffic, spatiotemporal causality may also become a key focus. This approach would establish a theoretical foundation for a comprehensive understanding of delay propagation mechanisms and improve network delay prediction accuracy.

2.3 Emerging deep learning methods

2.3.1 Neural network development stage

After reviewing classic methods, we now examine how network-based approaches evolved through distinct developmental phases. Past studies did not focus on the spatial correlation of network delay prediction. The airport network is a special graph structure with a small cosmopolitan nature and power-law, and the delay state of each airport has a strong spatiotemporal correlation (Guida and Maria 2007).

Recent research shows that advances in deep learning techniques help address traditional transportation problems. Compared with classic methods and traditional network-based prediction methods, deep learning can capture spatiotemporal dependencies through distributed and hierarchical feature representation. Therefore, researchers have applied deep learning technology to network delay prediction. To capture the spatial correlation in airport networks, scholars have conducted a series of studies.

Regarding delay prediction methods, most studies on spatiality used convolutional neural networks (CNN) and recurrent neural networks (RNN) (Pouyanfar et al. 2018). However, both have certain limitations. CNN is essentially suitable for Euclidean space, such as images, regular grids, etc., while the flight delay data on the entire network is a continuous time series distributed topologically, which is a typical non-Euclidean structured data (Kipf and Welling 2017). In addition, CNN has limitations for traffic networks with complex topology, so it cannot essentially characterize spatial correlation. The most widely used neural network model for processing sequence data is RNN, especially the LSTM model. Nevertheless, RNN models have limitations in long-term prediction due to defects such as vanishing gradient and exploding gradient.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

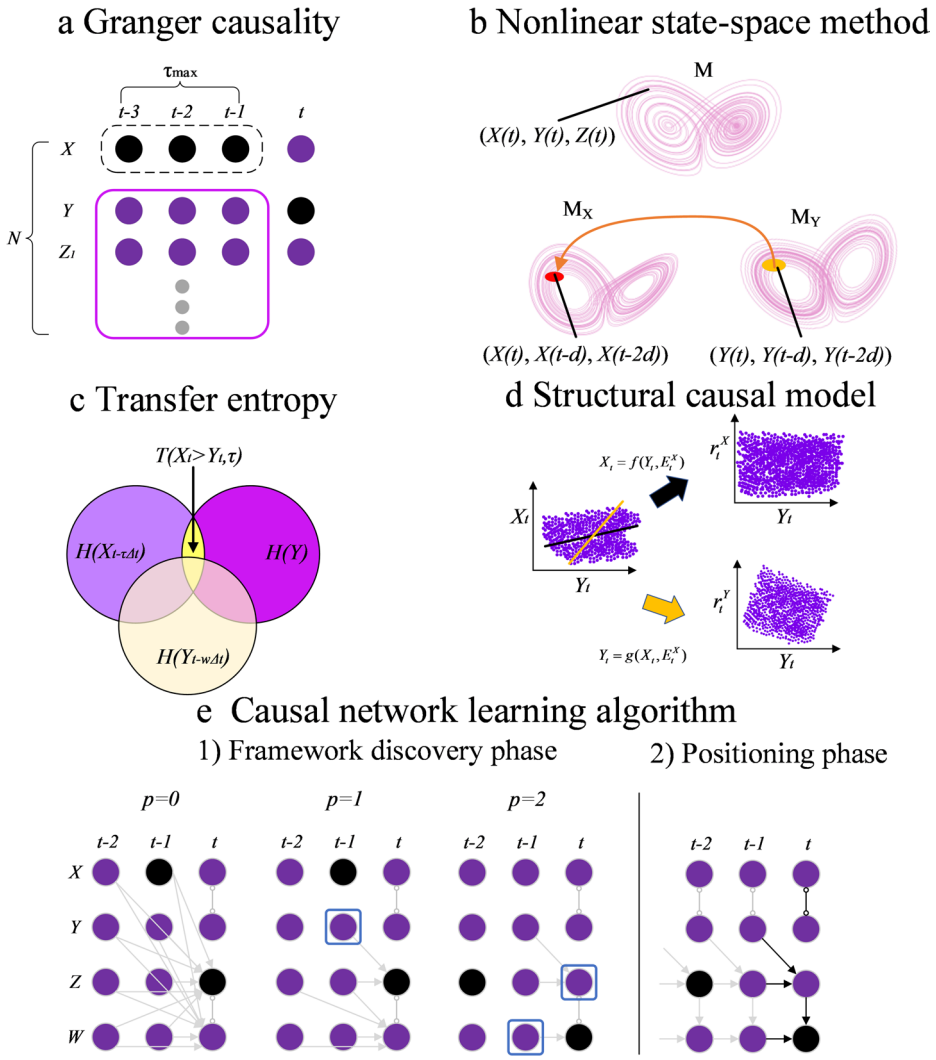

Fig. 5 Diagram of the development stages of the causal analysis theory (redrew) (Runge et al. 2019)

2.3.2 Development stage of emerging graph networks

Graph Neural Networks (GNNs) are nonlinear representation learning methods that leverage underlying graph structures to improve performance. Sperduti and Starita (1997) first applied neural networks to directed acyclic graphs, laying the groundwork for early GNNs research. Designed for end-to-end learning on graph-structured data, GNNs extract high-level representations explicitly (Wu et al. 2021). Their inherently localized and distributed nature makes them particularly well-suited for physical network applications. Zhang et al. (2023a) harnessed the flight conflict network to generate an air traffic situation map, employing GNNs and layered graph models to enhance the accuracy of air traffic complex-

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

ity assessment. Graph Convolutional Network (GCN) is increasingly popular as it pioneers the application of convolution operations from image processing to graph-structured data, providing detailed derivations. This involves complex spectral graph theory, as referenced in the work by Kipf and Welling (2017). For spatial correlation, GCN has shown effectiveness in embedding irregular data (Kipf and Welling 2017; Chen et al. 2021). Cai et al. (2022) offered a flight delay prediction method based on GCN from the perspective of a multi-airport network. Li and Jing (2022) proposed a novel spatial and temporal-random forest prediction framework for flight delay prediction. Tan et al. (2022) introduced a method for discovering the causality of CNNs, analyzing the causality of PDs of airport networks. Wang et al. (2022) constructed a new GCN-based deep learning model that introduced two GCNs to capture local and global spatial correlations between airports. Cai et al. (2023) considered a temporal attention-aware dual-graph convolution network (TAaDGCN) to predict air traffic flow across airspace sectors, which can capture correlations between the spatial dimension and the temporal dimension. Sun et al. (2024) built a GCN framework driven by traffic causal knowledge (i.e., SMTE) to address spatiotemporal dependencies and causality in network delays. This method's application principle in network delay prediction is illustrated in Fig. 6.

Graph Attention Networks (GATs) address the limitation of uniform neighbor weighting in GNNs by incorporating self-attention mechanisms. Unlike GCNs, GATs dynamically compute attention weights for neighboring nodes (Fig. 7), enabling adaptive feature aggregation and reduced structural dependency while emphasizing relevant node features.

Consequently, attention mechanisms show promise in flight delay prediction by effectively modelling complex flight relationships and leveraging multi-head attention mechanisms to capture various influencing factors, thus enhancing prediction accuracy and robustness. Vaswani et al. (2017) introduced the Transformer, the first sequence-to-sequence model based on self-attention, replacing recurrent layers in the encoder-decoder architecture with a multi-head attention mechanism. As is shown in Fig. 8, Guo et al. (2020) employed a spatiotemporal graph dual attention network (STGDAN) to quantify aircraft departure delay times. By representing airspace as a graph and employing dual attention mechanisms, they

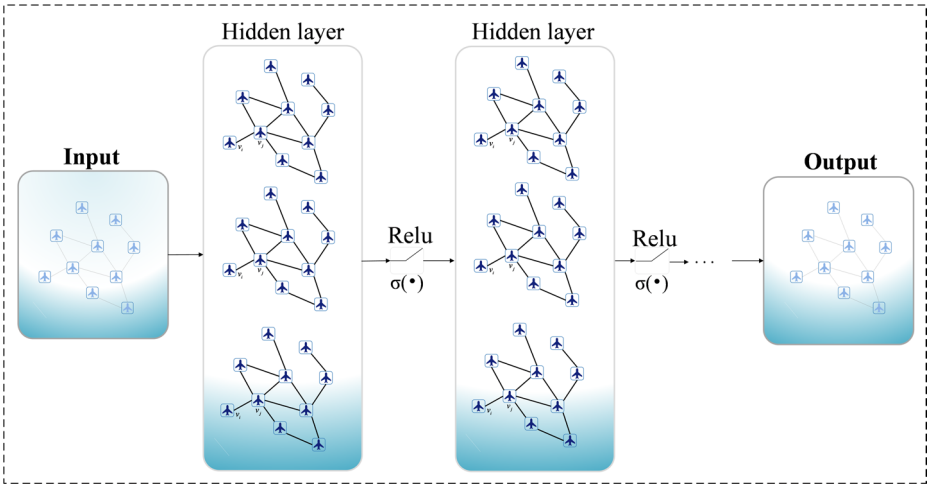

Fig. 6 Application of GCN in delay prediction (redrew) (Sun et al. 2024)

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

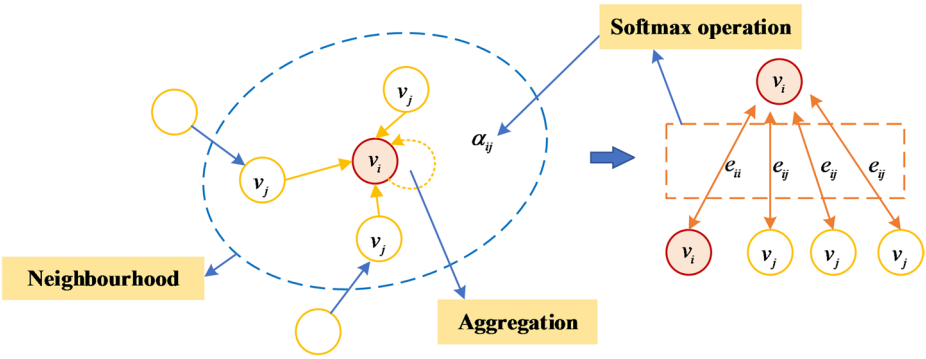

Fig. 7 Application of GAT in delay prediction (redrew) (Qi et al. 2023)

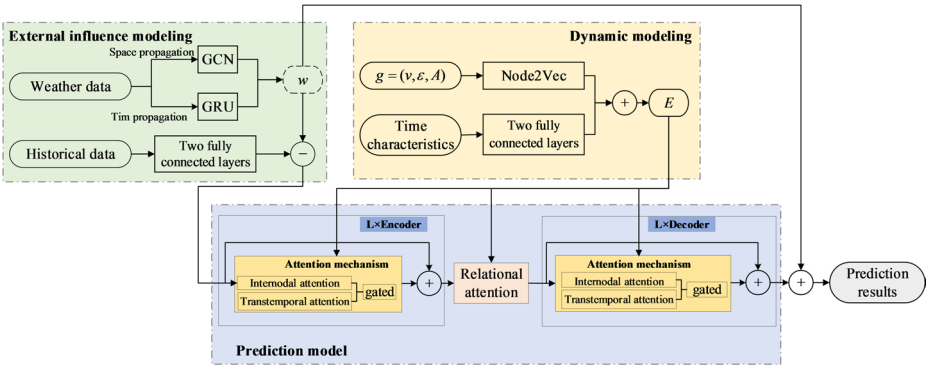

Fig. 8 SGDAN schematic diagram (Guo et al. 2020)

effectively modelled flight interactions, demonstrating high accuracy in predicting delays under 30 min. Du et al. (2023) built a novel hybrid deep learning model, combining CNNs for short-term temporal characteristics and Skip-LSTM for long-term temporal features. Xu et al. (2023) suggested a knowledge-based deep learning framework, Bayesian Ensemble Graph Attention Network (BEGAN), to model complex spatiotemporal variations in air traffic networks. Large-scale air traffic density prediction is achieved by dividing the airspace into grids, independently training models for each, and combining their posterior distributions. Xu et al. (2024) constructed the Physics-Informed Graph Attention Transformer (PIGAT), a deep learning framework combining graph attention-based spatiotemporal modules with temporal Transformers to capture dynamic dependencies in airport networks. Additionally, partial differential equations from fluid queuing theory are embedded in the loss function to improve prediction accuracy. Zheng et al. (2024) developed a spatiotemporal gated multi-head attention network to forecast network-wide flight delays over the next 24 h. Fofanah et al. (2025) proposed CHAMFormer, a model integrating Transformer, GCN, and GNN through a three-stage coupling mechanism to capture short-, mid-, and long-term traffic patterns for improved prediction.

Generally, research on emerging graph networks has become a hotspot, particularly in the air traffic system. The spatiotemporal hotspot map (Fig. 9) reveals current research trends in

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

network delay prediction. Keywords with a minimum frequency of five are displayed, with node time zones indicating their temporal emergence and usage patterns.

As shown in Fig. 9, node size reflects keyword frequency within the selected time window; larger nodes indicate higher prominence in the research. From 2017 to 2021, "machine learning," "artificial neural networks," and "flight delay prediction" emerged as core research hotspots, highlighting the growing dominance of deep learning methods in delay prediction during this period. Node color represents the time of keyword emergence: blue (1999–2001) denotes the earliest keywords, purple to orange-red (2002–2016) indicates mid-phase activity, and orange to yellow (2017–2025) marks the most recent developments. In addition, links represent keyword co-occurrence or citation paths, and lines of the same color indicate keyword co-occurrence within the same time period. Clusters represent groups of related keywords automatically classified into the same category. Particularly, keywords such as *traffic flow management*, *air traffic management*, and *airspace capacity* were concentrated between 2011 and 2016, aligning with Sect. 2.1 and representing the classic methods stage. From 2017 to 2021, terms like *flight delay prediction*, *optimization*, and *artificial neural networks* emerged, corresponding to Sect. 2.2 and the traditional network-based prediction methods stage. Since 2021, keywords such as *deep learning*, *reinforcement learning*, *networks*, *unmanned aerial vehicles*, *airport operations*, *and safety* have become prominent, aligning with Sect. 2.3 and the emerging deep learning methods stage. Notably, *prediction*, *deep reinforcement learning*, and *unmanned aerial vehicles* are expected to be key trends in future research.

Typically, in the classic methods stage, most research analyzes delay propagation mechanisms using statistical methods, operations research, traditional machine learning, and causal inference without network structure to examine relationships between initial and propagated delays (Kafle and Zou 2016). These approaches have low computational complexity and emphasize integrating theoretical models with practical applications for delay

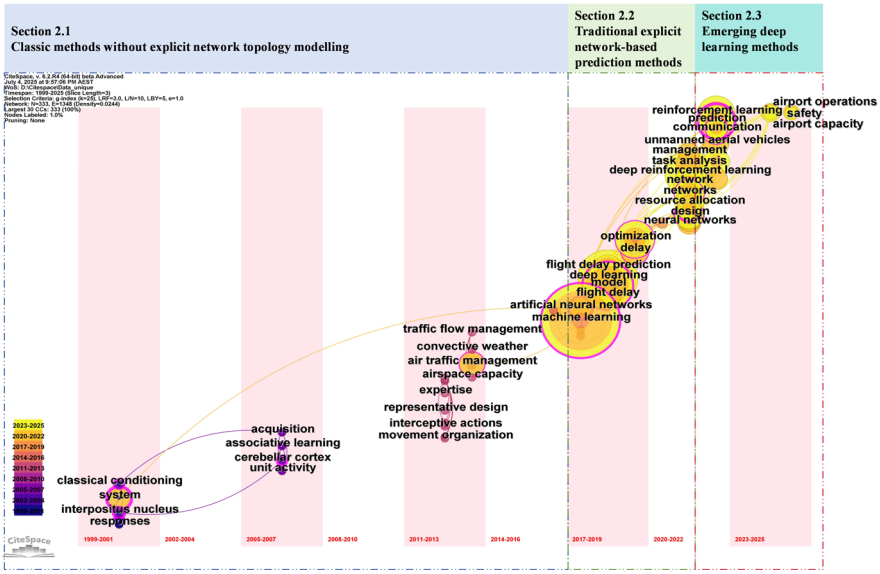

Fig. 9 The spatiotemporal hotspot map of network delay prediction [generated from Citespace]

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

identification and classification. Accordingly, network-based prediction and emerging deep learning methods focus on applying advanced techniques for delay forecasting. A systematic summary and comparison of these approaches is presented in Table 4.

As shown in Table 4, most flight delay prediction studies rely on flight operation data, with many utilizing the publicly available U.S. Bureau of Transportation Statistics (BTS) database. This dataset includes flight date, airline, aircraft type, origin and destination airports, scheduled and actual departure times, departure delays, wheel-off/on times, and schedule/actual flight durations. The CAAC provides similar operational data, though it is not publicly accessible and may differ in format. Most studies utilize datasets spanning over one year, with prediction scopes ranging from hourly to weekly forecasts, including binary delay classification. Common evaluation metrics include RMSE, MSE, and accuracy, with performance gains often highlighted over baseline models.

In addition to temporal dependencies, many studies also consider spatial correlations. However, challenges remain in reducing computational complexity and enhancing model diversity. Most research codes remain unpublished, likely due to sensitive data and proprietary model designs, limiting the reproducibility of experimental results. Although flight delay prediction studies show promising results, the absence of publicly available code and detailed implementation documentation hampers reproducibility and limits validation and extension by the broader research community. Encouraging open-source practices and standardized benchmarks is essential to ensure transparency and foster progress in this rapidly evolving field. Current applications focus on delay prediction at both individual airports and airport networks. Future directions include optimizing flight schedules, reducing fuel consumption, enhancing operational efficiency, and improving service quality. Based on Tables 2 and 4, and Fig. 9, future air traffic delay prediction is expected to focus on uncovering causal relationships in PDs and developing spatiotemporal frameworks using advanced deep learning methods. With the emergence of LLMs, the aviation sector is poised to develop domain-specific LLMs to support decision-making for airports, airlines, and air traffic control.

2.4 Large language models in transportation

With amazing progress in the fields of artificial intelligence and natural language processing, Large Language Models (LLMs), such as ChatGPT, have demonstrated their capability in content generation. This also shows great potential for air traffic management and control. As shown in Table 5, the application of LLMs in transportation is summarized into the early development stage and the rapid development stage. Although LLMs designed for ground transportation can theoretically be adapted to air transportation by modifying input data and operational rules, their practical application necessitates extensive experimentation to account for the unique complexities of air traffic. Some studies have begun exploring air traffic-specific LLMs, particularly in traffic flow management and operational safety. Jin et al. (2021) applied large-scale traffic data for prediction training, and a bidirectional encoder representation from transformers (BERT) framework was proposed to predict overall traffic flow. Xue et al. (2022) designed an AuxMobLCast pipeline, which used language-based models to mine timing patterns and was used to predict the passenger flow of each tourist attraction.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Gradually, the field of traffic management has developed a variety of traffic fundamental models (TFM) that focus on solving specific traffic problems. Wang et al. (2023d) introduced the TFM and integrated it with LLMs by incorporating traffic simulation principles into prediction tasks. Utilizing graph structures and dynamic graph generation algorithms, the model effectively captures complex interactions within the transportation system. However, traditional traffic-based models are typically limited to handling single-input, single-output scenarios.

Additionally, LLMs have certain limitations in traffic decision-making, including data privacy concerns, insufficient domain-specific training data, and a lack of interpretability (Kaddour et al. 2023). This is prominent when it involves digital data and direct interaction with traffic management and control systems. Therefore, Zhang et al. (2023b) proposed TrafficGPT, an AI framework integrating ChatGPT with traffic models to bridge the gap between LLMs and transportation systems. Its typical framework is shown in Fig. 10. By integrating multimodal data and predefined prompts, TrafficGPT extends ChatGPT's capabilities in processing traffic data, analyzing systems, and supporting decision-making. It enables stepwise decomposition of complex tasks via traffic models and enhances traffic control through natural language interaction, feedback integration, and iterative result refinement, improving system adaptability and reliability. Wang et al. (2023a) introduced TransWordNG, a traffic simulator leveraging number-driven algorithms and graph computing techniques to learn traffic dynamics from real-world data.

Lai et al. (2023) suggested LLMLight, an innovative traffic signal control method utilizing LLMs for enhanced traffic management through generalization and zero-shot reasoning abilities. Zheng et al. (2023) proposed a novel autonomous vehicle-following strategy by integrating cooperative adaptive cruise control (CACC) with twin delayed deep deterministic policy gradient (TD3) algorithms. Wang et al. (2023c) provided the AccidentGPT, a multi-modal model for accident analysis and prevention, establishing a framework for multi-sensor perception and integrated traffic safety. Chu et al. (2024) presented a trajectory generation framework based on the diffusion model (TrajGDM), aiming to construct a spatial intelligence foundational model with generalization and emergence capabilities. Pang et al. (2024) suggested iLLM-TSC, a traffic signal control (TSC) framework integrating LLMs with reinforcement learning (RL). Experiments showed that under degraded communication, iLLM-TSC reduced average waiting time by 17.5% compared to traditional RL, highlighting its potential for an intelligent transportation system. Liu et al. (2024) introduced the ST-LLM, a spatiotemporal large language model that encoded traffic timesteps as tokens via specialized embeddings. The model employed a partially frozen attention mechanism to retain pretrained knowledge while enabling traffic prediction adaptation. Experimental results demonstrated state-of-the-art performance, including few-shot and zero-shot capabilities, confirming LLMs' effectiveness for spatiotemporal learning tasks.

Intelligent transportation systems also face challenges such as data quality issues and limitations in simulation methods, which impact their effectiveness. Leveraging the advanced common sense, reasoning, and planning capabilities of LLMs, researchers have conducted studies to address these challenges. As shown in Table 5, early LLM development focused on traffic flow prediction, while recent advancements have shifted attention to traffic management and control, including urban traffic management, signal control, safety management, and pedestrian flow prediction. However, it is important to note that LLMs demonstrate strong language processing abilities, but they lack a deep understanding of

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 4 Summary of related works on dataset, quantitative metrics, strengths, weaknesses, applicable scenarios and code availability

Literature	Dataset	Data source	Publicly available	Data scale	Prediction horizon	Index	Strengths	Weaknesses	Applicable scenarios	Code
Khanmohammadi et al. (2016)	FO data	BTS	√	Jan. 2012: (1099 flights)	Delayed flights	RMSE, Memory usage, Run time	Effective for the prediction of defects	Complexity (variables, connections)	Predict delayed flights of specific airports	NA
Kim et al. (2016)	FO data	BTS	√	Jan. 2010 to Aug. 2015	The delay status of a day	Accuracy	Predict individual flight delays	Other deep architectures applications	Predict the delay status	NA
Li et al. (2018)	Airport operation data	Airport	NA	1 year (2016)	The level of delays	Calculation cost, actual accuracy, loss	Calculate the delay level	How to capture more information	The airport delay prediction	NA
Gui et al. (2020)	ADS-B	Data platform	NA	Dec. 2018 to May 2019	Individual flight delay	Accuracy, errors	Effective structure	The overfitting problem	Predict individual flight delays	NA
Bao et al. (2021)	FO	BTS	√	Jan. 2015 to Dec. 2019 27,080,431 records	1 h	RMSE, MAE	Better performance at some types of airports	Without considering the air route situation	Network-wide flight delay prediction	NA
Cai et al. (2022)	FO	CAAC	NA	1 year (2018)	Hourly delays	RMSE, MAE, MSPE	Markov property	Less comprehensive scenario	Time-varying patterns, spatial interactions	NA
Li and Jing (2022)	FO	VariFlight	Request	Jun and Aug 2016: 762,415 samples	Yes, no	ROC curves	A high-accuracy prediction model	Without consider congestion	Flight delay prediction from a temporal and spatial perspective	NA
Wang and Chen (2022)	FO	BTS	√	Jan. 2017 to Dec. 2021: 30,940,455 entries	Hourly delays	MAE	Spatial dependency	Hard to solve spatiotemporal forecasting tasks	Flight delay prediction	NA
Wang et al. (2022b)	FO	NA	NA	1 Jan. 2019 to 1 Jan. 2020: 606,139 flight records	Yes, no	RMSE, MAE, MSE, Var, Time	Help the airport improve its management ability	Some factors cannot be analysed	Classify direct and indirect factors of delays	NA

Springer
Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 4 (continued)

Literature	Dataset	Data source	Publicly available	Data scale	Prediction horizon	Index	Strengths	Weaknesses	Applicable scenarios	Code
Wang and Chen (2022)	FO	CAAC	NA	Jan. 2018 to Jan. 2019; 874,591 flights	from 10 to 30 min	RMSE, MAE, MAPE, SMAPE	Improve node feature representation ability	Lacking consideration of air traffic's operation rules	Multi-airport flight delay prediction	NA
Li et al. (2023)	FO	BTS	√	Jan. 1 and Dec. 31, 2019: 5,426,150 flights	Hourly delays	Accuracy; Precision; Recall; TPR; FPR	Spatial–temporal correlations	Determine an appropriate grid size	Spatial–temporal correlations for flight delay prediction	NA
Wu et al. (2024)	FO	BTS, China dataset	Partially	U.S. dataset: Seven-year: Jan. 2015, to Dec. 2021; China dataset: Apr. 2015, to May, 2017	6 h, 12 h,	RMSE, MAE, R^2	Spatiotemporal dependencies	The structure for handling missing data	Multi-step delay in large airport networks	NA
Sun et al. (2024)	FO	CAAC	NA	Mar. 2018 to Feb. 28, 2019: 1.8 million flights	Hourly delays	RMSE, MAE, Accuracy, R^2, Var	Causality, spatiotemporal dependence	Less comprehensive scenario	Propagated delay prediction	NA
Li et al. (2024b)	FO	BTS, xiecheng	√	BTS: Jan. 2015, to Dec. 31, 2021; China data: Apr. 2015, to May, 2017	Hourly delays	RMSE, MAE	Consider weather	Do not consider flight chains	Predict delay propagation in airport networks	NA
Bala Bisandu and Moulitsas (2024)	FO	BTS, xiecheng	√	Jan. 2021: 361,428 records. Nov. 2021: 547,559	Predict the total number of delays for each day	RMSE, MSE, MAE, MAPE	Improve the model performance	Data limitations	Optimize flight schedules, reduce fuel consumption, and enhance service quality	NA

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 4 (continued)

Literature	Dataset	Data source	Publicly available	Data scale	Prediction horizon	Index	Strengths	Weaknesses	Applicable scenarios	Code
Xu et al. (2024)	Trajectory data, FO	SDW	√	Agu. to Oct. 2019	Predict critical air traffic state parameters	RMSE, MAE, MAPE	Learning capabilities with higher data volumes	Without considering congestion scenarios	Air traffic state prediction, dynamic spatial-temporal dependencies	√
Franco et al. (2025)	FO; weather reports	ICEA; METAR and METAF reports	NA	42,336 observations	Yes, no	Accuracy, Precision, Recall, F1-Score	Improved fuel efficiency	Lacking the integration of real-time and weather data	Flight delays due to holding maneuvers	√

FO data: Flight operation data; the U.S. Bureau of Transportation Statistics (BTS) database; http://www.transtats.bts.gov; CAAC: Civil Aviation Administration of China; DMP-ANN: Defect of modules prediction artificial neural network (ANN); ADS-B: Automatic Dependent Surveillance-Broadcast; VariFlight: https://data.variflight.com; China dataset: xiecheng; SDW: Sherlock Data Warehouse; "√" denotes the dataset/code is fully publicly available and can be accessed directly from its official source; "Partially" indicates partially public data requiring additional approval for full historical access; "Request" denotes restricted data available only to authorized members/institutions; "NA" signifies non-public data/code

Springer Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 5 Application of LLMs in the field of transportation

Periods	Literature/Model	Application areas	Characteristics	Performance	Aviation applicability
Early stage of development (2016~2022)	Duan et al. (2019), CPPBTR	Urban crowd flow forecast	Two-stage framework based on Transformer	Superior effectiveness	+
	Jin et al. (2021), BERT	Traffic flow forecast	Multi-headed self-attention	High accuracy	+
	Xue et al. (2022), AuxMobLCast	Human activity prediction	Language basic model, POI classification	Comparable to advanced numerical methods	-
Rapid development stage (2023~)	Wang et al. (2023), TFM	Traffic flow forecasting	Dynamic graph	Accurately predict urban traffic	+
	Zhang et al. (2023b), TrafficGPT	Urban traffic management	Combined with ChatGPT	Intelligently complex tasks	+
	Wang et al. (2023a), TransWordNG	Traffic simulation and management	Data-driven, dynamic graph	Generate realistic traffic patterns	+
	Lai et al. (2023), LLMLight	Traffic signal control	LLMs, chain thinking reasoning	Advanced effect on world traffic data	-
	Zheng et al. (2023), CACC and TD3	Autonomous driving and car following strategies	CACC, TD3, adaptive adjustment	Collisions were reduced	+
	Wang et al. (2023c), AccidentGPT	Traffic accident analysis and prevention	GPT-4 V, environment awareness	Proactive warnings and effective accident analysis	+
	Chu et al. (2024), TrajGDM	Human mobility	Remove the uncertainty	Interpret the latent space	-
	Abdulhak et al. (2024), CHATATC	Air traffic flow management	LLM-driven with natural language interaction	Accurate for specific queries (e.g., airport, weather)	√
	Fox et al. (2024), VAE-LLM	Aviation safety	LLM, VAE	Accuracy is great (88.4%)	√
	Pang et al. (2024), iLLM-TSC	Traffic signal control	Combined with RL	Enhance the robustness	+
	Tabrizian et al. (2024), Evtol-LLM	Optimize flight paths	eVTOL operation	Dealing with human requirements	√
	Liu et al. (2024), ST-LLM	Traffic prediction	Global spatial-temporal dependencies	ST-LLM outperforms state-of-the-art models	+

+: It indicates that the research is based on a ground traffic framework but can be extended and adapted to air traffic applications

–: It indicates that the approach does not apply to air traffic

√: It indicates that the research is conducted about air traffic and applies to this domain

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

traffic-related complexities. Traffic management requires real-time responses at the millisecond level (e.g., traffic light control, tactical level air traffic flow management, tactical level conflict detection and resolution, and accident warnings), but the reasoning speed of LLMs is constrained by model complexity, making them unsuitable for low-latency tasks (Chu-Carroll et al. 2024). While LLMs excel at processing natural language, traffic systems depend on structured data (e.g., sensor data, GPS coordinates, ADS-B data, and flight operation data), requiring additional translation layers that complicate their application and limit their effectiveness in handling numerical data and interactive tasks.

Hence, to enhance traffic system operation, it is essential to integrate ChatGPT with basic traffic models and leverage multi-modal data for traffic pattern learning and decision support. Additionally, combining RL models with LLMs can optimize basic traffic problems, ensuring real-time performance and reliability while enhancing system flexibility and interpretability. In summary, the cross-application of AI with fundamental traffic problems can facilitate intelligent task decomposition, enable natural language dialogue in air traffic, and provide timely interactive feedback, improving system reliability and applicability.

3 Challenges and future research directions

This section outlines current research challenges and future directions in terms of causality, deep learning, and LLMs.

3.1 Causality in propagated delays

The air traffic system exemplifies a complex system, where causal analysis is often applied to delay sequences between airport pairs to uncover intrinsic temporal dependencies. Traditional explicit network-based prediction methods face challenges in analyzing the robustness and resilience of air traffic networks with time-varying topologies, as most conventional models assume static network structures when modelling propagated delays (Cai et al. 2020). However, air traffic networks are inherently dynamic, influenced by operational, environmental, and regulatory factors. This dynamic nature introduces complexity, as predictions must account for temporal variations in network connectivity. Additionally, flight delay time series exhibit both temporal and spatial variability due to aircraft movements between airports, expanding the spatiotemporal scope of analysis. These factors complicate delay propagation modelling and necessitate a deeper understanding of spatiotemporal dependencies and their causal relationships.

Current research on causality remains limited. GCT assumes temporal precedence of cause over effect and requires stationary data, often necessitating first- or second-order differencing (Hasan et al. 2023). It also assumes no confounders or instantaneous effects, implying that changes in one variable cannot simultaneously affect another at the same time point. In constructing delay propagation models using complex network theory and analyzing delay time series with GCT, Zanin et al. (2017) assumed that the delay time series from airports A to B are stationary. However, actual operations are more complex, as delays can arise from various factors such as airport operations, weather, and other events. Large airports, equipped with advanced facilities and robust emergency procedures, can promptly mitigate large-scale delays and limit their spread. In contrast, small airports, with less com-

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

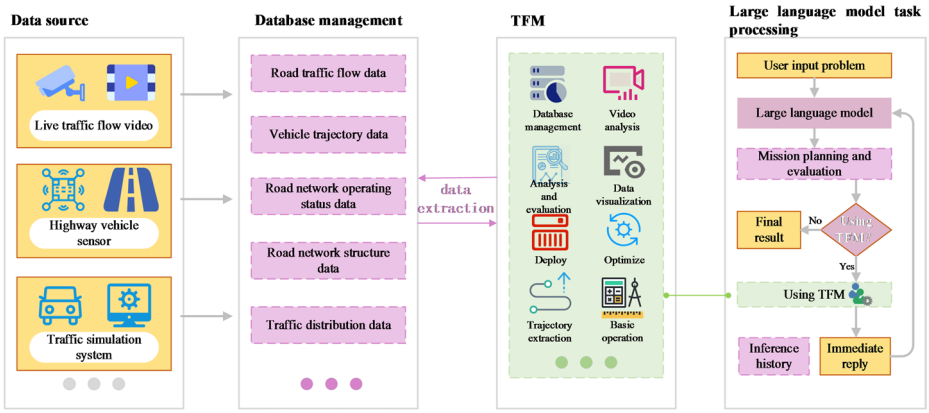

Fig. 10 Application framework of TrafficLLM (redrew) (Zhang et al. 2023b)

prehensive infrastructure and emergency response capabilities, may be less effective in managing disruptions, leading to increased risk of PDs. These operational differences influence the analysis of causality in delay propagation. For spatiotemporal analysis of flight delays in air traffic networks, Bombelli and Sallan (2023) used cases without extreme weather as a baseline, analyzing four U.S. airlines and two extreme weather events. Particularly, GCT was applied to assess causal relationships between hourly delays in airports. Nevertheless, the study only considered extreme weather, excluding other operational factors such as controller status, flow control, or special activities, which may influence propagated delays differently. Additionally, varying operational environments and airline alliance cooperation can alter delay patterns, suggesting that causal sufficiency requires further examination.

In contrast, transfer entropy (TE) serves as a nonlinear extension of Granger causality, capable of detecting both linear and nonlinear causal relationships in time series data. Oh et al. (2021) proposed a lag-specific transfer entropy (lag-specific TE), and explored the delay methods between roads under different congestion modes. Wang et al. (2022b) introduced a causal flight delay prediction model for a single airport using an attention-enhanced LSTM (LSTM-AM) to forecast delays and identify their primary causes. Tan et al. (2022) developed a deep temporal convolutional model with an attention mechanism to capture inter-airport delay dependencies within a unified network framework. This approach revealed both the causal links and the intensity of delay propagation within the airport network. Çelik et al. (2025) examined the air transport-economic growth causality in the ten largest air transport countries (1970–2021) using bootstrap tests. Results showed predominant unidirectional causality, with Fourier-based tests proving more robust and highlighting the limitations of single-dimensional air transport indicators.

Furthermore, a fully connected delay propagation network based on TE method was established to address time lags in delay propagation (Li et al. 2024c). A causality-biased random walk was introduced to explore the temporal cascade effect of root delays, generating a delay propagation tree for each airport. However, the study relied on annual flight operation data, encompassing delays and cancellations caused by weather, airline operations, air traffic control, and other factors. These factors can reduce airport capacity, create imbalances between capacity and demand, and increase the likelihood of delays, particularly under adverse weather conditions along flight routes. Additionally, the interconnected nature

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

of flights and shared resources such as aircraft, crew, and airports leads to network delays, as initial delays are transmitted to downstream flights. This complicates causal analysis, as not all system variables are observable, and the assumption of causal sufficiency is often unmet. What's more, the stationarity of delay time series cannot be guaranteed, particularly during periods of frequent disruptions such as summer thunderstorms. Delay causes may also interact, creating a "snowball effect" that amplifies system-wide delays. Consequently, strong assumptions, especially causal sufficiency, are required for causal inference, but these assumptions are often violated in actual operations, challenging the validity of such analyses. Accordingly, Peral (2009) introduced structural causal models (SCMs) to formally represent the structural knowledge of data-generating processes. SCMs are crucial tools for causal reasoning and decision-making, as they elucidate the underlying causal mechanisms (Kaddour et al. 2022). However, identifiability remains a key challenge; observational data alone rarely suffice to uniquely determine the causal graph, especially given the potential for multiple compatible structures and the complexity of high-dimensional data. Another significant limitation is the scarcity of benchmark datasets with known ground truth for training and evaluating causal models (Shimizu et al. 2006). Accurate assessment requires comprehensive public repositories of real causal graphs, which are currently lacking.

Consequently, future research can focus on developing methods and tools capable of capturing these dynamics, with particular emphasis on causal inference, to better understand PDs and mitigate congestion in adaptive network structures. Moreover, integrating neural networks with causal inference methods offers a promising approach for identifying causality in network delays. Given the temporal and spatial interdependencies within airport networks, combining graph neural networks with causal methods (e.g., GCT, TE) to model PDs at network levels represents a significant direction for future research. What's more, as causal discovery methods continue to expand, future research should prioritize the use of real data and account for operational limitations to enhance accuracy and reliability. Incorporating background knowledge, such as domain expertise and literature evidence, into causal models is also crucial for addressing current challenges. Resolving these issues will enable the development of more accurate and robust causal discovery methods.

3.2 Emerging deep learning in network delay prediction

Flight delays represent a persistent challenge in air transportation systems. Deep learning offers distinct advantages over classical and network-based prediction methods by effectively capturing spatiotemporal dependencies in network delays through distributed hierarchical feature representation. A key challenge in applying deep learning to network delay prediction is ensuring model applicability and scalability across diverse airports, countries, and datasets (Li and Jing 2021). Achieving high accuracy and robustness in varying operational environments requires comprehensive validation to assess model generalization. Enhancing scalability in this context can offer valuable insights into global delay propagation and contribute to more effective delay recovery management in international air traffic systems. In addition, deep learning methods for network prediction typically rely on large historical datasets and assume stable data distributions over time. However, airline and airport operations are subject to dynamic changes driven by unforeseen events, economic shifts, and environmental factors, leading to distributional shifts that deviate from conventional conditions. In such scenarios, data must be treated with consideration for these anomalies, and

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

models must adapt to evolving environments while maintaining performance (Carvalho et al. 2021). The continuous development of the aviation industry further demands predictive models that are both accurate and resilient to dynamic operational changes.

Spatiotemporal graph neural networks are well-suited for modelling complex traffic systems, as they capture both temporal and spatial dependencies in graph-structured data (Jiang and Luo 2022). By integrating spatiotemporal information, these models effectively represent interactions among nodes and their attributes in air traffic networks, enhancing performance in tasks such as delay prediction and propagation modelling. While notable progress has been made in spatiotemporal analysis of flight delays and traffic flow, further research is needed to improve predictive accuracy and deepen the understanding of spatiotemporal dependencies.

(i) Coupling spatiotemporal dependency: Network delay prediction is influenced by the airport connectivity and operational capacity across different scales. Large airports often have well-developed delay response strategies, enabling effective mitigation of large-scale disruptions. In contrast, small and medium-sized airports typically lack comprehensive response mechanisms and infrastructure. Enhancing delay prediction accuracy and establishing robust spatiotemporal frameworks (e.g., GCN, GAT) can improve overall airport operations, enabling large hubs to optimize responses and supporting smaller airports in anticipating and managing delays more effectively.
(ii) Dynamic graph structure: Existing studies predominantly model airport networks as static structures, overlooking their inherent dynamic nature arising from time-varying flight schedules and spatial variations in aircraft operations between airports. This fundamental limitation necessitates the development of dynamic spatiotemporal graph-based approaches to accurately represent the evolving nature of air traffic networks.
(iii) Multi-layer graph network structure: Air traffic networks demonstrate complex interrelationships among nodes, including spatial, causal, and flight schedules. Flight operations involve multiple phases (taxi-out, climb, cruise, descent, taxi-in) across three interconnected networks: terminal control networks, route network, and airport network. These networks interact dynamically under weather and operational uncertainties. To improve delay prediction accuracy and understand inter-network interactions, it is essential to develop predictive models based on a multi-layer graph network framework that accounts for such uncertainties.

3.3 A conceptual AirTraffic LLM

3.3.1 Overview

Although LLMs excel in general tasks, their domain-specific knowledge is limited. Fine-tuning is necessary to account for language-specific nuances and specialized requirements. Consequently, researchers have developed domain-specific LLMs to address these limitations. For example, BloombergGPT (Wu et al. 2023) and Xuan Yuan 2.0 (2023c) specialized in financial services, while LexiLaw (2024) focused on legal applications. Similarly, Lawyer Llama (Huang et al. 2023) catered to legal domains, whereas Huatuo (2023b) and ModioGLM (2024) served the medical field. These models play a crucial role in their respective domains.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Inspired by the TrafficGPT and its operation in ground transportation (Zhang et al. 2023b), this study proposes an AirTraffic LLM framework, shown in Fig. 11, a conceptual framework integrating structured prompts with ChatGPT to enhance interaction with aviation data (aviation-related books, reports, flight plan data, trajectory data, et al.) and systems. The framework combines LLM capabilities with fundamental air traffic models, including airspace models (Künnen and Strauss 2022) and aircraft performance models (full energy model, motion model, aerodynamics model, and thrust model). Specifically, the aircraft performance model preliminarily adjusts historical delay data to account for variations in PDs across different aircraft types. Historical flight operation data are used to identify commonly operated aircraft models between city pairs and airports. Performance parameters for each model- including climb rate, descent rate, cruise altitude, and speed- are extracted from EUROCONTROL's Base of Aircraft Data (BADA) (EUROCONTROL 2009). These parameters, combined with the original delay time series, are input into the spatiotemporal embedding module. Additionally, the airspace model incorporates airspace structure (including air routes and sectors), route congestion thresholds, and capacity limits, all of which influence the spatial dynamics of delay propagation between airports. Airspace capacity and route congestion serve as key factors in constructing the graph structure within the graph convolutional network (GCN). Accordingly, the GCN adjacency matrix is dynamically composed of two components: causal relationships and spatial structure. Through weighted integration and dynamic adjustment, the adjacency matrix is optimized, enhancing the accuracy of network delay predictions. These initiatives improve LLM's capabilities, enabling it to process air traffic data and systems and provide auxiliary decision-making support in the air traffic field.

3.3.2 Spatial-temporal causal LLM in network delay prediction

Flight delay propagation reflects the interconnectivity of airport networks, where airports serve as hubs for passenger and cargo transfers. To maximize resource utilization, airlines often schedule the same aircraft for multiple flight segments (typically 4–5 per day) (Kafle and Zou 2016), resulting in shared dependencies across flights involving aircraft, crew,

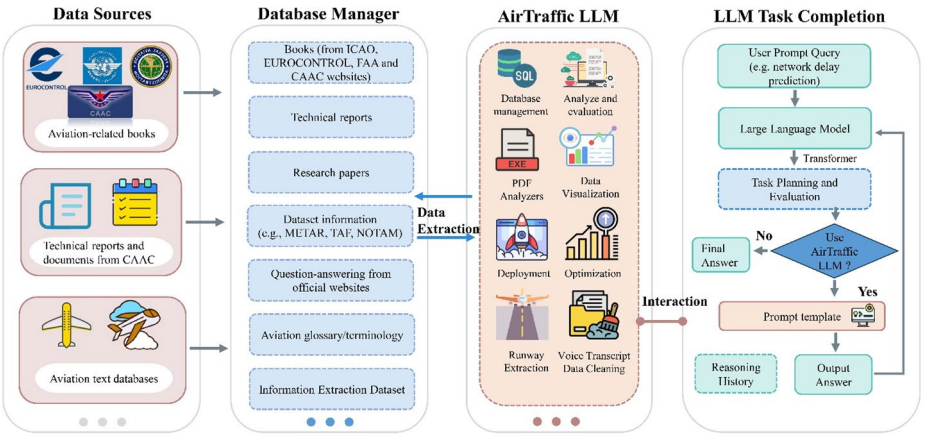

Fig. 11 Application framework of AirTraffic LLM

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

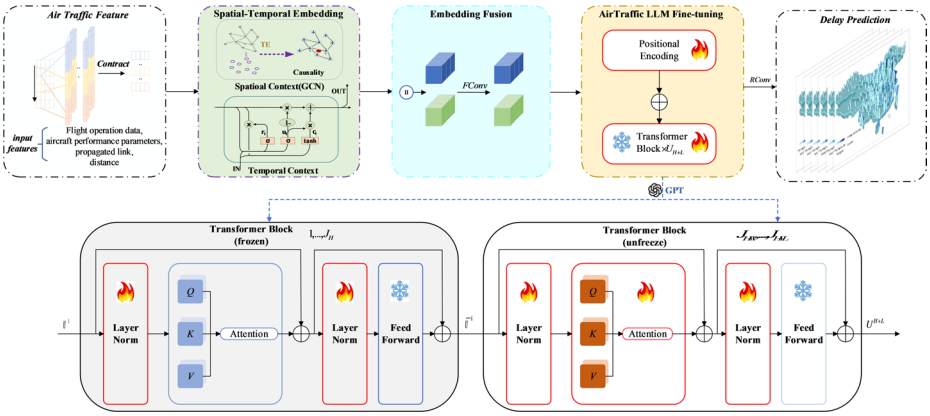

Fig. 12 STC-LLM framework for air traffic delay prediction

and passenger resources. In addition, traditional air traffic prediction models often lack interpretability due to their "black-box" nature (García-Sigüenza et al. 2023). Embedding causal knowledge enhances transparency, improving the accuracy and reliability of air traffic models. This facilitates a deeper understanding of causal relationships network delays, aiding civil aviation personnel and researchers in improving decision-making and operational efficiency. Hence, since delay propagation within flight chains directly reflects the causal dynamics of airport networks, applying causal discovery methods to identify delay relationships is highly valuable. This enables the identification of affected airports and key nodes within the delay propagation network, thereby enhancing the accuracy of network delay prediction.

Consequently, considering spatiotemporal dynamics (Liu et al. 2024), the AirTraffic LLM framework can be applied in air traffic delay prediction, a novel Spatial-Temporal Causal Large Language Model (STC-LLM) framework is proposed shown in Fig. 12. Notably, LLMs are typically pre-trained models with billions of parameters, primarily based on the Transformer decoder architecture (Min et al. 2023). The training is mainly divided into pre-training (base model) and post-training (instruct model), and then downstream applications are carried out through testing (reasoning). Between pre-training and post-training, fine-tuning is typically used to enhance task execution and generalization. By freezing the base model and updating only a small set of additional parameters, it reduces trainable parameters and memory usage while maintaining performance comparable to full fine-tuning.

As shown in Fig. 12, STC-LLM framework employs a spatiotemporal embedding architecture to capture the temporal and spatial dependencies of the airport delay propagation network. To support the practical implementation of the proposed STC-LLM framework, we specify key requirements in data, computation, and validation. The model integrates multi-source heterogeneous data (Pineda-Jaramillo et al. 2024), including historical flight operation data (Li et al. 2024a), distance, aircraft performance parameters (e.g., flight level, air pressure, true airspeed, and fuel burn rates, etc.) (EUROCONTROL 2009; Nuic et al. 2010), and domain-specific textual corpora (e.g., regulations, manuals) (NOAA 2005). Effective training requires tactical-level (15-minute to 60-minute) temporal resolution (Sun et al. 2022; Cai et al. 2022), airport or sector-level spatial granularity, and high-quality, synchronized data. Computationally, the framework demands high-performance GPU clusters

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

(e.g., NVIDIA A100) (Zhong et al. 2025), multi-terabyte storage, and substantial memory to handle long-context LLM inputs. A three-stage validation process is proposed (Amat Rodrigo and Escobar Ortiz 2023): (a) offline back-testing with historical data, (b) cross-validation across multiple airport networks to ensure generalizability, and (c) stress-testing under simulated congestion to assess robustness. These refinements aim to bridge the gap between conceptual design and real-world deployment.

By analyzing causal relationships in propagated delays between airports, the model enables spatiotemporal delay prediction across the network. Notably, the framework employs a GCN based on the TE method to capture causality in network delays, and a gated recurrent unit (GRU) to model temporal dependencies. A fusion convolution integrates spatial and temporal features, transforming the spatiotemporal embeddings into representations compatible with LLMs. These embeddings encode the network's delay causality, hourly delay patterns, and daily delay dynamics. Subsequently, a partially frozen attention (PFA) architecture with $\boldsymbol{H}$ frozen and $\boldsymbol{L}$ trainable layers is then applied in STC-LLM fine-tuning. The first frozen $\boldsymbol{H}$ layers follow the Frozen pretrained Transformer (FPT) to preserve prior knowledge, while the last unfrozen $\boldsymbol{L}$ layers enhance the model's ability to capture spatiotemporal dependencies. Given the large scale of GPT-4's Transformer architecture and its high computational demands, local fine-tuning faces practical limitations. Therefore, using smaller-scale models for preliminary framework validation is both practical and meaningful. While the full GPT-3 model offers strong performance, it suffers from slower response times, higher costs, and deployment challenges. In contrast, GPT-2 lags behind GPT-3 and its variants in terms of training scale, data quality, and learning capacity (Floridi and Chiriatti 2020). A smaller GPT-3 variant (e.g., Curie) offers a balanced trade-off between performance and computational efficiency (Braun 2022), making it well-suited for integration into composite frameworks due to its speed, cost-effectiveness, and ease of deployment. Thus, the STC-LLM adopts a Transformer framework and utilizes a small variant of GPT-3 (such as Curie). An additional normalization layer is introduced after the final multi-head attention layer. The model outputs the predicted delay time series, which is subsequently visualized on a map.

3.3.3 AirTraffic-reasoning LLM

After post-training, the instruction model is applied to downstream tasks by employing intermediate reasoning steps to solve complex problems. As conditional probability generators, LLMs benefit from chain-of-thought reasoning, which improves prediction accuracy. Models capable of producing extended reasoning chains are often referred to as reasoning models. In this study, AirTraffic-Reasoning is a reasoning model trained based on AirTraffic-Base, and the formal representation of the reasoning process is provided in formula (5).

Specifically, the proposed AirTraffic-Reasoning framework is designed to enhance the model's capacity for step-by-step inference in flight delay identification and causal propagation analysis. The framework integrates a reinforcement learning (RL) training objective with structured data sources, a task-specific reward design, and a guided training procedure (DeepSeek-AI et al. 2025; Cao et al. 2025). The RL objective aims to equip the base model with reasoning capabilities necessary for downstream supervised fine-tuning. Training data incorporates airport-pair information (e.g., distances and delay propagation links), delay time series with classification rules, airspace and aircraft performance models, expert knowl-

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

edge on causal mechanisms, and reference implementations from the literature. The reward function consists of two components: an accuracy reward, which evaluates delay detection, causal network construction, and physical-model-based reasoning; and a language reward, which promotes structured, interpretable outputs following the <think>→<answer> format. The model is trained using Guided Reward Preference Optimization (GRPO) (Christiano et al. 2023) over thousands of steps, with performance monitored through reward progression, reasoning trace development, and output length evolution. This setup enables the model to develop human-like explanatory reasoning in the context of network-level air traffic delay propagation.

$$AirTraffic - Reasoning(p) = \langle < think > r < /think >, < answer > \hat{y} < /answer > \rangle \quad (5)$$

In formula (5), AirTraffic-Reasoning is the trained reasoning-capable model, p shows the input prompt, such as "Determine whether the flight is delayed and explain why." <*think*> r </*think*> suggests the model's step-by-step reasoning trace, reflecting a thinking process (e.g., checking scheduled vs. actual times, causal indicators like transfer entropy values, etc.). <answer>$\hat{y}$ </answer> shows the final output, e.g., "Delayed, caused by propagation from origin airport."

To ensure operational safety and practical adoption, the AirTraffic-Reasoning LLM is deployed within a human-in-the-loop decision-support workflow (Mercer et al. 2016; Perott et al. 2019), providing advisory recommendations with concise rationales, calibrated uncertainty estimates (He et al. 2025), and interactive causal propagation visualizations (Hurter et al. 2022) integrated into air traffic management interfaces (Yeh and Ravikumar 2021; Zohrevandi et al. 2022). Controllers and dispatchers retain full decision authority, with their feedback systematically logged to support supervised fine-tuning and reinforcement learning from human feedback. By embedding interpretability, traceability, and iterative human feedback, the framework ensures trustworthy, transparent, and operator-aligned decision-making in real-world air traffic operations.

3.3.4 Advantages and limitations

Although LLMs are trained in natural language, they function as powerful autoregressive sequence learners capable of modelling token-level dependencies, contextual associations, and cross-modal embeddings. Recent studies show that LLMs can generalize to structured tasks—such as table prediction, graph embedding, and time series modelling—when equipped with suitable input mappings and embedding mechanisms. Nassar et al. (2022) proposed a novel table structure recognition approach by introducing a table cell object detection decoder, enabling direct extraction of table content from PDF files without custom OCR training. Replacing the LSTM decoder with a Transformer-based decoder further improved the evaluation score from 91% to 98.5%. Jiang et al. (2023) enhanced LLMs' reasoning ability on structured data through a unified Iterative Reading-then-Reasoning (IRR) framework for question answering. Experiments across three types of structured data demonstrated significant performance gains in both few-shot and zero-shot settings. Sui et al. (2024) investigated LLMs' capacity to process structured data, such as tables, by designing a benchmark to evaluate structural understanding across seven tasks, including cell lookup, row retrieval, and size detection. Evaluations using models like GPT-3.5 and

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

GPT-4 revealed performance sensitivity to input formatting. To address this, a self-reinforcing prompt method was proposed. The open-source benchmark and prompting strategy offer a simple and generalizable tool for future research.

Additionally, structured data is encoded into unified token sequences, aligning it with a language modelling framework and enabling adaptive input processing. In our proposed framework, outputs from the GCN (spatial causality) and GRU (temporal dependencies) are fused and serialized into spatiotemporal token sequences, which are then fed into a partially fine-tuned part. This integration bridges the structural gap and enables the LLM to capture higher-order, long-range dependencies beyond the capabilities of traditional models. In terms of model complementarity, STC-LLM offers unique advantages, such as long-context modelling, multi-path causal reasoning, and scalability across heterogeneous data, that complement the localized learning capabilities of GCNs and GRUs. The proposed PFA architecture enables efficient domain adaptation while retaining pretrained generalization.

In general, LLMs enhance the efficiency, intelligence, and sustainability of transportation systems (Long et al. 2025; Mahmud et al. 2025). However, their application in complex air traffic tasks remains limited. Challenges such as data privacy, quality, and model bias (Boateng et al. 2024) persist due to the substantial data requirements, posing obstacles to practical implementation. Carlini et al. (2019) demonstrated the extraction of personal identifiers (e.g., phone numbers, emails) using GPT-2. Feldman (2021) further noted that undetected personal data in training sets can be unintentionally retained and revealed by the model. What's more, LLMs face challenges in interpretability, as their decision-making processes remain opaque, limiting human trust (Zhang et al. 2024). Accuracy is affected by factors such as training data quality, model size, and irrelevant features, leading to higher error rates. Privacy risks arise from exposure to sensitive data during training and user interactions (Gan et al. 2024; Das et al. 2025), exacerbated by the models' complexity and opacity. Especially when applying LLMs to air traffic, several limitations emerge:

- Data privacy is a critical concern, particularly in the context of flight operation data. In most countries, such data is not publicly accessible and typically requires formal collaboration with airlines, air traffic control authorities, or civil aviation regulators, along with a rigorous approval process. These datasets often include sensitive operational details such as flight plans, trajectories, delay records, and dispatch orders. Moreover, some airlines embed passenger information (Carlini et al. 2019), such as contact details, email addresses, and membership numbers, within operational data. Without strict control over data inputs, models like LLMs risk exposing passenger privacy.
- Insufficient domain-specific training data: LLMs require large-scale pretraining to generalize effectively (Hadi et al. 2024). However, the scarcity of high-quality civil aviation data limits its reliability in this domain. As a result, outputs may reflect biases or inaccuracies that could misguide air traffic decisions. For instance, patterns learned from limited or non-specialized datasets may fail to capture the complexities of real-world airport operations.
- Lack of interpretability: As black-box models, LLMs offer limited transparency, posing challenges for critical tasks such as route planning and delay management. In air traffic, where safety and operational accountability are paramount, this opacity may hinder adoption by airlines and regulatory departments.
- Limited adaptability to real-time operational constraints: Civil aviation operates under

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

strict, dynamic rules that general-purpose LLMs often fail to capture, such as airspace restrictions or weather-induced diversions and alternate landings. These scenarios require real-time route re-planning, which LLMs are not inherently equipped to handle without domain-specific adaptation.

- Challenges in real-time deployment: Air traffic management demands real-time or near-real-time decision-making. However, LLMs, particularly large-scale models, incur high computational costs and latency, limiting their feasibility for real-time applications in operational environments.

In future research, to address the limitations of LLMs in air traffic applications, the following measures are proposed to support their integration into national air traffic systems:

First, to address data privacy concerns, as air traffic operations involve sensitive information, such as flight plans, operational dynamics, and dispatch commands, direct use of LLMs poses a risk of data leakage. Deploying lightweight LLMs locally, in combination with a federated learning framework, enables on-site training and updates without exposing sensitive data outside the operational domain (Zhang et al. 2024).

Second, issues such as delayed reporting, transmission interruptions, and false alarms in flight data can compromise LLM training and inference. Integrating a multi-source heterogeneous data fusion mechanism, combining ADS-B data, METER reports, flight plans, and more, enhances data quality through consistency checks and cleaning processes.

Third, LLMs may be overfit to major airports or high-frequency routes, overlooking marginal airports and atypical scenarios, thereby reducing prediction generalizability. Incorporating structured sampling and category balancing strategies during training ensures adequate representation of low-frequency airports and routes.

Fourth, key decisions in air traffic control, such as delay prediction and route recommendation, require human interpretability and traceability. The black-box nature of LLMs limits their applicability (Kaddour et al. 2023). To enhance model trustworthiness, post-hoc interpretability modules, such as attention-weight visualizations and causal path explanations, can be integrated.

Finally, due to complex airway structures and dynamic weather conditions, language models alone struggle to capture nonlinear spatiotemporal dynamics (Sun et al. 2024). Integrating LLMs with GCNs and Time Series Models (e.g., GRUs) enhances spatiotemporal modelling and improves prediction accuracy.

4 Conclusion

This paper combines different background knowledge with journal paper publishing data on air traffic network delay prediction to propose a systematic review. First, this study categorizes network delay prediction approaches into four aspects. It explores classical methods without explicit network topology modelling, traditional explicit network-based prediction methods, emerging deep learning methods, and the application of large language models in transportation. Second, traditional explicit network-based prediction methods are categorized into the early network development stage and causality modelling in network delays. This study emphasizes the latter, exploring the underlying operational mechanisms of air traffic. Representative technologies are then analyzed based on their characteristics. Third,

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

with advancements in AI and natural language processing, LLMs like ChatGPT exhibit strong reasoning, planning, and decision-making capabilities, offering significant potential for air traffic management and control. Therefore, considering LLM applications in transportation and the unique characteristics of air traffic, this paper proposes a conceptual framework for AirTraffic LLM. It aims to enhance network delay prediction and provide strategic decision support for airlines, airports, and air traffic control.

Notably, this study proposes a novel Spatial-Temporal Causal Large Language Model (STC-LLM) framework to address spatiotemporal causality in air traffic delay prediction. The framework integrates flight data, airspace models, and aircraft performance parameters, and leverages GCN, GRU, and transfer entropy methods to analyze delay propagation networks. A partially frozen attention mechanism is employed during fine-tuning, using a GPT-3 variant (e.g., Curie) with an additional normalization layer. The model outputs predicted delay time series, which are subsequently visualized on a map. The STC-LLM currently represents a conceptual architecture, with future work aimed at its empirical validation through large-scale experiments and operational implementation. Nevertheless, as LLMs in air traffic are domain-specific, their generalization to other fields is limited and less effective than general-purpose models. They also require extensive pretraining, yet data acquisition in air traffic is often constrained by privacy concerns, potentially leading to inaccurate outputs. Additionally, the black-box nature of LLMs hinders domain-specific interpretability, reducing transparency in decision-making for airlines and airport operators. While effective in multimodal tasks, LLM training remains resource-intensive, especially when processing large-scale multimodal data. Future research should focus on integrating emerging deep learning methods with causal methods (e.g., combine GCT with TE) for network delays. Promising directions include the development of spatiotemporal GCN that incorporate spatiotemporal dependencies, dynamic graph structures, and multi-layer architectures.

Acknowledgements This work was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX24_0600), Funding for Outstanding Doctoral Dissertation in NUAA (BCXJ24-16), China Scholarship Council (202406830101), China Postdoctoral Science Foundation Funded Project (2024M752347), Jiangsu High Level "Shuang-Chuang" Project (JSSCBS20220212), the Natural Science Foundation of Jiangsu Province (BK20230892), Talent Research Start-up Fund of NUAA (YAH22019).

Author contributions M.S. wrote the original draft, reviewed it, edited the draft, and provided the funding acquisition. Y.T. edited the draft, provided supervision, review, and editing, gave the funding acquisition, J.L. edited the draft, gave the funding acquisition, and performed the validation. C.W. reviewed and edited the draft. L.P. performed the data analysis. S.X. provided the resources and software. All authors reviewed the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

Abdulhak S, Hubbard W, Gopalakrishnan K, Li MZ (2024) CHATATC: large language model-driven conversational agents for supporting strategic air traffic flow management. arXiv, Singapore, pp 1-8.

AhmadBeygi S, Cohn A, Guan Y, Belobaba P (2008) Analysis of the potential for delay propagation in passenger airline networks. J Air Transp Manag 14:221–236. https://doi.org/10.1016/j.jairtraman.2008.04.010

Allan SS, Beesley JA, Evans JE, Gaddy SG (2001) Analysis of Delay Causality at Newark International Airport. Santa Fe, New Mexico, USA, p 1–11

Amat Rodrigo J, Escobar Ortiz J (2023) skforecast(version 0.17.0)

Bala Bisandu D, Moulitsas I (2024) Prediction of flight delay using deep operator network with gradient-mayfly optimisation algorithm. Expert Syst Appl 247:123306. https://doi.org/10.1016/j.eswa.2024.123306

Bao J, Yang Z, Zeng W (2021) Graph to sequence learning with attention mechanism for network-wide multi-step-ahead flight delay prediction. Transp Res C 130:103323. https://doi.org/10.1016/j.trc.2021.103323

Baspinar B, Koyuncu E (2016) A data-driven air transportation delay propagation model using epidemic process models. Int J Aerosp Eng. https://doi.org/10.1155/2016/4836260

Bauer M, Cox JW, Caveness MH et al (2007) Finding the direction of disturbance propagation in a chemical process using transfer entropy. IEEE Trans Control Syst Technol 15:12–21. https://doi.org/10.1109/TCST.2006.883234

Bergantino AS, Gardelli A, Rotaris L (2024) Assessing transport network resilience: empirical insights from real-world data studies. Transp Rev. https://doi.org/10.1080/01441647.2024.2322434

Boateng GO, Sami H, Alagha A et al (2024) A survey on large language models for communication, network, and service management: application insights, challenges, and future directions

Bombelli A, Sallan JM (2023) Analysis of the effect of extreme weather on the US domestic air network. A delay and cancellation propagation network approach. J Transp Geogr 107:103541. https://doi.org/10.1016/j.jtrangeo.2023.103541

Braun J (2022) Verbal epistemic uncertainty estimation for numeric values with GPT-3. Eberhard Karls Universität Tübingen

Cai Q, Alam S, Duong V (2020) On robustness paradox in air traffic networks. IEEE Trans Netw Sci Eng 7:3087–3099. https://doi.org/10.1109/TNSE.2020.3015728

Cai K, Li Y, Fang Y-P, Zhu Y (2022) A deep learning approach for flight delay prediction through time-evolving graphs. IEEE Trans Intell Transp Syst 23:11397–11407. https://doi.org/10.1109/TITS.2021.3103502

Cai K, Shen Z, Luo X, Li Y (2023) Temporal attention aware dual-graph convolution network for air traffic flow prediction. J Air Transp Manag 106:102301. https://doi.org/10.1016/j.jairtraman.2022.102301

Cao Y, Sheng QZ, McAuley J, Yao L (2025) Reinforcement learning for generative AI: a survey

Carlini N, Liu C, Erlingsson Ú et al (2019) The secret sharer: evaluating and testing unintended memorization in neural networks

Carvalho L, Sternberg A, Gonçalves LM et al (2021) On the relevance of data science for flight delay research: a systematic review. Transp Rev. https://doi.org/10.1080/01441647.2020.1861123

Çelik AK, Yalçınkaya Ö, Kutlu M (2025) The causal relationship between air transport and economic growth: evidence from top ten countries with the largest air transport volume. Transp Policy 162:521–532. https://doi.org/10.1016/j.tranpol.2025.01.002

Chen J, Cai K, Li W et al (2021) An airspace capacity estimation model based on spatio-temporal graph convolutional networks considering weather impact. In: 2021 IEEE/AIAA 40th digital avionics systems conference (DASC). pp 1–7

Chen N, Man Y, Ning W (2022) Knowledge graph of civil aircraft approach and landing flight safety research based on CiteSpace sustainability analysis. In: 2022 IEEE 4th international conference on civil aviation safety and information technology (ICCASIT). pp 363–369

Chickering DM (2013) Learning equivalence classes of Bayesian networks structures

Choi S, Kim YJ, Briceno S, Mavris D (2016) Prediction of weather-induced airline delays based on machine learning algorithms. In: 2016 IEEE/AIAA 35th digital avionics systems conference (DASC). pp 1–6

Christiano P, Leike J, Brown TB et al (2023) Deep reinforcement learning from human preferences

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Chu C, Zhang H, Wang P, Lu F (2024) Simulating human mobility with a trajectory generation framework based on diffusion model. International Journal of Geographical Information Science 38:847–878. https://doi.org/10.1080/13658816.2024.2312199
Chu-Carroll J, Beck A, Burnham G et al (2024) Beyond LLMs: advancing the landscape of complex reasoning
Cramer KL, O'Dea A, Clark TR et al (2017) Prehistorical and historical declines in Caribbean coral reef accretion rates driven by loss of parrotfish. Nat Commun 8:1–8. https://doi.org/10.1038/ncomms14160
Das BC, Amini MH, Wu Y (2025) Security and privacy challenges of large language models: a survey. ACM Comput Surv 57:152:1-152:39. https://doi.org/10.1145/3712001
DeepSeek-AI, Guo D, Yang D et al (2025) DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning
Du W, Li B, Chen J et al (2023) A spatiotemporal hybrid model for airspace complexity prediction. IEEE Intell Transp Syst Mag 15:217–224. https://doi.org/10.1109/MITS.2022.3204099
Duan W, Jiang L, Wang N, Rao H (2019) Pre-Trained Bidirectional Temporal Representation for Crowd Flows Prediction in Regular Region. IEEE Access 7:143855–143865. https://doi.org/10.1109/ACCESS.2019.2944990
Dück V, Ionescu L, Kliewer N, Suhl L (2012) Increasing stability of crew and aircraft schedules. Transp Res C 20:47–61. https://doi.org/10.1016/j.trc.2011.02.009
Dunbar M, Froyland G, Wu C-L (2012) Robust airline schedule planning: minimizing propagated delay in an integrated routing and crewing framework. Transp Sci 46:204–216. https://doi.org/10.1287/trsc.1110.0395
Elwert F (2014) Book review: causality: models, reasoning, and inference. Acta Sociol 57:369–371. https://doi.org/10.1177/0001699314551683
EUROCONTROL (2009) Base of aircraft data (BADA). https://www.eurocontrol.int/model/bada. Accessed 11 June 2025
EUROCONTROL (2024) EUROCONTROL Data Snapshot #44 on the causes of flight delays. https://www.eurocontrol.int/publication/eurocontrol-data-snapshot-44-causes-flight-delays. Accessed 13 Aug 2025
EUROCONTROL (2025) EUROCONTROL datalink performance and capacity analysis—2024 edition. https://www.eurocontrol.int/publication/eurocontrol-datalink-performance-and-capacity-analysis-2024-edition. Accessed 28 July 2025
FAA (1987) Types of delay—ASPMHelp. https://www.aspm.faa.gov/aspmhelp/index/Types_of_Delay.html?utm_source. Accessed 13 Aug 2025
Feldman V (2021) Does learning require memorization? A short tale about a long tail
Feng D, Hao B, Lai J (2024) Tracing delay network in air transportation combining causal propagation and complex network. Int J Intell Netw 5:63–76. https://doi.org/10.1016/j.ijin.2024.01.006
Fleurquin P, Ramasco JJ, Eguiluz VM (2013) Systemic delay propagation in the US airport network. Sci Rep 3:1159. https://doi.org/10.1038/srep01159
Floridi L, Chiriatti M (2020) GPT-3: its nature, scope, limits, and consequences. Minds Mach 30:681–694. https://doi.org/10.1007/s11023-020-09548-1
Fofanah AJ, Chen D, Wen L, Zhang S (2025) CHAMFormer: dual heterogeneous three-stages coupling and multivariate feature-aware learning network for traffic flow forecasting. Expert Syst Appl 266:126085. https://doi.org/10.1016/j.eswa.2024.126085
Fox KL, Niewoehner KR, Rahmes M et al (2024) Leverage large language models for enhanced aviation safety. In: 2024 integrated communications, navigation and surveillance conference (ICNS). pp 1–11
Franco JL, Neto MVM, Verri FAN, Amancio DR (2025) Graph machine learning for flight delay prediction due to holding manouver. In: arXiv.org. https://arxiv.org/abs/2502.04233v1. Accessed 12 June 2025
Frank MR, Obradovich N, Sun L et al (2018) Detecting reciprocity at a global scale. Sci Adv 4:eaao5348. https://doi.org/10.1126/sciadv.aao5348
Gan Y, Yang Y, Ma Z et al (2024) Navigating the risks: a survey of security, privacy, and ethics threats in LLM-based agents
García-Sigüenza J, Llorens-Largo F, Tortosa L, Vicent JF (2023) Explainability techniques applied to road traffic forecasting using graph neural network models. Inf Sci 645:119320. https://doi.org/10.1016/j.ins.2023.119320
Goater C (2025) IATA: global air passenger demand will hit a record high in 2024. IATA
Gopalakrishnan K, Balakrishnan H (2021) Control and optimization of air traffic networks. Annu Rev Control Robot Auton Syst 4:397–424. https://doi.org/10.1146/annurev-control-070720-080844
Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37:424–438. https://doi.org/10.2307/1912791
Gui G, Liu F, Sun J et al (2020) Flight delay prediction based on aviation big data and machine learning. IEEE Trans Veh Technol 69:140–150. https://doi.org/10.1109/TVT.2019.2954094
Guida M, Maria F (2007) Topology of the Italian airport network: a scale-free small-world network with a fractal structure? Chaos Solitons Fractals 31:527–536. https://doi.org/10.1016/j.chaos.2006.02.007

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Guo Z, Mei G, Liu S et al (2020) SGDAN—a spatio-temporal graph dual-attention neural network for quantified flight delay prediction. Sensors 20:6433. https://doi.org/10.3390/s20226433
Guo Z, Hao M, Yu B, Yao B (2022) Detecting delay propagation in regional air transport systems using convergent cross mapping and complex network theory. Transp Res E 157:102585. https://doi.org/10.1016/j.tre.2021.102585
Hadi MU, Tashi QA, Qureshi R et al (2024) Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects
Haitao L, Qingyao A, Qian D, Yiqun L (2024) Lexilaw: a scalable legal language model for comprehensive legal understanding
Hansen M (2002) Micro-level analysis of airport delay externalities using deterministic queuing models: a case study. J Air Transp Manag 8:73–87. https://doi.org/10.1016/S0969-6997(01)00045-X
Hao L, Hansen M, Zhang Y, Post J (2014) New York, New York: two ways of estimating the delay impact of New York airports. Transp Res E 70:245–260. https://doi.org/10.1016/j.tre.2014.07.004
Hasan U, Hossain E, Gani MO (2023) A survey on causal discovery methods for temporal and non-temporal data
He W, Jiang Z, Xiao T et al (2025) A survey on uncertainty quantification methods for deep learning
Hoyer PO, Janzing D, Mooij J et al (2008) Nonlinear causal discovery with additive noise models. In: Proceedings of the 21st International conference on neural information processing systems. Curran Associates Inc., Red Hook, pp 689–696
Huang Q, Tao M, Zhang C et al (2023) Lawyer LLaMA technical report
Hunter G, Boisvert B, Ramamoorthy K (2007) Advanced national airspace traffic flow management simulation experiments and vlidation. In: 2007 winter simulation conference. pp 1261–1267
Hurter C, Degas A, Guibert A et al (2022) Usage of more transparent and explainable conflict resolution algorithm: air traffic controller feedback. Transp Res Procedia 66:270–278. https://doi.org/10.1016/j.trpro.2022.12.027
IATA (2025) Global air passenger demand reaches record high in 2024. https://www.iata.org/en/pressroom/2025-releases/2025-01-30-01/. Accessed 1 Oct 2025
Jia Z, Cai X, Hu Y et al (2022) Delay propagation network in air transport systems based on refined nonlinear Granger causality. Transp B 10:586–598. https://doi.org/10.1080/21680566.2021.2024102
Jiang W, Luo J (2022) Graph neural network for traffic forecasting: a survey. Expert Syst Appl 207:117921. https://doi.org/10.1016/j.eswa.2022.117921
Jiang J, Zhou K, Dong Z et al (2023) StructGPT: a general framework for large language model to reason over structured data. In: arXiv.org. https://arxiv.org/abs/2305.09645v2. Accessed 20 June 2025
Jin K, Wi J, Lee E et al (2021) TrafficBERT: pre-trained model with large-scale data for long-range traffic flow forecasting. Expert Syst Appl 186:115738. https://doi.org/10.1016/j.eswa.2021.115738
Kaddour J, Lynch A, Liu Q et al (2022) Causal machine learning: a survey and open problems
Kaddour J, Harris J, Mozes M et al (2023) Challenges and applications of large language models
Kafle N, Zou B (2016) Modeling flight delay propagation: a new analytical-econometric approach. Transp Res B 93:520–542. https://doi.org/10.1016/j.trb.2016.08.012
Khanmohammadi S, Tutun S, Kucuk Y (2016) A new multilevel input layer artificial neural network for predicting flight delays at JFK airport. Procedia Comput Sci 95:237–244. https://doi.org/10.1016/j.procs.2016.09.321
Kim YJ, Choi S, Briceno S, Mavris D (2016) A deep learning approach to flight delay prediction. In: 2016 IEEE/AIAA 35th digital avionics systems conference (DASC). pp 1–6
Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks
Kugiumtzis D, Koutlis C, Tsimpiris A, Kimiskidis VK (2017) Dynamics of epileptiform discharges induced by transcranial magnetic stimulation in genetic generalized epilepsy. Int J Neural Syst 27:1750037. https://doi.org/10.1142/S012906571750037X
Künnen J-R, Strauss AK (2022) The value of flexible flight-to-route assignments in pre-tactical air traffic management. Transp Res B 160:76–96. https://doi.org/10.1016/j.trb.2022.04.004
Lai S, Xu Z, Zhang W et al (2023) Large language models as traffic signal control agents: capacity and opportunity
Li Q, Jing R (2021) Characterization of delay propagation in the air traffic network. J Air Transp Manag 94:102075. https://doi.org/10.1016/j.jairtraman.2021.102075
Li Q, Jing R (2022) Flight delay prediction from spatial and temporal perspective. Expert Syst Appl 205:117662. https://doi.org/10.1016/j.eswa.2022.117662
Li Z, Chen H, Ge J, Ning K (2018) An airport scene delay prediction method based on LSTM. In: Gan G, Li B, Li X, Wang S (eds) Advanced data mining and applications. Springer International Publishing, Cham, pp 160–169
Li Q, Guan X, Liu J (2023) A cnn-lstm framework for flight delay prediction. Expert Syst Appl 227:120287. https://doi.org/10.1016/j.eswa.2023.120287

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Li C, Mao J, Li L (2024a) Flight delay propagation modeling: data, methods, and future opportunities. Transp Res E 185:103525. https://doi.org/10.1016/j.tre.2024.103525
Li C, Qi X, Yang Y et al (2024b) FAST-CA: fusion-based adaptive spatial–temporal learning with coupled attention for airport network delay propagation prediction. Inf Fusion 107:102326. https://doi.org/10.1016/j.inffus.2024.102326
Li Y, Cai K, Zhu Y, Yang Y (2024c) Modeling delay propagation in airport networks via causal biased random walk. IEEE Trans Intell Transp Syst 25:4692–4703. https://doi.org/10.1109/TITS.2023.3321398
Liu C, Yang S, Xu Q et al (2024) Spatial-temporal large language model for traffic prediction
Long S, Tan J, Mao B et al (2025) A survey on intelligent network operations and performance optimization based on large language models. IEEE Commun Surv Tutor. https://doi.org/10.1109/COMST.2025.3526606
Mahmud D, Hajmohamed H, Almentheri S et al (2025) Integrating LLMs with ITS: recent advances, potentials, challenges, and future directions. IEEE Trans Intell Transp Syst. https://doi.org/10.1109/TITS.2025.3528116
Mamdouh M, Ezzat M, Hefny H (2024) Improving flight delays prediction by developing attention-based bidirectional LSTM network. Expert Syst Appl 238:121747. https://doi.org/10.1016/j.eswa.2023.121747
Markovic D, Hauf T, Röhner P, Spehr U (2008) A statistical study of the weather impact on punctuality at Frankfurt Airport. Meteorol Appl 15:293–303. https://doi.org/10.1002/met.74
Mercer J, Gomez A, Gabets C et al (2016) Impact of automation support on the conflict resolution task in a human-in-the-loop air traffic control simulation. IFAC-PapersOnLine 49:36–41. https://doi.org/10.1016/j.ifacol.2016.10.458
Min B, Ross H, Sulem E et al (2023) Recent advances in natural language processing via large pre-trained language models: a survey. ACM Comput Surv 56:30:1–30:40. https://doi.org/10.1145/3605943
MTPBC (2016) Flight normal management provisions. In: Normal flight management regulations. https://www.gov.cn/gongbao/content/2016/content_5115843.htm. Accessed 1 May 2024
Nassar A, Livathinos N, Lysak M, Staar P (2022) TableFormer: table structure understanding with transformers. In: arXiv.org. https://arxiv.org/abs/2203.01017v2. Accessed 20 June 2025
Neuberg LG (2003) CAUSALITY: MODELS, REASONING, AND INFERENCE, by Judea Pearl, Cambridge University Press, 2000. Econometric Theory 19:675–685. https://doi.org/10.1017/S0266466603004109
NOAA (2005) Ogimet home page. https://www.ogimet.com/index.phtml.en. Accessed 12 Aug 2025
Nuic A, Poles D, Mouillet V (2010) BADA: an advanced aircraft performance model for present and future ATM systems. Int J Adapt Control Signal Process 24:850–866. https://doi.org/10.1002/acs.1176
Oh Y, Kwak J, Kim S (2021) Time delay estimation of traffic congestion propagation due to accidents based on statistical causality. In: arXiv.org. https://arxiv.org/abs/2108.06717v3. Accessed 10 Jan 2024
Pang A, Wang M, Pun M-O et al (2024) iLLM-TSC: integration reinforcement learning and large language model for traffic signal control policy improvement
Papana A, Kyrtsou C, Kugiumtzis D, Diks C (2017) Financial networks based on Granger causality: a case study. Physica A 482:65–73. https://doi.org/10.1016/j.physa.2017.04.046
Pastorino L, Zanin M (2021) Air delay propagation patterns in Europe from 2015 to 2018: an information processing perspective. J Phys Complex 3:015001. https://doi.org/10.1088/2632-072X/ac4003
Pathomsiri S, Haghani A, Dresner M, Windle RJ (2008) Impact of undesirable outputs on the productivity of US airports. Transp Res E 44:235–259. https://doi.org/10.1016/j.tre.2007.07.002
Pearl J (2009) Causality, 2nd edn. Cambridge University Press, Cambridge
Perott A, Schader NT, Leonhardt J, Licu T (2019) White paper human factors integration in ATM system design. EUROCONTROL
Pineda-Jaramillo J, Munoz C, Mesa-Arango R et al (2024) Integrating multiple data sources for improved flight delay prediction using explainable machine learning. Res Transp Bus Manag 56:101161. https://doi.org/10.1016/j.rtbm.2024.101161
Pouyanfar S, Sadiq S, Yan Y et al (2018) A survey on deep learning: algorithms, techniques, and applications. ACM Comput Surv 51:92:1–92:36. https://doi.org/10.1145/3234150
Qi S, Liu Q, Liu C et al (2023) GA2T: traffic flow prediction model combined with graph attention networks. J Comput Aided Des Graph 1–9
Qin QL, Yu H (2014) A statistical analysis on the periodicity of flight delay rate of the airports in the US. Adv Transp Stud. https://doi.org/10.4399/978885487831010
Que Z, Yao H, Yue W (2018) Simulation analysis of the effect of initial delay on flight delay diffusion—IOPscience. In: IOP conference series: earth and environmental science. IOP Science, Boston, pp 1–10
Rebollo JJ, Balakrishnan H (2014) Characterization and prediction of air traffic delays. Transp Res C 44:231–241. https://doi.org/10.1016/j.trc.2014.04.007
Ringbauer M, Giarmatzi C, Chaves R (2016) Experimental test of nonlocal causality. Sci Adv 2:e1600162. https://doi.org/10.1126/sciadv.1600162

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Runge J, Bathiany S, Bollt E et al (2019) Inferring causation from time series in earth system sciences. Nat Commun 10:2553. https://doi.org/10.1038/s41467-019-10105-3
Schaefer L, Millner D (2001) Flight delay propagation analysis with the Detailed Policy Assessment Tool. In: 2001 IEEE international conference on systems, man and cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236), vol 2. pp 1299–1303
Schreiber T (2000) Measuring information transfer. Phys Rev Lett 85:461–464. https://doi.org/10.1103/PhysRevLett.85.461
Shannon CE (1948) A mathematical theory of communication. The Bell System Technical Journal 27:379–423
Shimizu S, Jp IA, Hoyer PO, et al (2006) A linear non-gaussian acyclic model for causal discovery. Journal of Machine Learning Research 7:2003–2030
Shu Y, Zhao J (2013) Data-driven causal inference based on a modified transfer entropy. Comput Chem Eng 57:173–180. https://doi.org/10.1016/j.compchemeng.2013.05.011
Sperduti A, Starita A (1997) Supervised neural networks for the classification of structures. IEEE Trans Neural Netw 8:714–735. https://doi.org/10.1109/72.572108
Spirtes P, Glymour C (1991) An algorithm for fast recovery of sparse causal graphs. Soc Sci Comput Rev 9:62–72. https://doi.org/10.1177/089443939100900106
Spirtes P, Glymour C, Scheines R (2001) Causation, prediction, and search. The MIT Press
Sugihara G, May R, Ye H (2012) Detecting causality in complex ecosystems. Science 338:496–500. https://doi.org/10.1126/science.1227079
Sui Y, Zhou M, Zhou M et al (2024) Table meets LLM: can large language models understand structured table data? A benchmark and empirical study. In: Proceedings of the 17th ACM international conference on web search and data mining. Association for Computing Machinery, New York, pp 645–654
Sun X, Wandelt S, Zhang A (2020) How did COVID-19 impact air transportation? A first peek through the lens of complex networks. J Air Transp Manag 89:101928. https://doi.org/10.1016/j.jairtraman.2020.101928
Sun J, Dijkstra T, Aristodemou C et al (2022) Designing recurrent and graph neural networks to predict airport and air traffic network delays. 1–8
Sun M, Tian Y, Wang X (2024) Transport causality knowledge-guided GCN for propagated delay prediction in airport delay propagation networks. Expert Syst Appl 240:122426. https://doi.org/10.1016/j.eswa.2023.122426
Tabrizian A, Gupta P, Taye A et al (2024) Using large language models to automate flight planning under wind hazards. In: 2024 AIAA DATC/IEEE 43rd digital avionics systems conference (DASC). IEEE, San Diego, pp 1–8
Tan X, Liu Y, Liu D (2022) An attention-based deep convolution network for mining airport delay propagation causality. Appl Sci 12:10433. https://doi.org/10.3390/app122010433
Tang Z, Huang S, Han S (2021) Recent Progress about Flight Delay under Complex Network. Complexity 2021:1–18. https://doi.org/10.1155/2021/5513093
U.S. Department of Transportation (2024) Flight delays. Office of Aviation Consumer Protection, Washington, DC.
Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In: Advances in neural information processing systems. Curran Associates, Inc., pp 5998–6008
Wandelt S, Chen X, Sun X (2025) Flight delay prediction: a dissecting review of recent studies using machine learning. IEEE Trans Intell Transp Syst 26:4283–4297. https://doi.org/10.1109/TITS.2025.3528536
Wang T, Chen S-C (2022) Multi-task local-global graph network for flight delay prediction. In: 2022 IEEE 23rd international conference on information reuse and integration for data science (IRI). pp 49–54
Wang PTR, Schaefer LA, Wojcik LA (2003) Flight connections and their impacts on delay propagation. In: Digital avionics systems conference, 2003. DASC '03, vol 1. The 22nd. p 5.B.4-5.1-9
Wang C, Hu M, Yang L, Zhao Z (2022a) Overview of research on air traffic delay prediction. Syst Eng Electron 3:863–874
Wang F, Bi J, Xie D, Zhao X (2022b) Flight delay forecasting and analysis of direct and indirect factors. IET Intell Transp Syst 16:890–907. https://doi.org/10.1049/itr2.12183
Wang T, Chen S-C (2022c) Multi-task local-global graph network for flight delay prediction. In: 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI). pp 49–54
Wang D, Wang X, Chen L et al (2023a) TransWorldNG: traffic simulation via foundation model
Wang H, Liu C, Xi N et al (2023b) HuaTuo: tuning LLaMA model with Chinese medical knowledge
Wang L, Ren Y, Jiang H et al (2023c) AccidentGPT: accident analysis and prevention from V2X environmental perception with multi-modal large model
Wang X, Wang D, Chen L, Lin Y (2023d) Building transportation foundation model via generative graph transformer

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Wieland F (1997) Limits to growth: results from the detailed policy assessment tool [air traffic congestion]. In: 16th DASC. AIAA/IEEE digital avionics systems conference. Reflections to the future. Proceedings. p 9.2–1
Woodburn A, Ryerson M (2014) Airport capacity enhancement and flight predictability. Transp Res Rec J Transp Res Board 2400:87–97. https://doi.org/10.3141/2400-10
Wu C-L (2005) Inherent delays and operational reliability of airline schedules. J Air Transp Manag 11:273–282. https://doi.org/10.1016/j.jairtraman.2005.01.005
Wu W (2016) Flight plan optimization based on airport delay prediction. J Transp Syst Eng Inf Technol 16:189
Wu C-L, Caves RE (2002) Modelling of aircraft rotation in a multiple airport environment. Transp Res E 38:265–277. https://doi.org/10.1016/S1366-5545(02)00010-8
Wu C-L, Maher SJ (2018) Airline capacity planning and management. In: The Routledge companion to air transport management. Routledge
Wu W, Wu C-L (2018) Enhanced delay propagation tree model with Bayesian network for modelling flight delay propagation. Transp Plann Technol 41:319–335. https://doi.org/10.1080/03081060.2018.1435453
Wu Z, Pan S, Chen F et al (2021) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32:4–24. https://doi.org/10.1109/TNNLS.2020.2978386
Wu S, Irsoy O, Lu S et al (2023) BloombergGPT: a large language model for finance
Wu Y, Yang H, Lin Y, Liu H (2024) Spatiotemporal propagation learning for network-wide flight delay prediction. IEEE Trans Knowl Data Eng 36:386–400. https://doi.org/10.1109/TKDE.2023.3286690
Xiao Y, Zhao Y, Wu G, Jing Y (2020) Study on delay propagation relations among airports based on transfer entropy. IEEE Access 8:97103–97113. https://doi.org/10.1109/ACCESS.2020.2996301
Xiong J, Hansen M (2013) Modelling airline flight cancellation decisions. Transp Res E 56:64–80. https://doi.org/10.1016/j.tre.2013.05.003
Xiong H, Sheng W, Yitao Z, Zihao Z (2024) DoctorGLM: fine-tuning your Chinese doctor is not a herculean task. In: ResearchGate. https://www.researchgate.net/publication/369760637. Accessed 5 Mar 2025
Xu Q, Pang Y, Liu Y (2023) Air traffic density prediction using Bayesian ensemble graph attention network (BEGAN). Transp Res C 153:104225. https://doi.org/10.1016/j.trc.2023.104225
Xu Q, Pang Y, Zhou X, Liu Y (2024) PIGAT: physics-informed graph attention transformer for air traffic state prediction. IEEE Trans Intell Transp Syst 25:12561–12577. https://doi.org/10.1109/TITS.2024.3386128
Xue H, Voutharoja BP, Salim FD (2022) Leveraging language foundation models for human mobility forecasting. In: Proceedings of the 30th international conference on advances in geographic information systems. Association for Computing Machinery, New York, pp 1–9
Yang F, Sirish LS, Xiao D (2010) Signed Directed Graph modeling of industrial processes and their validation by data-based methods. In: 2010 Conference on Control and Fault-Tolerant Systems (SysTol). pp 387–392
Yang Z, Chen Y, Hu J et al (2023) Departure delay prediction and analysis based on node sequence data of ground support services for transit flights. Transp Res C 153:104217. https://doi.org/10.1016/j.trc.2023.104217
Yeh C-K, Ravikumar P (2021) Objective criteria for explanations of machine learning models. Appl AI Lett 2:e57. https://doi.org/10.1002/ail2.57
Zanin M (2021) Simplifying functional network representation and interpretation through causality clustering. Sci Rep 11:15378. https://doi.org/10.1038/s41598-021-94797-y
Zanin M, Belkoura S, Zhu Y (2017) Network analysis of Chinese air transport delay propagation. Chin J Aeronaut 30:491–499. https://doi.org/10.1016/j.cja.2017.01.012
Zeng L, Wang B, Wang T, Wang Z (2022) Research on delay propagation mechanism of air traffic control system based on causal inference. Transp Res C 138:103622. https://doi.org/10.1016/j.trc.2022.103622
Zhang M, Zhou X, Zhang Y (2019) Propagation index on airport delays. Transp Res Rec J Transp Res Board 2673:536–543. https://doi.org/10.1177/0361198119844240
Zhang L, Yang H, Wu X (2023a) Air traffic complexity evaluation with hierarchical graph representation learning. Aerospace 10:352. https://doi.org/10.3390/aerospace10040352
Zhang S, Fu D, Zhang Z et al (2023b) TrafficGPT: viewing, processing and interacting with traffic foundation models. In: arXiv.org. https://arxiv.org/abs/2309.06719v1. Accessed 11 Jan 2024
Zhang X, Yang Q, Xu D (2023c) XuanYuan 2.0: a large Chinese financial chat model with hundreds of billions parameters
Zhang D, Zheng H, Yue W, Wang X (2024) Advancing ITS applications with LLMs: a survey on traffic management, transportation safety, and autonomous driving. In: Hu M, Cornelis C, Zhang Y (eds) Rough sets. Springer Nature Switzerland, Cham, pp 295–309
Zheng Y, Yan R, Jia B et al (2023) Adaptive Kalman-based hybrid car following strategy using TD3 and CACC. In: arXiv.org. https://arxiv.org/abs/2312.15993v1. Accessed 28 Dec 2023

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Zheng H, Wang Z, Zheng C et al (2024) A graph multi-attention network for predicting airport delays. Transp Res E 181:103375. https://doi.org/10.1016/j.tre.2023.103375

Zhong Q, Yu Y, Huang Y, Zhang T (2025) Prediction and optimization of civil aviation flight delays based on machine learning algorithms. Int J Comput Intell Syst 18:189. https://doi.org/10.1007/s44196-025-00932-2

Zhou F, Jiang G, Lu Z, Wang Q (2022a) Evaluation and analysis of the impact of airport delays. Sci Program 2022:e7102267. https://doi.org/10.1155/2022/7102267

Zhou Z, Yang Z, Zhang Y et al (2022b) A comprehensive study of speed prediction in transportation system: from vehicle to traffic. iScience 25:103909. https://doi.org/10.1016/j.isci.2022.103909

Zohrevandi E, Westin L, Lundberg J, Ynnerman A (2022) Design and evaluation study of visual analytics decision support tools in air traffic control. Comput Graph Forum 41:230–242. https://doi.org/10.1111/cgf.14431

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Mengyuan Sun[1,2] · Yong Tian[1] · Jiangchen Li[1,3] · Cheng-Lung Wu[2] · Liqun Peng[4] · Shucai Xu[5]

✉ Jiangchen Li
jiangchen@nuaa.edu.cn

Mengyuan Sun
smengyuan@nuaa.edu.cn

Yong Tian
tianyong@nuaa.edu.cn

Cheng-Lung Wu
c.l.wu@unsw.edu.au

Liqun Peng
liqun@ualberta.ca

Shucai Xu
xushc@mail.tsinghua.edu.cn

[1] Department of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

[2] School of Aviation, The University of New South Wales, Kensington, NSW 2052, Australia

[3] School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China

[4] Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB, Canada

[5] State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University, Beijing 100084, China

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com