Commun Transp Res

Review

Review of volume-delay functions integrating traditional models and emerging AI technologies

Yuyan (Annie) Pan[1], Xianbiao (XB) Hu[1,✉], George List[2], Xuesong (Simon) Zhou[3,✉]

Cite this article: Pan Y Y, Hu X B, List G, et al. *Commun Transp Res* 2026, **6**(1): 9640003. https://doi.org/10.26599/COMMTR.2026.9640003

ABSTRACT: Link performance functions (LPFs) underpin the estimation of travel times, delays, and congestion in transportation network models. As multimodal systems become increasingly data rich and feature connected and automated vehicles (CAVs), electric vehicles (EVs), and complex queuing phenomena, traditional LPFs face limitations in accuracy, physical consistency, and real-time adaptability. This review synthesizes the evolution of LPFs to address these modern challenges and bridge theoretical foundations, diverse data sources, and artificial intelligence (AI) techniques. We conduct a systematic literature survey of over 270 peer-reviewed studies, covering (i) classical empirical forms (e.g., the US Bureau of Public Roads (BPR), Davidson), (ii) theory-driven models (fundamental-diagram and fluid-queue approaches), and (iii) AI-augmented frameworks (long short-term memory (LSTM), graph neural network (GNN), transformer, and physics-informed learning). Models are classified by their mathematical assumptions, data requirements, integration with dynamic traffic assignment tools, and support for multimodal flow interactions. We further evaluate each class against emerging criteria: Queuing dynamics, CAV/EV impacts, data fusion complexity, and computational tractability, with the following: (1) We highlight the emergence of hybrid physics-machine learning (ML) models that enforce conservation laws while leveraging large-scale probe and sensor data. (2) We identify critical gaps in AI-related method evaluation, with many studies replying on point-error metrics rather than system-level key performance indicators (KPIs) (e.g., network delay, throughput, queue length, bottleneck durations). (3) We demonstrate that physics-informed neural networks (PINNs) achieve superior predictive accuracy in travel time estimation while maintaining physical consistency, outperforming both traditional LPFs and other data-driven approaches. Overall, by bridging theory, data, and AI, this review maps the LPF landscape, highlights open research directions (e.g., fully differentiable LPFs, uncertainty quantification, real-time adaptation), and provides a roadmap for robust and transparent models for next-generation multimodal transportation networks.

KEYWORDS: link performance functions; volume-delay functions; physically informed; fluid-queue-based; emerging transportation applications; multimodal systems; artificial intelligence (AI)

1 Introduction

Link performance functions (LPFs) have long served as a cornerstone of traffic assignment, network design, and policy evaluation. LPFs such as the US Bureau of Public Roads (BPR) function have traditionally provided simple yet tractable formulations that link volume to delay. However, with the proliferation of connected and automated vehicles (CAVs), the emergence of high-resolution sensing, and the increasing availability of multimodal, multisource traffic data, the limitations of conventional LPFs have become increasingly evident. Existing methods often fail to capture dynamic congestion phenomena, lack physical consistency with queuing processes, and exhibit poor generalizability across regimes and networks.

The rapid advancements in artificial intelligence (AI) and machine learning (ML) have offered new opportunities to address these shortcomings. Models such as long short-term memory (LSTM) networks, graph neural networks (GNNs), and transformers can learn complex nonlinear patterns and exploit spatiotemporal dependencies in modern transportation systems. However, purely data-driven models remain limited: They risk violating conservation principles, offer little interpretability, and may not be generalizable under distribution shifts. This tension highlights the need for hybrid approaches that integrate domain knowledge from traffic flow theory with the predictive power of AI.

Against this backdrop, the unique contribution of this review lies not only in synthesizing prior research but also, equally importantly, in proposing an original future framework and a unified pipeline for LPF modeling in the AI era. Specifically, we revisit LPFs through the lens of multiresolution traffic flow theory, clarify technical misconceptions, and introduce reproducible calibration procedures such as demand-to-capacity (D/C) ratio estimation. Building on these insights, we develop a physics-informed, AI-driven pipeline and articulate a prioritized roadmap

[1] Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park 16802-1408, USA. [2] Department of Civil, Construction and Environmental Engineering, North Carolina State University, Raleigh 27606, USA. [3] School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe 85281, USA.

✉ Corresponding authors. E-mail: X. Hu, xbhu@psu.edu; X. Zhou, xzhou74@asu.edu

Received: July 24, 2025; Revised: October 18, 2025; Accepted: October 23, 2025

for future research. To our knowledge, this is the first attempt to consolidate the evolution of LPFs into a forward-looking, technically grounded framework that is both comprehensive and actionable.

In recent years, the rapid rise of large-language models and AI-enabled traffic forecasting has spawned numerous review articles in ML communities. However, on the ground, transportation planners largely continue to rely on the classical four-step paradigm, which includes trip generation, distribution, mode choice, and assignment (Bliemer et al., 2017; Patriksson, 2015; Roughgarden and Tardos, 2002), even as many agencies experiment with activity-based, dynamic self-assignment, or simulation-based extensions. These methods often aim to improve link flow or density predictions via ML, but they lack robust, system-level key performance indicators (KPIs), such as network-wide delay, throughput, or multimodal performance, which are needed to evaluate large-scale investment and operational policies effectively.

At the heart of every assignment model lies the LPFs, also called a volume delay, travel time, or cost function, which relates flow (or demand) to link travel time (or delay). The purpose of a link performance function in traffic assignment is to model the relationship between traffic flow and travel time on a particular link in a transportation network. LPFs have many different names in the literature; they are also called: (1) link travel time functions (Beckmann et al., 1956; Fisk, 1991; Xiong and Davis, 2009), (2) link capacity functions (Branston, 1976; Kosun et al., 2016), (3) link performance functions (Patriksson, 2015; Ran et al., 1996; Sheffi, 1984), (4) link cost functions (Small et al., 2007) with related marginal cost functions (Fosgerau and Small, 2012; Van and Fosgerau, 2009), and (5) edge latency functions in computer science and game theory (Lianeas et al., 2018; Roughgarden and Tardos, 2002).

The early formulation of the LPFs can also be traced back to work by Duffin (1947) on electrical networks with a certain nonlinear resistance function. The classical work of Beckmann et al. (1956) (Beckmann, McGuire and Winsten (BMW)) produced mathematical programming models that formulate the static equilibrium traffic assignment problem and adopt a generic nonlinear LPF. Charnes et al. (1958) and Charnes and Cooper (1959) independently proposed formulations for traffic assignment problems with fixed origin-destination (OD) flows and piecewise LPFs. While many studies have been conducted on Static Traffic Assignments (STAs) with hard capacity constraints (Bliemer and Raadsen, 2020; Larsson and Patriksson, 1995; Nie et al., 2004; Raadsen et al., 2020), representing and calibrating an LPF for oversaturated freeway facilities remains a significant challenge. This challenge has motivated another line of research, dynamic traffic assignment, which aims to capture queue formation, propagation, and dissipation in congested networks (Brederode et al., 2019; Peeta and Ziliaskopoulos, 2001). Equally important for LPFs are examining the inherent theoretical connections between macroscopic LPFs and queue-theoretic models and mesoscopic and microscopic simulation models, such as POLARIS (Auld et al., 2016); MATSim (Balmer et al., 2008); Dynus-T (Chiu et al., 2010); SUMO (Behrisch et al., 2011); TRANSIMS (Nagel et al., 1999); AIMSUN (Barceló and Casas, 2005); VISSIM (Fellendorf and Vortisch, 2010); TransModeler (Caliper, 2009); and DynaMIT (Ben-Akiva et al., 1998). Insights from a cross-resolution perspective can contribute to achieving a higher level of consistency and accuracy in modeling complex demand and supply interactions, and the importance and development of LPFs has been highlighted in surveys by Branston (1976) and Saric et al. (2018).

This review aims to fill that gap by guiding both planners and AI researchers through a concise yet systematic journey:

1) Core LPF concepts (capacity, demand, and queue dynamics);

2) Evolution of formulations from simple linear and BPR curves to theory-driven and hybrid AI approaches;

3) Extensions to multimodal and cross-resolution applications;

4) Integration with AI frameworks for real-time adaptation, physical consistency, and large-scale KPI design.

By highlighting how even a basic linear LPF connects to fundamental diagrams (FDs), traffic-state estimation, and multimodal interactions and by showing where AI can be woven in, we aim to equip practitioners with the conceptual tools and evaluation metrics necessary to leverage AI without abandoning the rigor and transparency of established traffic-modeling practices.

This review progresses systematically from foundational concepts to emerging applications. Section 2 traces the theoretical foundations and historical evolution of LPFs across four phases (1950s–present). Section 3 introduces four core insights addressing key challenges: Distinguishing true demand from measured flow, achieving multiscale consistency, emphasizing congestion duration as a key delay factor, and enabling cross-resolution modeling for future technologies. Section 4 examines the shift toward physics-informed AI, identifies critical integration challenges, and presents a unified end-to-end AI-driven LPF framework that integrates data-driven approaches with physics-based constraints to balance theoretical consistency with cross-scenario generalization. Section 5 presents a case study that systematically compares empirical Volume Delay Functions (VDFs), theory-driven approaches, data-driven models, and physics-informed methods, covering the full process from case selection to model comparison and result analysis to provide a comprehensive and unified evaluation across paradigms. Section 6 outlines nine priority questions that guide future research, synthesizing findings and offering strategic recommendations for transitioning to next-generation physics-informed AI systems.

2 Fundamentals, evolution, and contemporary challenges of link performance functions

2.1 Bibliometric analysis

We conducted a systematic literature search across Web of Science, Scopus, Transport Research International Documentation (TRID), and IEEE Xplore, using keywords such as "link performance function", "volume-delay function", "fundamental diagram", and "queue theory". Each study was then coded according to four key dimensions: (1) its theoretical foundation (FD-based, queue-based, simulation-derived, or data-driven), (2) its temporal scope (static, quasi-dynamic, or fully dynamic), (3) its spatial scale (link-level, corridor, or network-wide), and (4) its application domain (highway, urban arterial, multimodal, or emerging mobility).

A bibliometric timeline highlights the field's evolution: During the 1950s–1970s, foundational piecewise-linear and BPR functions were established; the 1980s–1990s saw the integration of queue theory and advanced polynomial forms; the 2000s–2010s brought dynamic extensions and hybrid simulation frameworks; and from 2015 onward, machine-learning, AI, and physics-informed approaches have come to the forefront.

These papers are from more than 50 journals, and all these journals belong to "transportation". Two of the top journals with the highest volume of publications are *Transportation Research*

Part B: Methodological and *Transportation Research Record*. Articles published in *Transportation Research Part B: Methodological* account for 18.9% of the total number of publications, and *Transportation Research Record* accounts for 19.4% of the publications.

We analyzed approximately 270 references, which are displayed in Fig. 1. We organized these references chronologically, and Fig. 1a provides a good indication of the publication trend in this field. Our analysis revealed an increase in related publications after 2010. LPFs remain a fundamental research topic in the transportation modeling field and are expected to evolve in the coming years. In addition, we grouped the references into four thematic categories, as shown in Fig. 1b: “Theoretical foundations of LPFs”, “Calibration and validation techniques”, “Multi-Resolution modeling and consistency”, and “Emerging applications and future directions”. This categorization highlights the interconnections between different approaches and their practical applications, providing readers with a more comprehensive understanding of the matter of the LPFs. In the subsequent sections 3, 4, 5, we provide a detailed survey of the theoretical underpinnings, practical considerations, and applications of each category, offering a more insightful analysis. Additionally, we observed that the standard form of LPF has been widely utilized in highly cited studies for numerous emerging applications, such as road pricing, tradable mobility credits, and autonomous vehicle management.

We summarized the most frequently cited papers during the past few decades, which are shown in Table 1. Notably, the original report of the BPR function by the US Bureau of Public Roads has approximately 284 direct citations, whereas the majority of papers (approximately 3184) only use the abbreviation “BPR function” without citing the original reference. The highly cited papers that directly cite the BPR function are shown in Table 2.

Link performance functions trace their origins to FD-based traffic flow theory, which relates flow, density, and speed to characterize free-flow, transitional, and congested regimes. By analytically mapping these regimes, FD models enable explicit delay formulations under varying demand and capacity conditions (e.g., Kucharski and Drabicki, 2017; Moses et al., 2013). Subsequent applications have leveraged this approach to study bottleneck formation and breakdown phenomena in urban networks (Geroliminis and Sun, 2011; Huntsinger and Rouphail, 2011).

2.2 Historical evolution of link performance functions

1) Early linear and piecewise linear forms

In the 1950s, practitioners adopted simple linear approximations for tractability. The Chicago Area Transportation Study introduced a two-segment piecewise-linear LPF based on the volume-capacity ratio (V/C) (Campbell, 1968). The detailed notation and symbol definitions are provided in Appendix A.

$$tt = \begin{cases} t_{\mathrm{f}}, \dfrac{V}{C} \leq 0.6 \\ t_{\mathrm{f}} + \alpha\left(\dfrac{V}{C} - 0.6\right), \ \text{else} \end{cases} \tag{1}$$

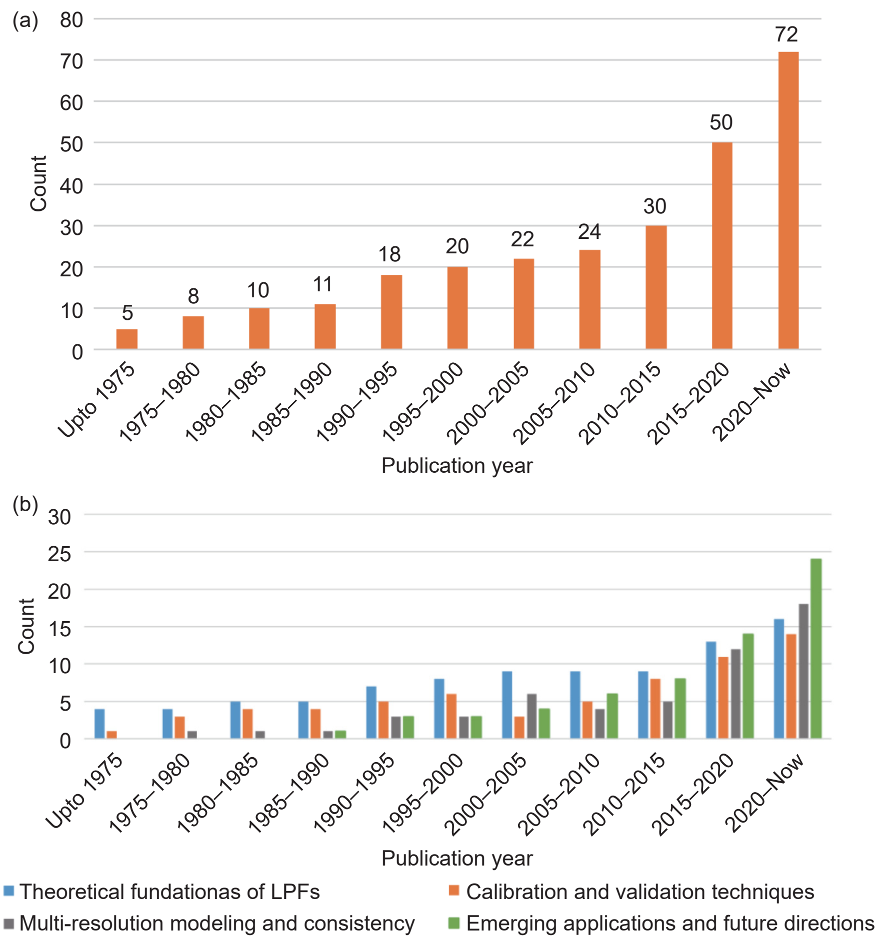


Fig. 1 Analysis of literature publications related to LPFs: (a) temporal distribution of the total publications and (b) temporal distribution of publications by topic group.

Table 1 Frequently cited papers on the topic of VDFs

Rank	Journal title	Total citation	Paper title	Ref.
1	*US Bureau of Public Roads*	3184 (284)	Traffic assignment manual for application with a large, high-speed computer	US Bureau of Public Roads (1964)
2	*Transportation Research Board*	2805	Traffic signal settings	Webster (1958)
3	*Transportation Research Part B: Methodological*	485	Comparison of delay estimates at undersaturated and oversaturated pretimed signalized intersections	Dion et al. (2004)
4	*Transportation Research Part A: Policy and Practice*	472	Traffic assignment and signal control in saturated road networks	Yang and Yagar (1995)
5	*Transportation Research*	444	Link capacity functions: a review	Branston (1976)
6	*Transportation Science*	396	Conical volume-delay functions	Spiess (1990)
7	*Australian Road Research Board*	360	Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and alternative travel time function	Akçelik (1991)
8	*Australian Road Research Board*	360	A flow travel time relationship for use in transportation planning	Davidson (1966)
9	*Transportation Science*	356	Traffic queues and delays at road junctions	Kimber and Hollis (1979)
10	*Journal of Urban Economics*	349	The incidence of congestion tolls on urban highways	Small (1983)
11	*ITE journal*	245	The highway capacity manual delay formula for signalized intersections	Akçelik (1988)
12	*Australian Road Research Board*	205	Time-dependent expressions for delay, stop rate and queue length at traffic signals	Akçelik (1980)
13	*Transportation Research Record*	158	Signalized intersection delay models-a primer for the uninitiated	Hurdle (1984)
14	*Transportation Research Part B: Methodological*	151	Estimation of delays at traffic signals for variable demand conditions	Akçelik and Rouphail (1993)
15	*Highway Research Board Bulletin*	125	An iterative assignment approach to capacity restraint on arterial networks	Smock (1962)

Table 2 Highly cited papers that directly cite the original BPR paper

Rank	Journal title	Total citation	Paper title	Ref.
1	*Mathematics of Operations Research*	574	Selfish routing in capacitated networks	Correa et al. (2004)
2	*Transportation Science*	427	The convergence of equilibrium algorithms with predetermined step sizes	Powell and Sheffi (1982)
3	*Journal of Transportation Engineering*	321	Pedestrian speed/flow relationships for walking facilities in Hong Kong, China	Lam and Cheung (2000)
4	*Transportation Science*	185	Shelter location and evacuation route assignment under uncertainty: A benders decomposition approach	Bayram and Yaman (2018)
5	*Transportation*	183	A study of the bidirectional pedestrian flow characteristics at Hong Kong, China signalized crosswalk facilities	Lam et al. (2002)
6	*International Transactions in Operational Research*	123	Fragile networks: identifying vulnerabilities and synergies in an uncertain age	Nagurney and Qiang (2012)
7	*Mathematics of Operations Research*	121	On the existence of pure Nash equilibria in weighted congestion games	Harks and Klimm (2012)
8	*Transportation Research Record*	115	Flow breakdown and travel time reliability	Dong and Mahmassani (2009)
9	*Transportation Science*	108	Wardrop equilibria with risk-averse users	Ordóñez and Stier-Moses (2010)
10	*Transportation Research Part B: Methodological*	100	A bilevel model of the relationship between transport and residential location	Chang and Mackett (2006)
11	*Journal of Intelligent Transportation Systems*	95	An evaluation of environmental benefits of time-dependent green routing in the greater Buffalo Niagara region	Guo et al. (2013)
12	*Networks and Spatial Economics*	94	Hybrid evolutionary metaheuristics for concurrent multiobjective design of urban road and public transit networks	Miandoabchi et al. (2012)

where tt is the average link travel time, t_f is the free-flow travel time, V is the hourly volume, C is the hourly capacity, and α is a parameter.

This linear form was also proposed earlier, although it is not based on the V/C ratio but instead expressed as the difference between volume and capacity $(V-C)$. The formulation was later extended to include a third straight line that represents the oversaturated region, as illustrated as Eq. (2):

$$tt = \begin{cases} t_p + \alpha\left(V - C_p\right), V < C_p \\ t_p + \beta\left(V - C_p\right), C_p \leq V \leq C_u \\ t_u + \gamma\left(V - C_u\right), V > C_u \end{cases} \tag{2}$$

where $t_u = t_p + \beta\left(C_u - C_p\right)$. t_p is the base (or free-flow) travel time, corresponding to low-volume conditions; t_u is the travel time at the end of the capacity interval; C_p has a lower critical capacity (onset of congestion threshold); below this threshold, the travel time increases only slightly with volume. C_u is the upper capacity threshold; beyond this threshold, traffic enters the oversaturated region with severe queuing. α, β, γ are parameters.

While the linear form is simple and easy to apply, it is unfortunately insufficient for capturing the curve-like patterns observed in field data, as travel time does not increase linearly with traffic volume.

2) Exponential and logarithmic form

To capture nonlinear congestion onset, exponential and logarithmic LPFs emerged in the 1960s. Smock (1962) fitted a natural exponential of V/C in Detroit, and Soltman (1966) used a base-two exponential for Pittsburgh, with a generalized exponential form as

$$tt = t_f \cdot \alpha^{(V/C)^{\beta}} \tag{3}$$

A logarithmic form was also developed, for example, by Mosher (1963), and expressed as $tt = t_f + \ln(\alpha) - \ln(\alpha - V)$, where α represents the saturation flow of a link.

$$tt = \begin{cases} t_f + \beta \ln(\alpha) - \beta \ln(\alpha - V), V \leq C \\ t_f + \beta \ln(\alpha) - \beta \ln(\alpha - C) + \dfrac{\beta V}{\alpha - C}, \text{ else} \end{cases} \tag{4}$$

3) BPR function and polynomial-based form

The BPR function, in its 1964 Traffic Assignment Manual (US BPR, 1964), introduced one of the most well-known and widely adopted VDFs:

$$tt = t_f \cdot \left[1 + \alpha \left(\frac{V}{C}\right)^{\beta}\right] \tag{5}$$

where parameter α represents the ratio of travel time per unit distance at practical capacity to that at free flow, whereas β controls how sharply the curve rises from the free-flow travel time. For V/C ratios less than 1.0, the increase in travel time is gradual, but it accelerates rapidly once the V/C ratio exceeds 1. Higher values of β intensify the onset of congestion effects.

4) Advanced polynomial forms

Several later studies identified shortcomings in the BPR function (Dowling et al., 1998; Skabardonis and Dowling, 1997; Spiess, 1990). The calibrated function may have a nonnegligible error for urban streets because of their unique characteristics, such as intersections, bus transit lanes, stop signs, and signals. To address these issues, several VDFs in polynomial form have also been proposed. For example, Spiess (1990) proposed the conical VDF, expressed as $tt = t_f \cdot \left[2 + \sqrt{\beta^2\left(1 - \frac{V}{C}\right)^2 + \alpha^2} - \beta\left(1 - \frac{V}{C}\right) - \alpha\right]$. In this formulation, α is directly related to β, defined as $\alpha = \dfrac{2\beta - 1}{2\beta - 2}$, with the condition that $\beta > 1$.

Additionally, Akçelik (1991) introduced another volume-delay function that employs a time-dependent form derived from a coordinate transformation technique. It is given by $tt = 0.25t_f \cdot \left[\frac{V}{C} - 1 + \sqrt{\left(\frac{V}{C} - 1\right)^2 + 8\frac{J_A V}{C^2 T_f}}\right]$, where T_f is the flow period (the analysis period, e.g., 15 min, 1 h), and J_A is a delay parameter (accounting for queueing and coordination effects).

5) System-oriented extensions

A truly system-oriented LPF framework must go beyond isolated volume–delay curves to account for interactions with land use, travel behavior, and environmental objectives. Empirical studies have demonstrated that surrounding land-use patterns fundamentally alter link sensitivities (Ma and Lo, 2012; Szeto et al., 2015), whereas heterogeneity in traveler value-of-time and route preferences can reshape the effective delay experienced on a link. Moreover, considerations of environmental factors such as emissions and fuel consumption are increasingly integrated into LPF design to promote sustainable network management (Bagherian et al., 2017; Benedek and Rilett, 1998).

In practice, agencies routinely recalibrate or extend the canonical BPR function to capture local signal timing, intersection density, and transit priority effects (Dowling and Skabardonis, 1993). For dynamically oversaturated corridors, Tisato (1991) modified Davidson's VDF to embed a congestion duration term, whereas Boyce et al. (1981) warned that asymptotic growth in LPFs can produce counterintuitive detours by over penalizing highly congested links. More recently, Zhou et al. (2022) revisited Spiess's well-behaved congestion criteria and infused them with queuing-theoretic constraints, thereby improving LPF realism under sustained oversaturation.

6) Link performance functions for heterogeneous traffic flows

The presence of heterogeneous traffic flows composed of various vehicle classes, such as cars, trucks, buses, and trains, introduces critical complexity to LPFs by affecting average speeds, stopping behaviors, and capacity distributions (Montanino et al., 2021; Qian et al., 2017). Early LPFs that accounted for truck impacts include Kim and Mahmassani's (1987) truck-volume delay model and Müller and Schiller's (2015) German-motorway extension. Mesbah et al. (2011) adapted the Akçelik cost function for mixed car-transit priority, whereas Chen et al. (2011) introduced a time-dependent queuing model for stochastic truck service at gates. In rail applications, Petersen (1974) and Prokopy and Richard (1975) developed multiclass train delay estimators; Crainic et al. (1990) proposed polynomial LPFs for yards and lines; Lai and Barkan (2009) fitted exponential delay-volume curves for single-track networks; and Sogin et al. (2016) quantified the capacity gains of double-tracking via nonlinear delay functions. On highways, Zhang and Waller (2018) modified the BPR formula for high occupancy vehicle (HOV) lanes versus general purpose (GP) lanes, Hörcher et al. (2017) modeled crowding-dependent train times, and Anupriya et al. (2023) leveraged station-level fundamental diagrams to capture passenger boarding–alighting dynamics. Collectively, these studies show that explicitly embedding multivehicle-type considerations into LPFs produces more accurate, context-sensitive delay models essential for both road and rail network analysis. A summary of the key VDFs and related publications is provided in Appendix B.

2.3 Core misconceptions and technical challenges in link performance function studies

Drawing on our comprehensive review, Fig. 2 synthesizes the full demand-capacity mapping across different regimes, and this end-to-end perspective dispels three persistent misconceptions about LPFs. It traces the demand–capacity (D/C) ratio through three linked regimes: Regime A (uncongested): Speed decreases gradually as flow increases toward capacity; Regime B (congested): Beyond the capacity peak, queue formation makes the speed-flow relationship non-monotonic; Regime C (oversaturated outflow): Observed outflow caps at the link's service rate μ even as demand increases.

Misconception 1: LPFs belong only to the regional-scale, static traffic assignment.

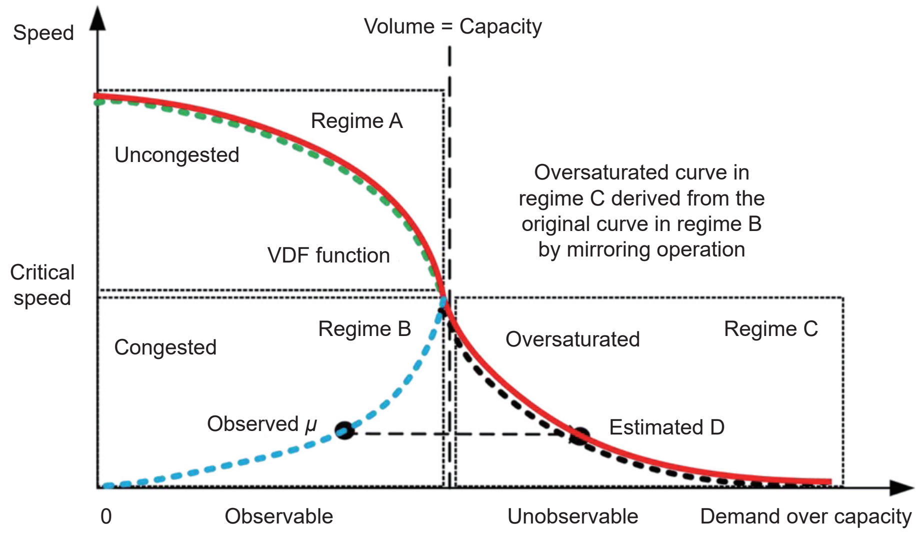


Fig. 2 Fundamental diagram-derived speed-flow relationships (Greenshields), illustrating Regime A (uncongested), Regime B (congested), and Regime C (oversaturated/unobservable demand). True demand *D* lies beyond the measured outflow and must be imputed.

1) Common belief: As the final component of the four-step process, LPFs are often calibrated only once annual or peak-hour traffic counts are used and are applicable solely to individual corridors.

2) Why it is misleading and how to address it: LPFs are grounded in fundamental diagrams and queue-theoretic principles across micro- to macro-scope scales. By calibrating across shoulders, peak, and oversaturated regimes, the same functional (or nonfunctional) form links Regimes B→C and infers demand *D*. This enables multimodal, regional analyses of delay, safety, and queuing, rather than isolated corridor costs.

Misconception 2: High-fidelity simulation is the only way to model congestion.

1) Common belief: Detailed microsimulation or dynamic assignment software is essential for capturing spillback effects, queuing behavior, and the impacts of traffic control, as analytical LPFs are unable to accurately represent real-world congestion dynamics.

2) Why it is misleading and how to address it: While simulators capture the Regime A→B transition and Regime C queues, they incur high computational cost and offer limited marginal cost insight. An analytical LPF, extended by mirroring the congested branch into Regime C, efficiently reproduces key KPIs (throughput, delay, congestion duration). A hybrid workflow, which uses LPFs for KPI-driven calibration and targeted simulation for local validation, balances speed with fidelity across all Regimes.

Misconception 3: Black-box AI models eliminate the need for LPFs.

(1) Common belief: If one can forecast link speed or flow with LSTM, GNN, or transformer models, one no longer needs an LPF.

(2) Why it is misleading and how to address it: Forecasting models (LSTM, GNN, transformers) excel in Regime A but typically ignore queue spillback (B→C) and unobserved demand (observed flow→inferred demand *D*). Without embedding the physical D/C mapping from Fig. 2, a model trained only on Regime A will underperform once capacity is exceeded. Rather than relying solely on predictive models, the integration of ML into a multistage LPF framework comprising fundamental diagram fitting, traffic-state estimation, and dynamic α and β parameter updates enables the extraction of corridor- and region-level insights. These include demand inference, capacity estimation, and real-time system adaptation, which are typically unattainable through standalone forecasting approaches.

3 Key insights from the multiresolution traffic flow modeling perspectives

The development of multiresolution methodologies (MRMs) for traffic analysis has gained significant attention, as highlighted in the comprehensive Department of Transportation (DOT) studies by Zhou et al. (2021) and Hadi et al. (2022). These methodologies, which have been successfully applied in computational fluid dynamics, offer a framework for expressing short-range, mid-range, and long-range relationships consistently and explicitly. In the context of traffic flow, a well-defined VDF that satisfies MRM requirements can reduce the computational complexity that would otherwise be required for detailed simulation and better capture long-range demand–supply interaction phenomena, improving numerical robustness. Notably, many researchers in the field of traffic flow theory also recognize that the traditional BPR formula has limitations in describing flow evolution when capacity is reached. As a result, the subject of macroscopic fundamental diagrams for urban traffic has recently received significant attention (Daganzo and Geroliminis, 2008; Gayah and Daganzo, 2011; Geroliminis and Sun, 2011; Leclercq et al., 2014).

3.1 Key questions driving the field

On the basis of studies by Dowling et al. (1998) and Dowling et al. (2011), we identified critical questions that continue to shape VDF research:

1) How can temporal resolutions and appropriate levels of service (LOSs) be selected to define capacity consistently between design capacity, practical capacity, and ultimate capacity.

2) How can the maximum possible value for capacity be defined while recognizing its stochastic distribution.

3) How can inflow demand be defined and queue discharge rates distinguished from design capacity, especially under heavy congestion.

4) How can daily volume, peak hour volume, and demand duration congestion be mapped through widely used parameters.

5) How can we obtain a consistent average delay from the VDF with the underlying queueing models.

These questions frame the four key insights and actionable takeaway messages that structure the following review.

3.2 Insight 1: Understanding true demand versus measured flow

3.2.1 Measurement problem: Evidence from the literature

The distinction between measured flow and true demand has been recognized by numerous researchers. Akçelik (1996), Dowling and Skabardonis (1993), and Fambro and Rouphail (1997) highlighted the importance of recognizing underlying traffic flow states when addressing demand over capacity conditions. As stated by Chiu et al. (2010), the dynamic traffic assignment (DTA) research community recognized that the V/C ratio does not directly correlate with physical traffic state measurements that describe congestion (e.g., speed, density, or queue). Small and Verhoef (2007) provided a particularly illuminating analysis, highlighting that a certain level of allowed flow can result from either a congested or a hypercongested speed and density. They demonstrated that the BPR function does not account for how long traffic exceeds capacity and proposed a piecewise-linear relationship model to estimate travel time patterns using congestion duration.

We summarize the different modeling definitions of demand D in published papers for congested conditions, as shown in Table 3.

3.2.2 Theoretical foundations from queue theory

Rakha and Zhang (2005) used shockwave and queuing theory procedures to analyze bottlenecks and proposed that the area between demand and capacity curves represents the total delay. Building on this foundation, Huntsinger and Rouphail (2011) derived a demand-to-capacity function by analyzing bottlenecks and queues via freeway detector data. Their key insight is using demand at capacity plus the queue as the demand under congestion conditions. Recently, Bliemer and Raadsen (2020) reformulated the BPR function to propose a model with capacity and storage constraints for static traffic assignment models with possible residual queues and spillback.

3.2.3 Capacity definitions from traffic bottleneck identification perspectives

The complexity of capacity definitions has been extensively studied. Branston (1976) provided an early comprehensive analysis that highlights the critical importance of time interval selection for capacity estimation, whereas Horowitz (1991) and Highway Capacity Manual (2000) established widely used frameworks. The literature reveals multiple capacity concepts:

Practical capacity: Defined by Branston (1976) as the maximum number of vehicles that can pass a given point during a specified time period. Dowling et al. (1998) reported that the practical capacity is 80% of the actual capacity.

Ultimate capacity: Horowitz (1991) emphasized providing consistent procedures to find a “design capacity” that yields reasonable delay estimates.

Stochastic capacity: Jia et al. (2011), Lorenz and Elefteriadou (2000), and Zhang et al. (2025) demonstrated the probabilistic nature of capacity.

A traffic bottleneck is a result of a specific physical condition, often related to the design of the road, a saturated traffic intersection, sharp curves, or incidents. Before the transition from uncongested to congested flow occurs, capacity decreases in the lanes that experience the most traffic. This capacity breakdown phenomenon has been investigated systematically in many studies (Banks, 1991; Cassidy and Bertini, 1999; Hall and Agyemang-Duah, 1991; Kerner, 2000; Lorenz and Elefteriadou, 2000; Han et al., 2025). Hall and Agyemang-Duah (1991) proposed that capacity is not the observed absolute maximum flow. Instead, it is the flow rate that can be repeatedly achieved, day in and day out. This is based on the premise that capacity has a distribution, not a single value achieved on all similar days. Cassidy and Bertini (1999) suggested that enhanced cumulative count curves constructed from counts and occupancies measured at neighboring locations along a roadway can serve as diagnostics for a bottleneck. Zhang and Levinson (2004) suggested that the capacity at a bottleneck should be the weighted sum of the long-run average queue discharge flows and the long-run average prequeue transition flows, both of which follow normal distributions. Kim et al. (2010), Mahmassani et al. (2013), and Yeon et al. (2009) discussed many types of flow rate measures related to capacity, including the prebreakdown flow rate, the postbreakdown flow rate, breakdown flow, and maximum queue discharge flow. On the basis of our literature review findings, the following definitions were considered for data from July 5, 2017 (Wednesday), on a corridor along USA I-405 in Los Angeles, California, on the basis of data from the Performance Measurement System (PeMS) (https://pems.dot.ca.gov/) database, as shown in Fig. 3.

Definition A (free-flow rate): The flow rate corresponding to the free-flow speed.

Definition B (before breakdown rate): A per-lane equivalent hourly rate, observed immediately before the onset of traffic breakdown.

Table 3 Various definitions of demand variables for congested conditions in the literature

Category	Ref.	Detailed definition
Highest flow rates associated with traffic intensity	Kimber and Hollis (1979)	The definition of traffic demand in a stream is q and the capacity available to it μ (both expressed in average vehicles per unit time) at a given stage, traffic intensity is defined as $\rho = V/\mu$
Peak-period inflow volume	Small (1983)	Peak-period inflow volume, uniform rate and delay result from queuing behind a single bottleneck with a constant capacity
Approximated overflow rate	Moses et al. (2013)	D = capacity + (capacity − volume)
Queued demand	Huntsinger and Rouphail (2011)	D = demand at capacity + queue
Approximation through density	Kucharski and Drabicki (2017)	$D = \frac{k}{k_c} \cdot C$
Total volume during the congested period	Cheng et al. (2022); Newell (1982)	$D = \sum_{t_0}^{t_3} q(t)$, $t \epsilon [t_0, t_3]$, t_0 is start time of congestion period, t_3 is end time of congestion period, $q(t)$ is flow at time t Demand is the total volume during a congestion period

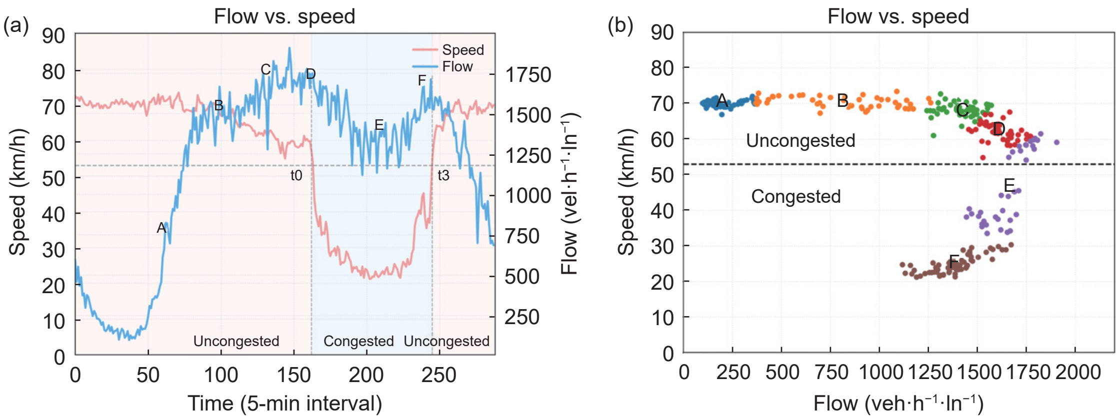


Fig. 3 Different definitions of capacity-related flow rates measured at a bottleneck (5-min intervals): (a) flow and speed along with the time series and (b) flow and speed relationship. (Reproduced with permission from Pan et al., 2025).

Definition C (maximum prebreakdown flow as the ultimate capacity): The highest 5 min flow within 2 h before breakdown during undersaturated conditions. For example, suppose that there is a previous breakdown within the 2 h limit. In that case, the maximum prebreakdown flow is the 5 min flow per lane that occurred after the last congested conditions ended and before the subject breakdown flow started.

Definition D (breakdown): The 5 min flow per lane for the interval before the breakdown event (i.e., before the abrupt speed drop).

Definition E (lowest queue discharge flow): The lowest 5 min flow during the congestion period.

Definition F (recovery flow): The 5 min flow per lane observed immediately after the traffic recovery.

In terms of the evolution of flow and speed measured along with the time series, we considered six states from A to F, accordingly, in one peak period of a bottleneck, as shown in Fig. 3.

Note: Extract cumulative arrival and departure curves to compute D/C ratios and use these ratios in place of the raw V/C ratio when calibrating delay functions. Substituting the D/C ratio for the V/C ratio ensures that the delay function calibration reflects the true demand, yielding more accurate travel-time predictions and network performance assessments.

Recent studies include Belezamo (2020), who introduced a polynomial point-queue model leveraging cumulative arrival/departure data to estimate congested travel times; Wu et al. (2021), who developed a queue-theoretic extension of the BPR function capturing both free-flow and oversaturated Regimes; and Pan et al. (2022), who formulated a dynamic volume-delay function calibrated on two loop-detector corridors via a rolling-horizon framework. Azizi and Beagan (2022, 2023) approached the VDF problem from a practitioner-facing reliability perspective, drawing on lessons from applied planning projects where regime shifts around the capacity boundary matter. Across their two related papers, they proposed a reliability-enhanced volume–delay function (VD(R)F) that augments the traditional mean-based VDF with a reliability measure (e.g., Planning Index), and they show that this perspective helps explain route diversion and assignment outcomes under heavy congestion. Importantly, even though the two papers use different functional specifications, their shared contribution is the recognition that the critical transition is governed by demand relative to capacity (and associated reliability budgets), rather than treating observed volume/capacity alone as the congestion driver. Accurate determination of the D/C ratio, particularly under oversaturated or bottleneck conditions, is critical for reliable traffic assignment and planning. These advancements also create new opportunities for the demand-driven calibration of LPF.

3.3 Insight 2: Fluid-queue models for multiscale consistency

The traditional link performance functions are primarily concerned with average travel time measurements. By examining how time-dependent performance measures are derived from queueing models and numerical methods for solving fluid approximation equations, we aim to shed light on the potential connection between the time-varying delay and its aggregated form. The following section describes the evolution from queueing theory to fluid approximations.

3.3.1 Stochastic queueing with undersaturated conditions

A wide range of link delay functions are adopted in related non-transportation disciplines, such as data communication networks (Bertsekas and Gallager, 2021), and many closed-form link performance functions can be seamlessly embedded in the upper layers of flow control and access control. Notably, these performance functions typically consider undersaturated conditions, the V/C ratio is expressed in terms of utilization rates as $\rho = \lambda/\mu < 1$, and the analytical form of the expected waiting or response time can be derived on the basis of specific queueing systems such as M/G/1 queue (a signal-server queue with Poisson arrivals and a general service-time distribution) (Berman et al., 1987). In the traffic engineering domain, Davidson's function (Davidson, 1978) was developed on the basis of the queueing model with random arrivals and an Erlang service distribution, and Akçelik (1991) proposed an alternative time-dependent queueing function to improve the estimation of intersection delays.

3.3.2 Bottleneck models with piecewise-constant arrival and departure rates

Assuming deterministic, piecewise-constant arrival rates and fixed discharge capacities, the point-queue framework developed by Vickrey (1969) provided closed-form, piecewise-linear travel-time profiles for oversaturated bottlenecks without explicit storage constraints. Embedding early- and late-arrival penalties around a user's preferred arrival time further produces dynamic departure-time equilibrium solutions under a constant value of time. Extensions by Small et al. (2007) introduced hypercongestion effects, in which excessive density reduces flow, to account for queue spillback and upstream blockage. Moreover, Akçelik (1980) applied the same deterministic assumption to signalized intersections, deriving analytical expressions for time-dependent queue length, delay, and stop rates on the basis of effective green-time ratios g/C (where g is the effective green time and C is the signal cycle length), and saturation flows, as demonstrated in Fig. 4.

3.3.3 Fluid approximation queue with polynomial arrival rates for oversaturated conditions

To analyze the queue evolution process, Newell (1968a, 1968b, 1982) proposed fluid-based queueing models with time-dependent arrival rates, characterized by quadratic functions, under oversaturated conditions. As noted by Hall and Agyemang-Duah (1991), this type of deterministic fluid approximation model has also been used as an essential tool to represent nonstationary queueing systems, especially with "predictable fluctuations", which tend to overwhelm random changes in arrival and service processes. Schwarz et al. (2016) systematically reviewed the performance of time-dependent queueing systems and classified them into three groups: Numerical and analytical, piecewise constant, and modified system models. Recently, Cheng et al. (2022) extended Newell's queue model to consider cubic arrival rates to analytically capture the evolution of queue saturation with closed-form delay functions.

It is crucial to understand how polynomial arrival queue (PAQ) models represent different system performance characteristics on the basis of a parsimonious set of critical parameters. Such an understanding can help planners establish a consistent definition and unified modeling framework to map time-varying traffic performance observations to different modeling elements in traditional VDFs. As indicated in Table 4 and Fig. 5, we offer the following four observations. (1) In a given analysis period $[t_s, t_e]$, where t_s is start time of analysis period, t_e is end time of analysis period, the inflow demand $\lambda(t)$ can be represented by the

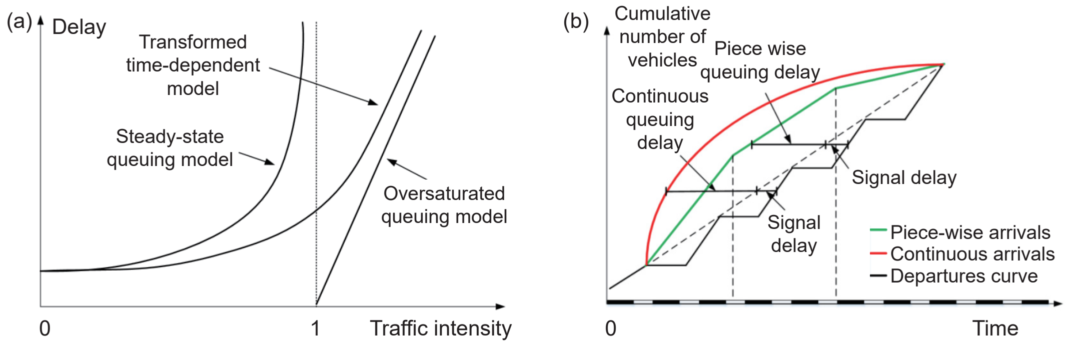


Fig. 4 Deterministic queuing analysis at a signal intersection (Kimber and Hollis, 1979): (a) steady-state delay, deterministic queuing delay, and transformed delay curves and (b) arrival and departure curves for a saturated link.

Table 4 Key parameters and performance measurements among the different queueing models

Model	Loading ratio	Arrival process	Departure process	Travel time output	Modeling of interaction
Stochastic queue	λ/μ for undersaturated conditions	Random arrival processes such as Poisson distribution	Determined by service time distribution	Average waiting time or response time	Delay is caused by interference among vehicles/customers.
Standard point queue in bottleneck model	A queue will form if the outflow rate exceeds the maximum discharge rate (capacity flow)	Piecewise constant time-varying arrival rates	Constant maximum discharge rate	Closed-form solutions for time-varying travel time	Outflow depends on its own inflow, bottleneck capacity; vehicles stack vertically, and the queue occupies no space and does not affect upstream links.
Fluid approximation queue	D/μ for oversaturated conditions	Polynomial arrival rates with quadratic or cubic form, specifically for the congested duration only. The arrival rate outside congestion duration is not modeled due to possible approximation errors	Constant discharge rate	Closed-form solutions for time-varying travel time	During congestion duration, the vehicle's space-time trajectories are mapped from the fluid approximation dynamic point queueing process to spatial queues with constant congestion speed.
Link travel time function	Time-dependent flow rate	Allow form-free time-dependent arrival rates and capacity-based inflow restrictions	Link exit flow function with link capacities as an upper bound	The explicit function forms for time-dependent travel time; link travel times reach an upper bound after spillback	Capture travel time by modeling delays without capturing queues explicitly, a special type of FD lacking the capability of withholding traffic flows when exceeding the physical link capacity.
Kinematic wave related models	Determined by initial and boundary conditions in terms of cumulative vehicle numbers	Allow form-free time-dependent arrival rates during discretized time intervals	Outflow depends on its own inflow, bottleneck capacity, and downstream spatial capacity supplies	Indirectly derived from speed in numerical solutions, link travel time could reach an upper bound after spillback	Simplified kinematic wave propagation principle for modeling shock waves.

Note: Newell's closed-form method is used to translate the time-dependent queue length $Q(t)$ into time-dependent travel times rather than relying solely on empirical VDFs. Given the many underlying queuing representations shown in Table 4, modelers must therefore strike an application-driven balance between representational fidelity and parsimonious model requirements, choosing the simplest formulation that still captures the key dynamics needed for their real-world context.

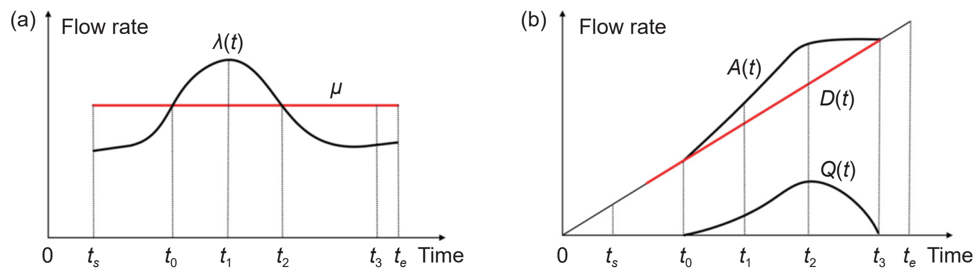


Fig. 5 Deterministic queuing analysis of a bottleneck: (a) polynomial arrival and constant discharge rate and (b) cumulative arrival and departure rates.

congestion duration P and the queue discharge rate μ during the congestion period $[t_0, t_3]$. (2) PAQ is able to provide a closed-form average delay $\overline{w}$ that is consistent with the time-dependent queueing delay $w(t)$. (3) There are subtle but important differences between the macroscopic parameters of V/C in a VDF and the parameters D/μ. (4) The time-dependent travel time $tt(t)$ in the underlying fluid approximation queue model should satisfy the first-in-first-out (FIFO) property and capacity constraints.

Balance of Fidelity and Parsimony: While Newell's closed-form relationship offers lightweight, direct mapping, the more sophisticated fluid-queue and kinematic-wave schemes (e.g., the cell transmission model proposed by Daganzo (1994), the link transmission model proposed by Yperman et al. (2005), and modified input-output diagrams) provide richer spatiotemporal detail. However, these methods require greater computational effort and more extensive data inputs for dynamic traffic assignment (Ran et al., 1996; Ran and Boyce, 1996). Lawson et al. (1997) used a modified input-output diagram to connect the space-time extent of a queue and the delay of a bottleneck to simplify the evaluation of the waiting time, queue length, and travel time in a jam. As highlighted by Carey (2004) and Carey et al. (2014), the link travel time function should satisfy certain properties, particularly uniqueness and continuity, FIFO, causality, and time-flow consistency, and there are consistent research efforts in this direction (Long et al., 2011; Raadsen and Bliemer, 2019).

3.4 Insight 3: Congestion duration as the primary delay driver for traffic bottlenecks

3.4.1 *Beyond instantaneous metrics: Literature evidence*

Travel planning models need to forecast future travel demand and load that demand onto a roadway network. Typically, in engineering practice, a peak hour factor (PHF) is used to convert the peak hour traffic volume into the flow rate representing the busiest 15 min of rush hour in the assignment period (e.g., AM or PM). Previous research has indicated that the PHF can strongly impact traffic analysis results (Tarko and Perez-Cartagena, 2005). It can be used to quantify the effects of short-term traffic peaking, which can lead to congestion.

In traffic assignment applications, to obtain a more temporally comprehensive understanding of congestion impacts, planners can use a peak spreading method (Horowitz et al., 2014) to expand the peak period of traffic from the traditional one-hour peak period to a multihour peak period. In a recent study by Wu et al. (2021), the threshold speed is selected first so that congestion durations and queued demand volume for each link are computed more precisely. The relationships among the PHF, peak spreading characteristics, and assignment period are shown in Fig. 6. On the basis of speed data, one can determine which hours of traffic surpass capacity, and this hour is typically defined as the peak hour. There are essential differences between congestion duration and the heaviest volume hour: All the hours in which inflow demand volumes exceed capacity have been identified in congestion duration.

Notably, in the literature on traffic delay functions for signalized intersections, the degree of oversaturation is expressed as $x =$ traffic peaking intensity (Rouphail et al., 1992), which is in turn represented by the ratio of the highest volume and capacity. From the perspective of fluid approximation queues with quadratic arrival rates, Newell (1982) provides a closed-form solution as

$$\frac{D}{\mu} = P = 3 \times \left(\frac{\lambda(t_1) - \mu}{\gamma}\right)^{\frac{1}{2}} \tag{6}$$

where $\lambda(t_1)$ is the highest volume at time t_1, as shown in Fig. 5a, γ is the inflow demand curvature parameter, and P is the congestion duration.

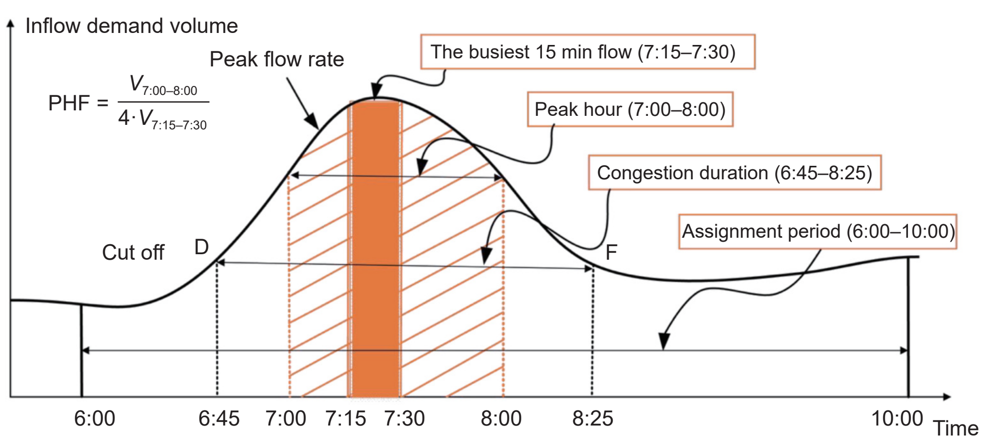


Fig. 6 Relationships among the peak flow rate, congestion duration, and assignment period; points D and F correspond to states D and F in Fig. 3. The vertical axis represents the inflow demand volume, as the observed flow count between states D and F is lower than the ultimate hourly capacity, as illustrated in Fig. 3.

As discussed in the trip scheduling literature (Small and Verhoef, 2007), the key variables of congestion duration P and longest waiting time in the queuing-based volume-delay function (QVDF) (Zhou et al., 2022) should be integrated into the departure time choice behavioral model while considering the preferred arrival time and schedule delay in bottleneck models. Recently, Wu et al. (2021) and Zhou et al. (2022) proposed the use of the total inflow demand volume during the congestion duration as the D variable in the D/C ratio. By introducing two types of elasticity functional forms, Zhou et al. (2022) further demonstrated connections from the macroscopic inflow D/C ratio to the congestion duration of a bottleneck, from the congestion duration to the magnitude of speed reduction.

Traditional metrics such as speed, volume, and the peak V/C ratio often overlook the fact that congestion duration ($P = t_3 - t_0$) explains far more delay variation than does the peak V/C ratio alone. We therefore recommend fitting a fluid-queue model to identify $[t_0, t_3]$ and adopting Δt (or P) as a core KPI in both VDF calibration and signal-timing analyses. As illustrated in Fig. 7, starting from the inflow rate $\lambda(t)$, the diagram contrasts two Regimes:

1) Uncongested conditions: A higher V/C ratio leads to a nonlinear increase in travel time (tt). Increasing the link capacity C reduces the V/C ratio, thereby decreasing the travel time tt.

2) Congested conditions: The demand-to-capacity ratio D/C governs key traffic state variables, including the queue length $Q(t)$, discharge rate μ, and congestion duration P. As the D/C ratio increases, $Q(t)$ increases, μ decreases, and P increases, which amplifies the overall delay and increases the time-dependent travel time $tt(t)$.

This causal map encapsulates how duration, not just intensity, underpins bottleneck delay and highlights where parsimonious fluid-queue models can improve LPF accuracy.

3.4.2 Operational framework for D/C ratio calibration and implementation

With the advancement of transportation sensing techniques, it has become increasingly important to systematically utilize available data to select appropriate VDF forms and calibrate their associated parameters. The general workflow for developing and calibrating VDF models can be summarized as follows, as illustrated in Fig. 8.

1) Fundamental diagram calibration: The process of calibrating the fundamental diagram involves determining and adjusting coefficients, including the road capacity c, critical speed v_c, critical density k_c, and free-flow speed v_f, on the basis of traffic flow models. Many empirical studies (Cheng et al., 2021; Ni et al., 2016; Wang et al., 2011) have confirmed that the FD provides a clearer representation of traffic flow characteristics and can be used to distinguish between various traffic states on the basis of important parameters.

2) Computation of the V/C or D/C ratios: Determining the V/C or D/C ratio involves analyzing existing congestion bottlenecks by determining the congested period and average waiting time and computing the demand during peak periods. The key parameters, the V/C ratio and the D/C ratio can be calculated for different facility types.

3) Calibrating volume-delay functions: Develop and calibrate one or multiple functional forms of volume-delay functions (e.g., α, β in the BPR function), taking into consideration the probabilistic distributions of capacity, demand, and travel time parameters.

4) Traffic assignment: The calibrated VDFs are embedded into traffic assignment models for network design or policy analysis. The traffic assignment results are compared against baseline data via a specified origin–destination matrix, and model outputs are validated against observed traffic conditions in the base year.

5) Evaluating measures of effectiveness (MOE): On the basis of the assignment outputs, additional MOEs such as queue length, delay, and emissions can be derived for both short-term and long-term planning tasks.

3.5 Insight 4: Microscopic-to-Macroscopic cross-resolution modeling-extending LPFs for connected and automated vehicle evaluation

Connected and automated vehicles introduce unique microscopic traffic flow characteristics that challenge traditional assumptions embedded in LPFs. Understanding how these microscopic behaviors scale to macroscopic traffic dynamics is crucial for designing efficient and sustainable transportation systems. We first synthesize the key literature and then demonstrate how microscopic CAV behaviors map to macroscopic LPFs via concrete examples.

3.5.1 Literature on microscopic dynamics and car-following model integration

An early study by Chen et al. (2011) applied the describing function method, substituting a nonlinear car-following model with a linear model. This process begins with Newell's parsimonious car-following model, which correlates the reaction time and minimum safe distance. Li (2022) emphasized the crucial role of microscopic car-following models in balancing

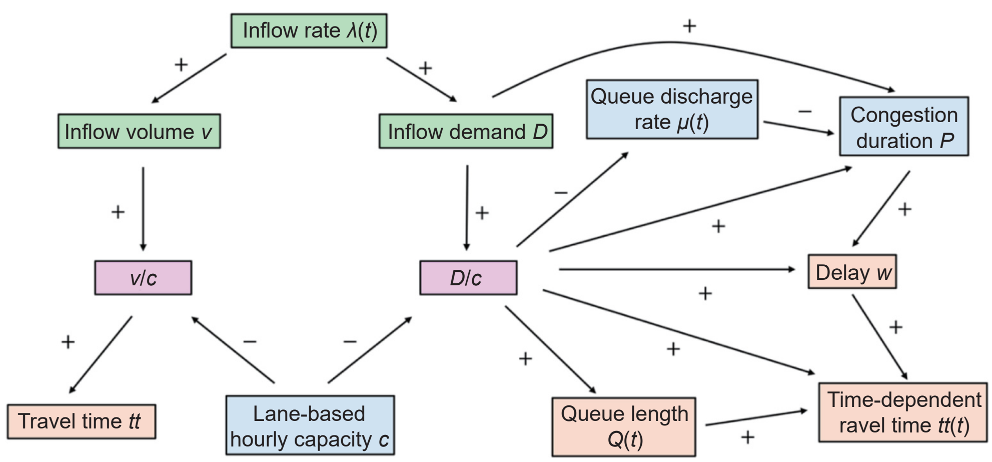


Fig. 7 Causal interactions of key variables in the LPF framework.

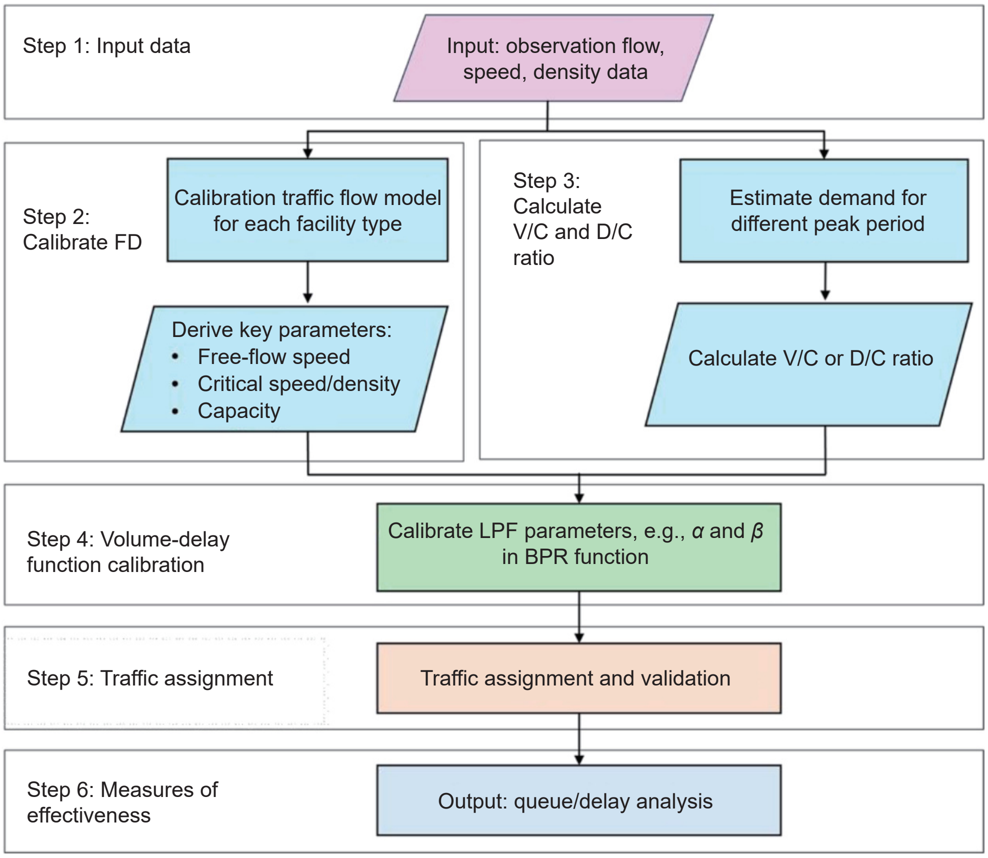


Fig. 8 Modeling calibration frameworks for characterizing VDFs.

safety, mobility, and stability for commercial AVs. Nonlinear effects within AV control may amplify traffic oscillations, posing challenges for modeling such dynamics. Huang et al. (2023a), Levin and Boyles (2015), and Mahmassani (2016) examined how the minimum safe distance can be reduced in CAV scenarios. More recent work has focused on these foundations: Vehicle-to-infrastructure (V2I)-enabled automation can reduce intersection saturation headways by more than 20% (Hajbabaie et al., 2024), whereas mixed-traffic throughput gains only emerge past certain CAV thresholds (Li et al., 2025). Moderate AV penetration improves urban travel-time reliability (Samaranayake et al., 2024), although intersection capacity still hinges on market share and signal logic (Song and Fan, 2023), and platooning models extend turbo-roundabout capacity forecasts (Guerrieri, 2024).

3.5.2 Literature on future transportation network design

Transportation network design is an essential planning task for improving transportation networks' speed and travel times by selecting the optimal projects among alternatives. Through a case study in Stockholm, Engelson and van Amelsfort (2015) highlighted the importance of refining the VDF for the forecasted year in evaluating network design to avoid overestimation or underestimation of the flow/travel time changes. In addition, the design of a new road network or the upgrading of existing roadways would require a good understanding of the uncertainty involved and the impact on system-wide performance (Chen et al., 2011). Many studies, such as Correa et al. (2004), Nagurney and Qiang (2012), Wei et al. (2019), Wong and Wong (2016), Yang and Wang (2011), and Zhang and Levinson (2004) have focused on evaluating various problems related to transportation network design problems.

3.5.3 Some key questions need to be considered in future studies

1) How can consistent definitions and a unified modeling framework be developed for multiresolution modeling?

Future studies should use an integrated supply-side parameter calibration package with consistent definitions of traffic flow models and macroscopic VDFs. They can capture the relationships among discharge rates, the demand-to-capacity ratio during the congestion period, and queue spillback.

2) How can the VDF function be developed to analyze performance on both the demand and supply sides?

It is important to determine the cost equilibrium point where demand equals supply. While traditional static models can significantly underestimate network congestion levels in traffic networks (Boyles et al., 2008), how to connect VDFs with available DTA models to account for variable demand and traffic dynamics can be a critical future research direction, especially under a policy of time-dependent congestion pricing.

Using Fig. 9 and Table 5, we examine three distinct scenarios: The baseline with zero AV penetration (blue dotted line), high AV penetration (solid red line), and the managed demand scenario (solid green line).

1) Microscopic car-following dynamics. Our analysis begins with Newell's parsimonious car-following model, which correlates the RT and minimum safe distance. Fig. 9a shows trajectory transitions from free-flow to congested speeds, whereas Fig. 9b displays the resulting piecewise linear approximation trajectories. Consistent with previous research (Huang et al., 2023a; Levin and Boyles, 2015; Mahmassani, 2016), the minimum safe distance decreases from 7 m at baseline to 5 m under high AV penetration, directly influencing the backward wave speed. Fig. 9c shows the linear relationship between the distance headway and speed in Newell's model, where the slope represents the reaction time and the intercept signifies the minimum distance headway. This

Microscopic-to-Macroscopic linkage is essential for understanding how individual vehicle interactions influence broader traffic dynamics.

2) Macroscopic flow and capacity. Figs. 9d and 9e show the macroscopic traffic flow and capacity characteristics approximated via a triangular fundamental diagram (TFD). The backward wave speed in the TFD aligns significantly with the wave speed of the microscopic car-following model. While the TFD provides

Table 5 Parameter transitions and data requirements across micro, meso, and macro scales

Scale	Key parameter	Parameter transition	Data requirement
Micro-level	Reaction time τ_0, minimum gap d_0, acceleration distribution	Driving behavior affects car-following and lane-changing dynamics	High-resolution CAV trajectory data, simulation outputs
Macro-level	Critical density k_c, capacity C, backward wave speed w	τ_0 and d_0, translate into k_c and w; fundamental diagram fitting yields capacity	Loop detector flow and speed data, probe vehicle measurements
Meso-level	Discharge rate μ, D/C ratio, flow rate q and congestion duration P	k_c, q and w result in μ, D/C, P derived from cumulative curves	Cumulative arrival-departure curves, OD demand, and bottleneck data

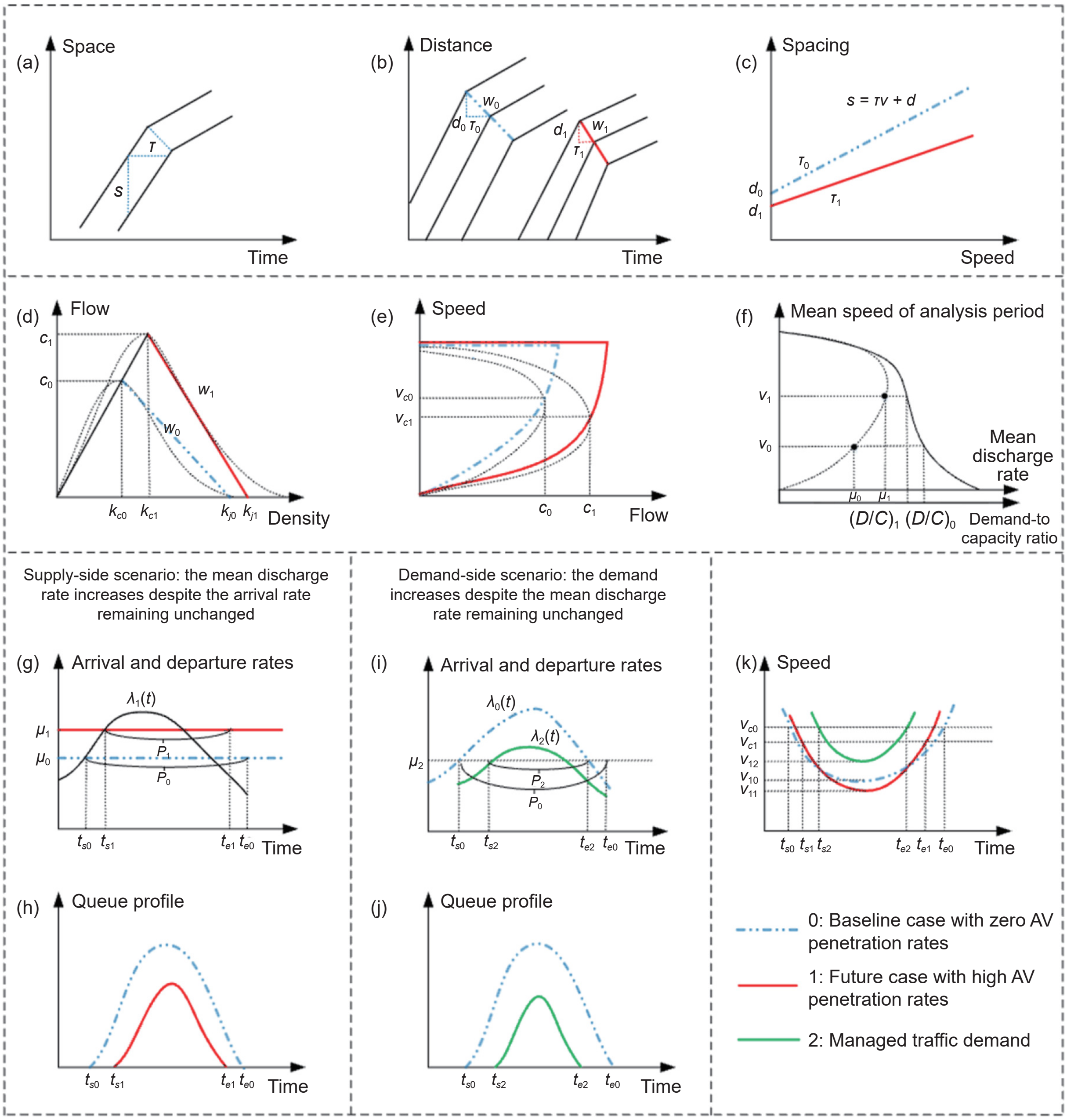


Fig. 9 Comprehensive LPF modeling framework: A micro- to macro-level perspective across diverse scenarios: (a) piecewise linear approximation to vehicle trajectories with a speed change, (b) illustration of vehicle following trajectories for different AV penetration rates, (c) relation between spacing and speed, (d) flow-density relationship in triangular and S3 model, (e) speed-flow relationship in S3 model, (f) speed-D/C relationship based on FD model, (g) time-dependent arrival rates and average discharge rate, (h) time-dependent queue length, (i) arrival and average discharge rate with different demand-side scenario, (j) time-dependent queue length and (k) time-dependent speed during congestion period.

straightforward approximations, other fundamental diagram shapes, such as the S3 model (Cheng et al., 2021), can be represented (dotted line). The abrupt transitions from free-flow to congested states in the triangular flow-density curve (Fig. 9e) are generally undesirable for calibrating traffic volume-delay functions. Traditional highway capacity studies demonstrate smoother transitions through LOS A–D, mirroring the S3 model's speed-flow relationship. Fig. 9f illustrates the "mirroring" perspective linking fundamental diagrams and linking performance functions by mapping congested states to Regimes with demand-to-capacity ratios greater than 1.

3) System-level implications. At the macroscopic level, we explore supply-side and demand-side scenarios for regional policy decision support. Supply-side improvements through higher CAV penetration enhance capacity, resulting in shorter congestion durations despite constant arrival rates (Figs. 9g and 9h). In demand-side scenarios employing active traffic management strategies, including time-varying tolling and advanced reservation systems, arrival rates are reduced, queue lengths are shortened, and discharge rates remain consistent, as illustrated in Figs. 9i and 9j. Fig. 9k illustrates time-dependent speed variations across all three scenarios during congestion periods, with the corresponding parameters detailed in Tables 6 and 7. These visualizations collectively demonstrate the intricate relationships between flow rates and system characteristics across various temporal and spatial scales.

To operationalize the proposed multiresolution framework, microscopic CAV parameters must be systematically transitioned into mesoscopic delay representations. A reduction in the reaction time τ_0 and minimum gap d_0 at the microscopic level translate into a higher critical density k_c and a modified backward wave speed w in the fundamental diagram. These macroscopic parameters directly influence mesoscopic delay estimation through changes in the discharge rate μ and congestion duration P. Calibration is performed hierarchically: (1) microscopic parameters are estimated from CAV trajectory data or simulations; (2) macroscopic parameters are obtained by fitting the fundamental diagram; and (3) mesoscopic parameters such as the discharge rate and D/C ratios are extracted from cumulative arrival–departure curves. Data requirements thus vary across scales, ranging from high-resolution trajectory datasets at the microlevel to loop-detector or probe-based aggregate flows at the macrolevel to OD demand and bottleneck-specific measurements at the meso-level. A summary of these parameter transitions and calibration procedures is provided in Table 5.

The four insights collectively highlight the need for physically informed and scalable delay models, which motivates the AI-driven framework presented in Section 4.

4 Physics-informed AI for delay modeling: Research challenges in an integrated demand-supply framework

This section presents an original, unified end-to-end AI-driven LPF framework that integrates data-driven approaches with physics-based constraints in VDF modeling. Whereas the preceding sections primarily synthesize literature, the framework proposed here constitutes an original contribution designed to address shortcomings in current practices, which include the lack of physical consistency in purely data-driven models and the limited flexibility of traditional formulations. Our framework combines the strengths of both paradigms: Embedding traffic flow conservation constraints into learning architectures while also introducing regime-aware benchmarking and a standardized evaluation protocol to systematically compare model performance. In addition, the framework leverages feature

Table 6 Comparison of microscopic and macroscopic parameter differences between the baseline and future scenarios

Parameter	0: Baseline case with zero AV penetration rates	1: Future-year case with high AV penetration rates	Ref.
Minimum distance between the leading and following vehicle d (m)	$d_0 = 7$	$d_1 = 5$	Huang et al. (2023a); Levin and Boyles (2015); Mahmassani (2016);
Reaction time τ (s)	$\tau_0 = 1.5$	$\tau_1 = 1$	Levin and Boyles (2015); Talebpour and Mahmassani (2016); Shi and Li (2021); Wei et al. (2017)
Backward wave speed w (km/h)	$w_0 = -16.8$	$w_1 = -18$	$w = \frac{d}{\tau}$
Free-flow speed v_f (km/h)	$v_{f0} = 112$	$v_{f1} = 112$	Shi and Li (2021); Zhou et al. (2022)
Capacity C (veh·h^{-1}·ln^{-1})	$c_0 = 1800$	$c_1 = 2160$	
Jam density k_j (veh·km^{-1}·ln^{-1})	$k_{j0} = 143$	$k_{j1} = 200$	$k_j = \frac{1000}{d}$
Critical density k_c (veh·km^{-1}·ln^{-1})	$k_{c0} = 36$	$k_{c1} = 80$	$k_c = k_j - \frac{C}{w}$
Critical speed v_c (km/h)	$v_{c0} = 50$	$v_{c1} = 27$	$v_c = \frac{C}{k_c}$
Reference queue-based VDF parameters	$\alpha = 0.21$ $\beta = 1.87$ $f_p = 0.23$, f_p is constant in elasticity function for mapping congestion duration to the magnitude of speed reduction. $f_d = 1.37$, f_d is constant in elasticity function for mapping D/C ratio to congestion duration. $s = 1.64$, s is elasticity coefficient of speed reduction magnitude in response to congestion duration changes, i.e., duration-to-speed reduction elasticity. $n = 1.14$, n is elasticity coefficient of congestion duration in response to D/C changes, i.e., oversaturation-to-duration elasticity.		Zhou et al. (2022)

Table 7 Comparative analysis of demand- and supply-side parameters across the three scenarios

Typical value	0: Baseline case with zero AV penetration rates	1: Future-year case with high AV penetration rates	2: Manage active traffic demand	Ref.
Demand D (veh/ln)	$D_0 = 3600$	$D_1 = 3600$	$D_2 = 2880$	Assumed for illustration with 20% congested demand decrease
Demand over capacity ratio D/C	$\frac{D_0}{C_0} = 2$	$\frac{D_1}{C_1} = 1.67$	$\frac{D_0}{C_0} = 1.6$	Zhou et al. (2022)
Congestion duration P (h)	$P_0 = 3.02$	$P_1 = 2.45$	$P_2 = 2.34$	$P = f_{\mathrm{d}} \cdot \left(\frac{D}{C}\right)^n$
Mean speed during congestion duration $\bar{v}$ (km/h)	$\bar{v}_0 = 28.51$	$\bar{v}_1 = 17.47$	$\bar{v}_2 = 33.47$	$\bar{v} = \frac{v_{\mathrm{c}}}{1 + \alpha\left(\frac{D}{C}\right)^{\beta}}$
Queue discharge rate μ (veh·km^{-1}·ln)	$\mu_0 = 1192$	$\mu_1 = 1468$	$\mu_2 = 1230$	$\mu = \min[\frac{C}{f_{\mathrm{d}}\left(\frac{D}{C}\right)^{n-1}}, C]$
Lowest speed v_l (km/h)	$v_0 = 20.93$	$v_{l1} = 13.49$	$v_{l2} = 26.14$	$v_l = \frac{v_{\mathrm{c}}}{f_{\mathrm{p}} {f_{\mathrm{d}}}^{s} \left(\frac{D}{C}\right)^{ns} + 1}$
Start congestion time, t_{s}	$t_{\mathrm{s}0} =$ 7:00	$t_{\mathrm{s}1} =$ 7:10	$t_{\mathrm{s}2} =$ 7: 20	$P = t_{\mathrm{e}} - t_{\mathrm{s}}$
End congestion time, t_{e}	$t_{\mathrm{e}0} =$ 10:01	$t_{\mathrm{e}1} =$ 9:37	$t_{\mathrm{e}2} =$ 9:42	

Note: The cross-resolution methodology enables comprehensive evaluation of emerging transportation technologies by integrating microscopic car-following models, mesoscopic link performance functions, and macroscopic fundamental diagrams. This approach captures critical operational metrics beyond travel time, including congestion duration, network resilience, and capacity variations, while incorporating real-world constraints such as technology penetration rates and infrastructure limitations. For practitioners, this framework provides evidence-based insights for strategic deployment (planners), infrastructure design optimization (engineers), and robust analytical tools that bridge behavioral dynamics with system performance (modelers), ensuring that technology assessments reflect operational reality rather than idealized scenarios.

engineering, graph neural networks, and attention mechanisms to integrate roadway attributes with external factors, enabling the prediction of link capacity and the updating of LPF parameters. In doing so, it balances theoretical consistency with cross-scenario generalization.

4.1 From empirical LPFs to AI-driven delay models

Traditional LPFs are typically based on empirical observations and theoretical assumptions about traffic flow characteristics, such as steady-state conditions and uniform congestion propagation. However, these assumptions may limit their accuracy in complex urban environments. With advances in ML and AI, researchers have begun exploring data-driven approaches to improve the predictive accuracy of LPFs and enhance their adaptability to dynamic and complex traffic conditions. ML techniques can capture nonlinear traffic flow relationships and integrate heterogeneous data from multiple sources, such as sensor-based traffic counts, GPS trajectories, and real-time incident reports. Studies have shown that ML models such as neural networks (Wang et al., 2018), support vector regression (Wu et al., 2004), random forests (Cheng et al., 2019) and LSTM (Duan et al., 2016) can reduce prediction errors and better adapt to different congestion patterns, outperforming traditional LPFs.

In addition to improving predictive accuracy, AI-enhanced LPF studies have made progress in traffic management applications, such as congestion pricing optimization (Genser and Kouvelas, 2022), dynamic route choice (Zhao and Liang, 2023), and adaptive signal control (Srinivasan et al., 2006). These studies indicate that integrating AI with LPFs not only overcomes the constraints of static empirical formulas but also enables the development of adaptive, real-time traffic models. Despite these advancements, AI-driven LPFs still face challenges in practical applications, including data sparsity, the interpretability of black-box models, and computational efficiency issues.

Fig. 10 illustrates the positions of three LPF modeling approaches within a three-dimensional space defined by reliance on traffic-flow theory (vertical axis), utilization of empirical data (depth axis), and ability to generalize across networks (horizontal axis). Classical physics-driven LPFs are located in the high-theory, low-data, high-generalization region. They are firmly grounded in fundamental diagram theory but have limitations in adapting to real-time and complex traffic patterns. Purely data-driven LPFs are situated in the low-theory, high-data, low-generalization region. They rely on sensor and GPS data to capture nonlinear relationships but often lack interpretability and transferability.

Physics-informed neural networks (PINNs) are a class of deep learning algorithms that can seamlessly integrate data and abstract mathematical operators, including Partial Differential Equations (PDEs) with or without missing physics (Karniadakis et al., 2021). PINN-based LPFs appear at the top-right corner of the cube as a more comprehensive and enriched framework that integrates traffic-theory constraints with large-scale data learning. The arrows pointing from the green and blue regions toward the red region indicate that PINNs draw on the strengths of both physics-

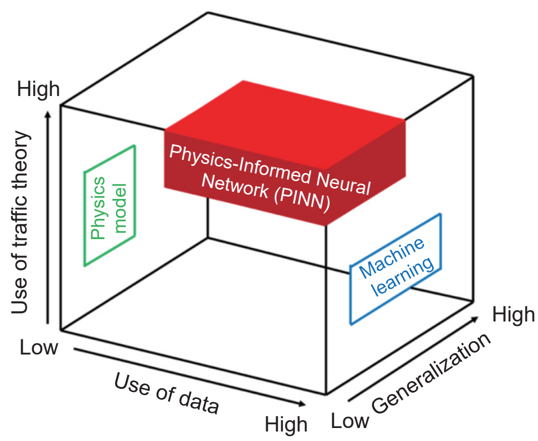


Fig. 10 Modeling paradigms for link performance functions: Balancing theory, data, and generalization.

based and data-driven approaches, enabling them to balance interpretability, adaptability, and cross-scenario performance, thereby forming an ideal “sweet spot”. Fig. 10 does not present a strict comparison of the absolute level of theory contained in each model. Instead, it highlights how traffic theory is incorporated into different modeling paradigms: Traditional physical models rely directly on traffic theory, whereas ML models typically do not explicitly include it, whereas PINNs embed theoretical constraints more deeply into the learning process, which explains their slightly greater placement in Fig. 10.

In practice, PINNs are not merely positioned conceptually at the intersection of theory and data but rather operationalize this integration by embedding physical residuals directly into the training loss. For traffic applications, residual terms derived from the Lighthill-Whitham-Richards (LWR) continuity equation, queue balance relationships, and FD consistency are added as penalty components to the loss function. This explicit formulation ensures that predicted density, flow, and delay trajectories evolve in compliance with conservation laws and remain physically interpretable, thereby distinguishing PINNs from conventional ML approaches that optimize purely against data-fitting error.

4.2 Physics-informed hybrid architectures and open challenges

Various ML methods have been adopted in transportation studies (Rowan et al., 2025) to capture complex spatiotemporal patterns in traffic data. Representative examples include k-nearest neighbor algorithms (Tak et al., 2014), convolutional neural networks (Han et al., 2022), LSTM networks (Ma et al., 2015), GNNs (Cui et al., 2020; Liu and Meidani, 2024), and reinforcement learning approaches (Zhou and Gayah, 2023). In contrast to traditional physics-based models, ML methods are capable of learning subtle and nonlinear relationships among traffic state variables while accommodating data variability.

Several studies have laid the foundation for deep learning and hybrid modeling in transportation, progressively advancing temporal, spatial-temporal, and physics-informed approaches to LPFs. The first application of LSTM to traffic speed prediction was by Ma et al. (2015). Ke et al. (2017) extended LSTM with convolutional layers in their fusion convolutional long short-term memory network (FCL-Net) to model short-term passenger demand, effectively fusing spatial and temporal dependencies. Wu et al. (2018) introduced a hierarchical flow network (HFN) that uses computation graphs to represent multilevel travel demand variables, including trip generation, OD matrices, path/link flows, and behavior parameters, thereby enabling the structured propagation of errors across planning layers.

In the broader context of delay modeling, Shi et al. (2022) developed a physics-informed deep learning framework incorporating a fundamental diagram learner (PIDL+FDL), which integrates fundamental diagram estimation, state estimation, and parameter identification within a unified neural framework for highway traffic. Recent studies have extended PINNs by integrating them with various traffic flow theories and models to enhance physical consistency and interpretability. Notable efforts include the incorporation of car-following models (Liu et al., 2023), first-order partial differential equations (Huang and Agarwal, 2023b), fundamental diagrams (Pan et al., 2024), macroscopic fundamental diagrams (Tang et al., 2024), fluid queue theory (Doll et al., 2024; Xu et al., 2024), deterministic queuing models (Lu et al., 2023; Pan et al., 2025), the bathtub model (Ka et al., 2024), Newell’s theory (Sengupta and Guler, 2025), Kalman filter-based frameworks (Deshpande and Park, 2025), and vehicle trajectory prediction (Long et al., 2024, 2025).

We present the following open challenges to the ML community, encouraging contributions that integrate VDF-based D/C analysis with multi-source data and AI methods for transportation network analysis:

1) Data and AI integration: Beyond black-box approaches. Traditional ML approaches in traffic state estimation bifurcate into model-driven and data-driven categories, each suffering from either deficient physics or insufficient data. The emergence of the PINN represents a paradigmatic shift, offering a deep learning framework that embeds knowledge of the physical laws governing traffic flow datasets in the learning process. The computational complexity of transportation networks demands sophisticated AI architectures that can handle temporal patterns and spatial correlations simultaneously. Physics-aware neural networks using VDF embedded in network architectures could yield traffic state estimations physically constrained by macroscopic traffic flow models, addressing the fundamental limitation of black-box approaches that fail to capture the underlying physics of traffic behavior.

2) Sensor coverage, link proportion matrices and VDF calibration. Real networks rely on sparse loop detectors and intermittent GPS probes, yet accurate VDFs demand fine-grained link proportion matrices for OD mapping and modular network design. Reliable calibration of time-varying VDF parameters (α, β, $\mu(t)$) requires robust denoising, synthetic data augmentation, and transfer learning across corridors to ensure that dynamic D/C profiles reflect true area-level flows rather than isolated bottlenecks.

3) Differentiable VDF formulations with area-scale D/C and queue spillback**.** Classical VDFs treat V/C at a single link; hybrid architectures must generalize to area-scale D/C ratios and model queue spillback across adjacent links. Smooth, gradient-friendly representations of congestion duration P (oversaturated cycles) and macroscopic flow-density-speed relationships enable LSTM and other models to learn delay functions adhering to flow conservation and interpretability.

4) Online, multiday, multiresolution inference. The delay dynamics span the signal cycle (s), queue buildup (min), and demand fluctuation (day). Hybrid VDFs should bridge multiple spatial and temporal scales by employing tensorized flow representations or hierarchical time grids. This enables support for real-time and multiday inference in traffic simulators and operational platforms, allowing continuous adaptation to varying inflow rates and control actions.

5) Multimodal data fusion and physical spatial behavior. In addition to vehicle counts, VDFs should incorporate pedestrian flows, transit schedules, EV charging profiles, CAV trajectories, weather conditions, and incident data, integrating them within a unified computational graph. This fusion allows AI modules to capture how diverse demand and supply streams jointly influence delays, reflecting spatial correlations that black-box methods often overlook.

6) Parsimonious AI and control-oriented design. Complex models risk overfitting and slow inference. A parsimonious approach begins with simple ML models, such as LSTM integrated with physics-based priors, to estimate delay functions. It then incrementally incorporates hybrid components, including noisy data filtering, event embedding, and flow-through tensor representations, to support lightweight control policies such as signal timing and dynamic pricing under system-optimal objectives.

A key advantage of PINNs lies in their ability to incorporate domain-specific knowledge into the training objective, thereby ensuring that the predicted traffic states adhere to first principles.

In the context of traffic flow modeling, several classes of physical constraints can be systematically embedded:

1) Conservation constraints. Vehicle conservation can be enforced by incorporating the residuals of the Lighthill–Whitham–Richards (LWR) continuity equation:

$$\frac{\partial k(x,t)}{\partial t}+\frac{\partial q(x,t)}{\partial x}=0 \tag{7}$$

This penalization ensures that the predicted density $k(x,t)$ and flow $q(x,t)$ evolve consistently with conservation laws, thereby preventing unrealistic outcomes such as vehicles appearing or disappearing within the network.

2) Queue balance constraints. For bottleneck or fluid-queue models, the consistency between cumulative arrivals and departures is as

$$Q(t)=A(t)-D(t)\geq 0 \tag{8}$$

where $A(t)$ represents the cumulative arrival count and $D(t)$ represents the cumulative departure count. The queue length could also represent the cumulative imbalance between the inflow and outflow rates from the onset of congestion t_0 to time t.

$$Q(t)=\int_{t_0}^{t}(\lambda(\tau)-\mu(\tau))\,\mathrm{d}\tau \tag{9}$$

where $\lambda(t)$ is the time-dependent inflow demand rate and $\mu(t)$ is the time-dependent outflow rate. Equation (9) is introduced as an additional training constraint. This formulation guarantees that the queue length $Q(t)$ reflects the cumulative imbalance between arrivals and departures, ensuring physically meaningful and bounded queue evolution.

3) Fundamental diagram consistency. The relationship between flow and density can be explicitly enforced by penalizing deviations from a calibrated FD, e.g., the triangular FD:

$$q(t)=\begin{cases} v_{\mathrm{f}}\cdot k(t),\, k(t)\leq k_{\mathrm{c}} \\ w\cdot[k_{\mathrm{j}}-k(t)],\, k(t)>k_{\mathrm{c}} \end{cases} \tag{10}$$

which restricts predictions to physically feasible traffic regimes. where w is the backward wave speed, k_c is the critical density, k_j is the jam density, $q(t)$ is the flow rate at time t, and $k(t)$ is the density at time t. This restriction ensures that the predicted states remain within physically feasible regimes, preventing violations, such as flows exceeding capacity.

4) Feasibility parameterization. Nonnegativity and monotonicity can be preserved through parameterization strategies in the neural network outputs. For example, applying exponential or softplus activation functions ensures that flow and delay remain nonnegative, whereas parameter bounds maintain monotonic behavior near capacity.

Following the classical review by Karniadakis et al. (2021), three major pathways for embedding physics into neural networks are identified: (1) through the data (observational bias), (2) through the training process (learning bias), and (3) through the model structure (inductive bias). The four physics-informed constraints summarized in this study are not homogeneous but belong primarily to the second and third categories. Specifically, Types (A–C) correspond to the "training-level" integration, where physical residuals are incorporated into the loss function to guide optimization, whereas Type (D) aligns with the "model-structure" integration, embedding physics directly into activation functions or network architectures.

Moreover, physical constraints can be divided into two subcategories: Hard physical laws (e.g., conservation or continuity equations), which act as strict regularization terms ensuring physically consistent outputs, and soft empirical relations (e.g., fundamental diagrams or car-following relations), which serve as guiding priors that may introduce bias when the underlying empirical model is imperfect. This distinction echoes observations from Mo et al. (2021) and Long et al. (2024), who demonstrated that iterative calibration or refinement of empirical models is crucial when soft physical priors are integrated into PINNs. Table 8 summarizes how each constraint type corresponds to its embedding level (data, training, or structure) and physical category (hard or soft), enhancing conceptual clarity and theoretical consistency. Furthermore, by explicitly distinguishing between hard physical laws and soft empirical relations, this classification clarifies how different constraint types correspond to varying integration levels between physics and neural networks.

Compared with conventional ML approaches such as the LSTM, GNN, or transformer models, which can effectively reduce error metrics such as the RMSE in delay estimation or LPF fitting, PINNs exhibit a distinct advantage in terms of physical consistency. Purely data-driven models often violate conservation laws or generate unrealistic queue dynamics, limiting their transferability across networks and traffic regimes. By embedding residual terms derived from the LWR continuity equation, queue balance relationships, and FD consistency directly into the loss function, PINNs ensure that the predicted delay and capacity patterns remain physically interpretable.

Although the PINN method comes at the cost of higher training complexity and computational overhead, the resulting models demonstrate more robust and generalizable performance in delay estimation and LPF calibration tasks. This integration provides a principled mechanism to combine data-driven learning with traffic flow theory, retaining predictive flexibility while reducing the risk of physically inconsistent outcomes, such as negative delays or violations of capacity constraints.

Table 8 Integration of physics-informed constraints for delay modeling

Constraint type	Integration level	Nature of physics constraint	Description and role in model
Conservation constraint	Training-level	Hard law	Penalizes violation of vehicle conservation; prevents mass-gain/loss artifacts in predicted density and flow fields.
Queue balance constraint	Training-level	Hard law and soft behavioral relation	Ensures inflow-outflow consistency and physically meaningful queue evolution; anchors temporal dynamics near observed arrival/departure patterns.
Fundamental diagram consistency	Training-level	Soft empirical relation	Constrains predictions within feasible flow-density regimes; introduces domain-specific bias reflecting traffic equilibrium relations.
Structural/activation constraint	Model-structure	Hard/soft	Embeds physics into network design itself to enforce nonnegativity and capacity-bounded behavior without explicit loss terms.

4.3 Limitations of the PINNs

Despite their potential, the application of PINNs to traffic flow modeling remains subject to several limitations.

First, the computational cost is substantially greater than that of conventional ML models, as the evaluation of additional physics-based residuals (e.g., conservation and queue balance equations) introduces considerable overhead during training, particularly for large-scale networks. Second, the design of physics terms requires strong prior knowledge from the domain expert, and alternative formulations of the same constraint may lead to divergent performance outcomes. Third, PINNs exhibit a high degree of sensitivity to hyperparameters, such as the relative weighting between data-fitting and physics-based losses, learning rate schedules, and collocation point sampling strategies; improper configurations may result in unstable convergence or suboptimal solutions. Finally, PINNs often suffer from performance degradation under sparse or noisy data, which are prevalent in real-world sensing environments. These issues collectively highlight the need for robust training algorithms, adaptive loss-weighting mechanisms, and scalable implementations to make PINNs practical tools for transportation applications.

4.4 A unified AI-driven LPF framework

Rather than treating LPFs as fixed algebraic forms, future research will emphasize hybrid architectures that fuse physics-based traffic flow models with AI-driven learning. This fusion aims to produce formulations that are both interpretable and broadly applicable across network contexts. Building on these insights, we modify the framework proposed by Huo et al. (2022) and propose an original unified end-to-end AI-driven LPF pipeline, as shown in Fig. 11. The pipeline proceeds as follows:

1) Data preparation: Raw observations are split into 80% training and 20% test sets; normalization, scaling, and feature engineering are applied; and categorical or high-dimensional inputs are mapped through embedding layers.

2) For feature extraction, core variables such as the observed mean travel time, free-flow travel time, and nominal link capacity are derived.

3) Interpretability: Interpretability is strengthened through SHapley Additive exPlanations (SHAP) analysis and attention weight visualization to identify the relative importance of link attributes, exogenous factors (e.g., weather, incidents), and network topology. This allows the framework to provide not only accurate predictions but also actionable insights for planners.

4) Capacity estimation: This method employs GNNs, transformers, and attention mechanisms to ingest link attributes plus exogenous factors (e.g., weather, incidents) and predict a dynamic link capacity profile.

5) Iterative loop: Rather than treating capacity estimation and travel time prediction as isolated tasks, the pipeline adopts an iterative, physically informed modeling process. The predicted delays are fed back to update the flow distributions, which in turn refine the capacity estimation. This feedback loop operationalizes the dynamic coupling between demand, supply, and delay in real-world networks.

6) Model training and physical consistency: To ensure theoretical validity, we incorporate physics-informed regularizers into the training objective. Specifically, conservation-based residuals require that inflows equal outflows plus changes in queue length, monotonicity constraints guarantee that volume–delay curves remain nondecreasing, and nonnegativity constraints prevent unrealistic negative delays or capacities. These conditions are embedded into the loss function alongside conventional error metrics, thereby preserving physical consistency across predictions.

7) Travel-time prediction: Combine the predicted capacity, traffic volume, and free-flow time into a generalized LPF to compute time-dependent travel times.

8) Evaluation and robustness testing: Beyond standard accuracy metrics (the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE)), the framework incorporates uncertainty quantification through prediction intervals and conformal prediction methods. Robustness is further evaluated under distribution shifts, including cross-temporal and cross-network validation, to ensure generalizability across diverse traffic regimes.

This unified AI-driven LPF framework is specifically tailored to the unique challenges of LPF modeling. At its core, it enforces physical consistency, embeds capacity estimation and delay prediction into an iterative loop, and integrates uncertainty quantification with interpretability modules. As such, the framework moves beyond black-box learning to provide a physically grounded, reproducible, and extensible methodology. By embedding ML and AI into each stage of the process, from parameter calibration to travel time estimation, it enables LPFs to adapt to evolving traffic patterns, enhances predictive accuracy, and supports more effective congestion management and strategic transportation planning.

5 Case study

5.1 Data and study site

In this section, we select traffic data from the Los Angeles I-405 corridor, which exhibits both recurring and nonrecurring congestion patterns across freeway segments, providing a solid basis for evaluating the generalizability of the model. The dataset, obtained from the PeMS system, contains flow, speed, and occupancy measurements collected at 5-min intervals via loop detectors from April 1, 2017 to July 31, 2017, covering a total of 81 weekdays and excluding holidays and weekends. The spatial extent of the study corridor is shown in Fig. 12a, while the corresponding bottleneck speed heatmap is presented in Fig. 12b.

5.2 Performance metrics

To evaluate the performance of various traffic flow prediction models, it is crucial to define clear criteria that both measure prediction accuracy and guide methodological improvements. The evaluation process consists primarily of comparing the predicted traffic states with the observed real-time conditions. In this study, we focus on three widely used error metrics: MAE, RMSE, and MAPE.

The MAE quantifies the average magnitude of errors by computing the mean of the absolute differences between the observed and predicted values. Formally, it can be expressed as

$$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}|y_n - \widehat{y}_n| \tag{11}$$

where y_n is the measured traffic flow value for observation m, $\widehat{y}_n$ is the estimated traffic flow value for observation n, and N is the total number of flow counts.

The RMSE is defined as the square root of the average of the squared differences between the predicted values produced by the model and the corresponding observed values. Mathematically, it can be expressed as

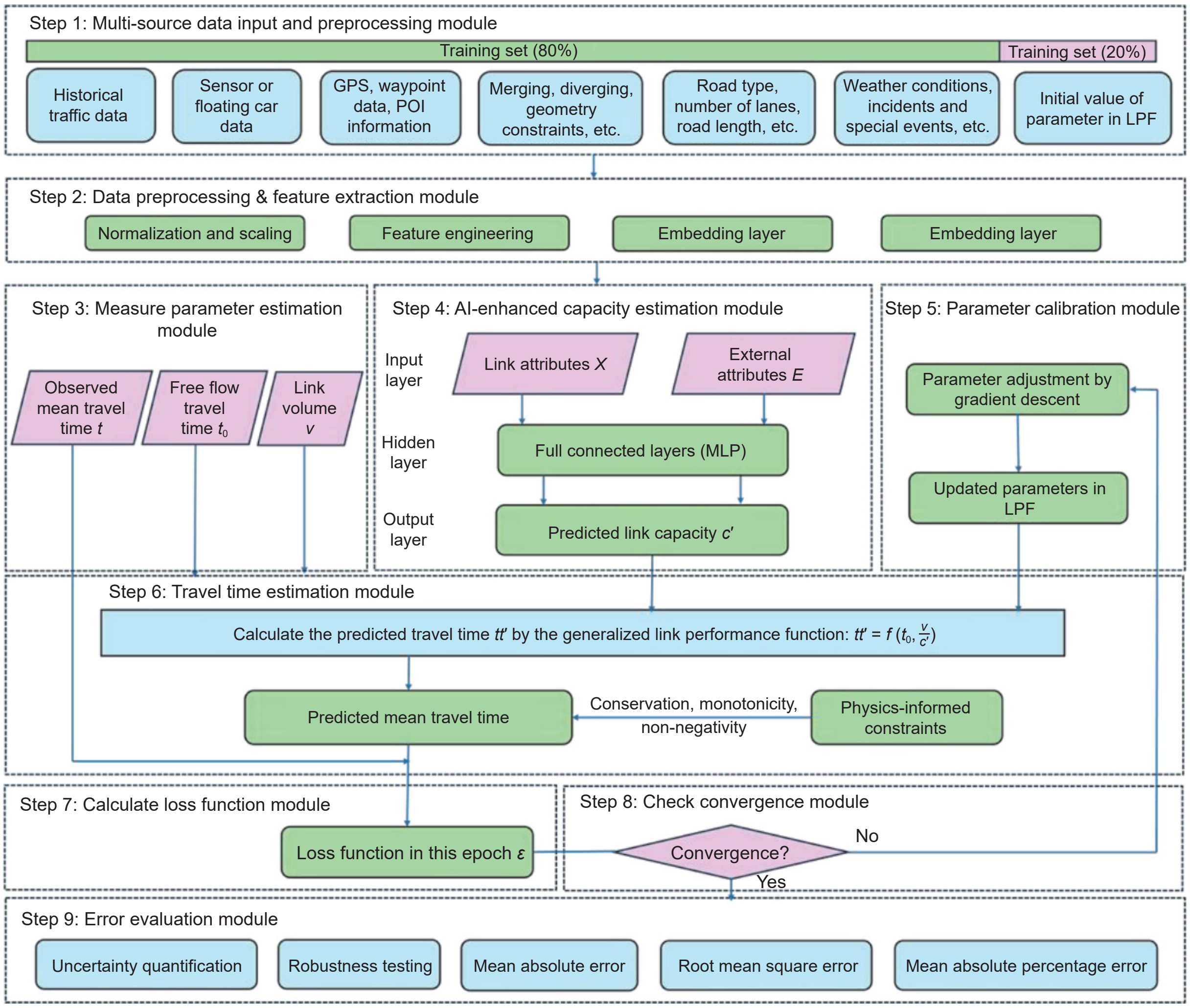


Fig. 11 Physics-informed neural network and LPF-based travel time prediction framework.

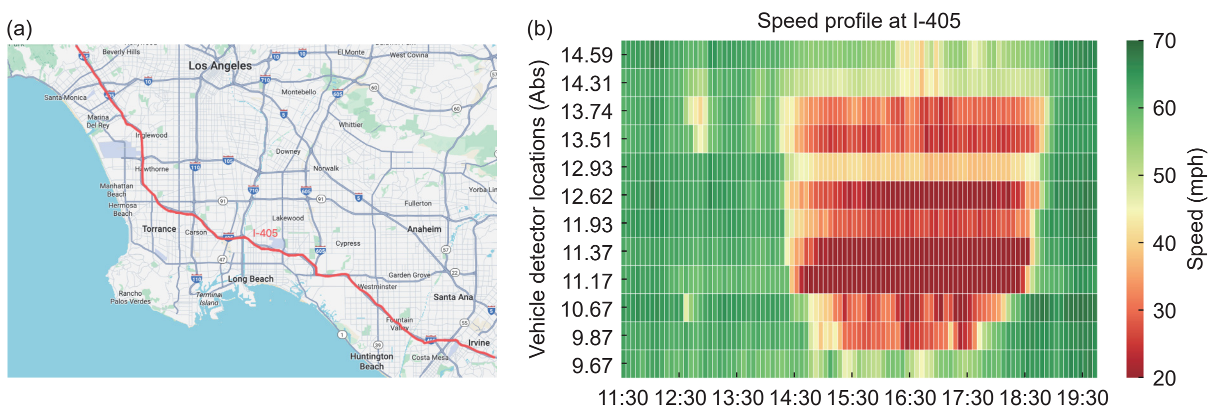


Fig. 12 Areas of Los Angeles I-405 corridors: (a) corridor location; (b) speed heatmap of the I-405 corridor.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{n=1}^{N} (y_n - \widehat{y}_n)^2}{N}} \tag{12}$$

The MAPE is frequently used in traffic forecasting because it helps reduce the disproportionate influence of instances with very low traffic flow on the overall error measurement. Unlike absolute error metrics, the MAPE incorporates both the magnitude of the error and its relative size by expressing the discrepancy between the predicted and observed values as a proportion of the observed value. Formally, it is defined as

$$\mathrm{MAPE} = \frac{100\%}{N} \sum_{n=1}^{N} \left| \frac{y_n - \widehat{y}_n}{y_n} \right| \tag{13}$$

5.3 Representative models for comparison

This case study is based on PeMS data from the I-405 corridor and compares different types of VDF models uniformly in the travel time dimension. The outputs of all the models are converted into travel times or delays through speed-time or flow-density relationships and then compared against the observed travel time series.

The study considers six representative models, as illustrated in Table 9, covering empirical, theory-driven, data-driven, and physics-informed approaches. The BPR function is simple and widely applied, but its accuracy is limited, and it lacks physical consistency. FD-based models (e.g., the Greenshields model) offer better interpretability and remain consistent with fundamental traffic flow principles, whereas fluid-queue models can further capture oversaturation and congestion duration with a strong physical foundation.

1) The classical BPR function, which is formulated as Eq. (5), is a flow-based model that can estimate travel time with the observed data. We also test the following density-based modified BPR function (Kucharski and Drabicki, 2017):

$$tt = t_f \cdot \left[1 + \alpha\left(\frac{k}{k_c}\right)^{\beta}\right] \tag{14}$$

where k is the density and k_c is the critical density.

2) The mathematical relationship in the Greenshield FD model is the speed-density relationship, which can be used to derive the travel time as

$$tt = \frac{L}{v} = \frac{L}{v_f \cdot \left(1 - \frac{k}{k_j}\right)} = \frac{t_f}{1 - \frac{k}{k_j}} \tag{15}$$

where L is the link length and v is the speed.

3) The time-dependent speed of the fluid queue-based volume-delay function can be represented as

$$v(t) = \frac{L}{\frac{L}{v_c} + w(t)} \tag{16}$$

If we use the quadratic PAQ model, then we have the time-dependent waiting time:

$$w(t) = \frac{\gamma}{3\mu} \cdot (t - t_0)^2 \cdot (t_3 - t) \tag{17}$$

The travel time can be derived as

$$tt = \frac{L}{v_c} + w(t) = \frac{L}{v_c} + \frac{\gamma}{3\mu} \cdot (t - t_0)^2 \cdot (t_3 - t) \tag{18}$$

The travel time of the fluid queue-based volume-delay function can be represented by the sum of the free-flow travel time and the waiting time as

$$tt = t_f + w = t_f \cdot \left[1 + \alpha\left(\frac{D}{C}\right)^{\beta}\right] \tag{19}$$

4) In the LSTM and transformer methods, we can convert the observed time series speed into the time series travel time and then input it into the data-driven method.

$$tt = \sum_i \frac{L_i}{v_i} \tag{20}$$

where L_i is the length of link i and v_i is the observed speed on link i.

5) Finally, to unify all the models under the "travel time/delay" dimension, we adopt a PINN. In addition to conventional data features (including flow, speed, density, and historical sequences), this method integrates traffic flow theory-based variables such as free-flow travel time, segment flow, estimated capacity, congestion duration, and queue length as additional inputs or feature engineering variables into the neural network.

A key feature of the proposed PINN is the explicit incorporation of physical constraints into the learning process. Specifically, the travel time is formulated as $tt \geq t_f$, ensuring that the predicted travel time never falls below the physically feasible lower bound. The additional delay component $\varphi_t(D, P, Q(t), C, X, E)$ is constrained to be nonnegative and to reflect dynamic traffic flow characteristics, such as the time-dependent queue length. While the neural network function $\varphi_t(\cdot)$ captures complex nonlinear dependencies among demand, congestion duration, queue dynamics, capacity, and exogenous factors, the overall structure remains consistent with queueing theory, thereby improving both predictive accuracy and physical interpretability.

$$tt = t_f \cdot [1 + \varphi_t(D, P, Q, C, X, E)] \tag{21}$$

where D is the demand, P is the congestion duration, Q is the queue length, C is the capacity, X represents the exogenous factors, E represents the environmental or error terms, and D, P, Q, and C are the positive.

Data-driven methods predict travel time by inputting observed temporal features such as flow, speed, density, time of day, and

Table 9 Performance comparison of empirical, theory-driven, data-driven, and physics-informed models

Model category	Representative method	Accuracy	Physical consistency	Interpretability	Remark
Empirical VDF	BPR	Low (fit simple trends, inaccurate under oversaturation)	Low (rely only on V/C, lack conservation and queue dynamics)	Medium (parameters α, β partially interpretable)	Simple and widely applied in planning practice
Theory-driven	FD-based (e.g., Greenshields model)	Medium (capture free-flow to critical flow)	High (consistent with traffic flow theory)	High (parameters v_f,k_j, w have clear physical meaning)	Effective for capacity estimation and regime distinction
	Fluid-queue based (e.g., Newell model)	High (accurate for oversaturated travel time curves)	Very high (explicitly models queues, congestion duration P and μ)	High (parameters μ, P, D are directly interpretable)	Emphasizes congestion duration as the main driver of delay
Data-driven	LSTM	High (strong short-term prediction)	Medium (no explicit physical constraints)	Medium (black-box with limited interpretability)	Sensitive to data distribution shifts
	Transformer	Very high (long-sequence modeling is strong)	Low-medium (may violate physical rules without constraints)	Medium (attention weights provide partial interpretability)	Data-hungry, computationally costly
Physics-informed	PINN	High (balance between prediction and constraints)	Very high (conservation, capacity and nonnegativity embedded)	High (physical terms interpretable, flexible prediction)	High training cost, prior knowledge needed, hyperparameter-sensitive

day of week and then transforming point-based speed/flow data into link-level travel time through spatial aggregation. LSTM is well suited for capturing short-term temporal dependencies and can effectively predict dynamic fluctuations in travel time, whereas the transformer excels at modeling long sequences, accounts for spatiotemporal dependencies across multiple road segments, and is suitable for large-scale prediction. These methods are capable of capturing complex nonlinear relationships and handling high-dimensional inputs (e.g., speed, flow, weather). However, their limitation lies in the lack of physical constraints, which may lead to unrealistic results (such as negative delays or excessively high speeds). In contrast, the core idea of the PINN method is to embed traffic flow physical principles (e.g., fundamental diagram relations, queueing delay constraints) directly into the loss function of the neural network, enabling the model to learn the mapping between speed/flow and travel time while ensuring physical consistency in predictions. This approach combines the high predictive accuracy of data-driven models with the interpretability and rationality of theoretical models, thereby achieving more robust travel time prediction under limited data and complex traffic conditions.

5.4 Results

The dataset is first divided chronologically into a training set and a testing set. For the BPR volume-based, density-based and fluid queue-based methods, the unknown parameters α and β are calibrated on a daily basis via observed data from the training period. The resulting daily estimates are then robustly aggregated to obtain a stable set of link-specific parameters, which are considered representative of the link when accounting for observational variability. These aggregated parameters are subsequently applied to the testing set to generate travel times, which are compared against observations to evaluate the predictive performance via the MAE, RMSE, and MAPE.

For the FD-based model, the free-flow speed $v_{\rm f}$ and jam density $k_{\rm j}$ are estimated each day from density-speed data during the training period. The density values are then mapped to speed and travel time via the fitted Greenshields function. Robust aggregation of daily parameter estimates yields stable values of $v_{\rm f}$ and $k_{\rm j}$, which are applied to the testing set to compute the estimated travel time, and the corresponding errors are assessed.

For data-driven sequence models such as LSTM and Transformer, the training samples are generated via a sliding-window approach, where a sequence of past observations with sequence length is mapped to the next time step. Normalization is performed on the basis solely of training set statistics. After training, rolling prediction is carried out on the testing set to generate travel time trajectories, which are compared with the ground truth. For the PINN framework, model training is conducted on the training set by minimizing a composite loss that integrates data fidelity, physics-based residuals, and boundary/initial conditions. An out-of-sample evaluation of the testing set is performed via the selected error metrics. Table 10 summarizes the predictive performance of different models in terms of the MAE, RMSE, and MAPE.

Table 10 Comparative prediction performance of different models

Model	MAE	RMSE	MAPE (%)
BPR volume-based	4.335	7.959	16.4600
BPR density-based	0.629	0.965	3.343
Fluid-queue based	1.912	3.079	9.709
FD-based	1.540	1.973	8.359
LSTM neural network	0.480	0.988	2.110
Transformer	0.296	0.471	1.377
PINN (proposed)	0.077	0.176	0.261

Fig. 13a presents the evaluation performance comparisons across different methods, including the metrics of the MAE, RMSE, and MAPE. The classical BPR volume-based model, with an MAE of 4.335, applies a single V/C ratio to produce a constant speed value across the entire time interval, which results in poor performance when compared against time-varying observed travel times. Consequently, Fig. 13b does not include the classical BPR volume-based results, as they would appear as a flat horizontal line that cannot capture the dynamic variations shown in the time-dependent travel time profiles.

The fluid-queue model with an MAE of 1.912 demonstrates the theoretical advantage of embedding congestion duration and D/C ratios, producing structurally consistent time-dependent travel times through its analytical framework. While its MAE appears higher, this model provides physically meaningful insights into congestion dynamics that purely empirical approaches cannot offer.

The density-based BPR modification, with an MAE of 0.629, and the FD-based model, with an MAE of 1.540, are still time series generators that use a general demand-to-speed form to produce speed from immediate observations. LSTM, which achieves an MAE of 0.480, and the transformer, with an MAE of 0.296, are more sophisticated time series generators that can extrapolate from immediate observations and make short-term predictions. The high accuracy of these methods reflects their ability to capture temporal patterns but always raises questions about their transferability to different networks or demand scenarios.

The proposed PINN framework, with an MAE of 0.077, an RMSE of 0.176, and a MAPE of 0.261, demonstrates the potential of integrating physical constraints with learning algorithms. However, sensitivity testing is needed to assess the impact reasonableness of D/C-to-speed and queue-to-speed relationships. The functional forms with many interrelated parameters (D/C ratios, queue dynamics, and external factors) suggest that more customized and time-dependent formulations of time-dependent travel times deserve future research. The better utilization of external factors to explain model results and the incorporation of short-term time series to explain how the model absorbs time-dependent demand variations remain important research directions.

Fig. 13b illustrates these paradigmatic differences. The fluid queue-based method produces smooth, physically consistent travel time profiles critical for understanding system behavior and policy evaluation. Data-driven methods excel at reproducing observed patterns but could struggle with scenario analysis where demand patterns differ from historical observations. The PINN framework offers a promising direction, although further development is needed to fully realize its potential for unified modeling.

6 Future research

The role of VDFs in transportation research and practice is critical, as highlighted by this review. This study offers a comprehensive analysis of the theoretical fundamentals, practical deployment, and emerging applications of VDFs. While these findings highlight significant advancements, they also reveal persistent gaps that limit the full potential of VDFs in addressing evolving transportation challenges. To provide actionable guidance, we structure the future research agenda into three layers of priority: (1) short-term priorities in calibration and congestion identification with multisource data to improve operational reliability and reproducibility, (2) medium-term priorities in

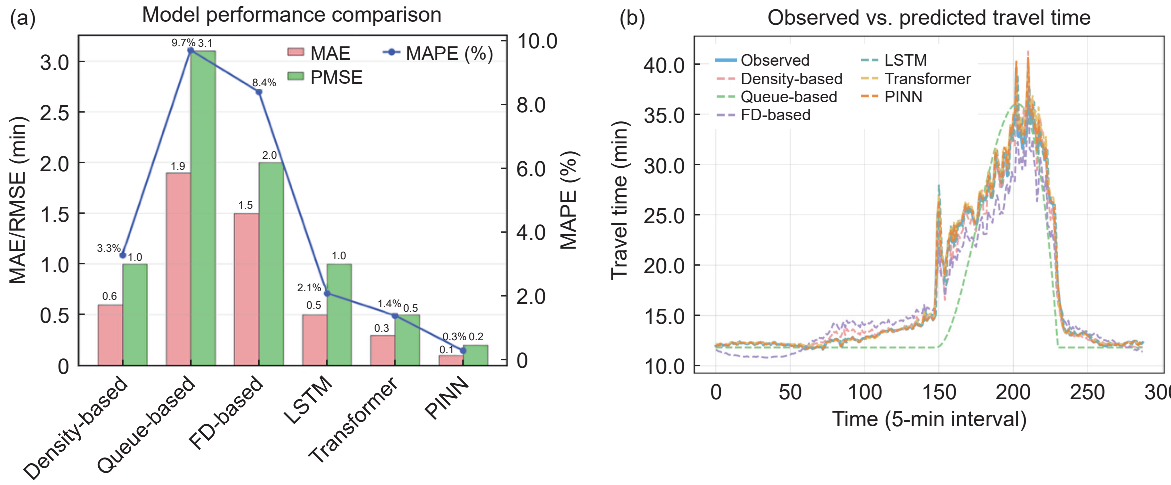


Fig. 13 Comparison of model performance and travel time prediction: (a) model performance comparison and (b) comparison between observed and predicted travel times in the time series.

incorporating reliability and stochasticity into VDFs to ensure robustness across varying traffic regimes, and (3) long-term priorities in leveraging AI-driven and ML-enhanced VDFs to enable adaptive, scalable, and real-time applications in intelligent transportation systems.

Each layer is detailed below, with specific research questions and methodological pathways.

1) VDF calibration with congestion identification and multisource data: The development of reliable VDFs for operational planning demands a deeper understanding of bottleneck congestion dynamics, queue formation, and the variability of traffic flow. We pose the following future research questions (FRQs) to be explored in future research.

FRQ 1: How should current VDFs be enhanced to incorporate queue spillback effects and downstream constraints?

Pathway: Integrate cumulative arrival–departure curves and shockwave-based models into LPF calibration, embedding spillback as boundary conditions.

FRQ 2: In what ways can the breakdown probability be formulated as a function of the D/C ratio and inflow demand curvature to support reliability-aware VDFs?

Pathway: Use stochastic capacity formulations and survival analysis techniques to link D/C ratios with breakdown probability distributions.

FRQ 3: Can advanced data-driven methods, such as deep learning or Bayesian inference, improve real-time VDF calibration while quantifying uncertainty from heterogeneous data inputs?

Pathway: Bayesian neural networks and variational inference are employed to capture parameter uncertainty, combined with online learning for real-time updates.

2) VDFs with reliability considerations: Building upon the calibration foundation, it is essential to integrate reliability considerations into VDFs to ensure robustness across varying traffic conditions and demand scenarios. VDFs that reflect temporal and spatial variability in travel time can support planning tools such as congestion pricing, adaptive signal control, and active demand management. Additionally, incorporating reliability into VDFs enables system-level performance evaluation by linking link-level variability with path- or OD-level indicators.

We propose the following FRQ to be explored in future research:

FRQ 4: How can traffic flow variability and reliability metrics (e.g., travel time standard deviation, trip rate distribution) be consistently integrated into VDF formulations for both recurring and nonrecurring conditions?

Pathway: Extend VDFs with stochastic terms calibrated from empirical reliability metrics and apply Monte Carlo simulation to capture nonrecurring disruptions.

FRQ 5: How can reliability-aware VDFs be scaled up to the OD and route levels to assess system-wide performance, and what are the implications for equilibrium modeling under stochastic demand?

Pathway: Embed uncertainty assessment methods into the calibration process and integrate online updating to enable rapid response to dynamic traffic conditions.

FRQ 6: How do within-day and day-to-day demand fluctuations impact travel time reliability, and how can VDFs be adapted to account for these temporal dynamics?

Pathway: This pathway employs time-varying stochastic processes and day-to-day dynamic traffic assignment frameworks to explicitly capture demand variability.

3) Machine learning-enhanced VDF modeling for intelligent transportation systems: The emergence of high-resolution traffic data and ML tools presents new opportunities for dynamic and scalable VDF modeling. However, challenges remain in data heterogeneity, anomaly handling, computational scalability, and real-time system integration. Efficiently leveraging ML, deep learning, and connected vehicle data can significantly enhance VDF accuracy and responsiveness, especially under dynamic traffic conditions.

We propose the following FRQ to be explored in future research:

FRQ 7: How can data fusion algorithms be developed to integrate heterogeneous data sources effectively (e.g., probe data, sensors, CAVs) for VDF training and calibration?

Pathway: Develop graph-based data fusion and tensor decomposition methods to align temporal and spatial dimensions across diverse sources.

FRQ 8: How can graph-based learning methods (e.g., GNNs) and distributed computing architectures be employed to scale ML-based VDF models to large urban networks while maintaining accuracy and computational efficiency?

Pathway: Adopt distributed GNN frameworks and parallel computing platforms to achieve scalability in metropolitan-scale networks.

FRQ 9: How can VDFs be dynamically adapted to reflect the real-time impact of emerging technologies such as CAVs, and how can reinforcement learning be leveraged to integrate VDF modeling with real-time traffic control decisions?

Pathway: Implement reinforcement learning agents that adjust

VDF parameters online, informed by streaming CAV and IoT data.

As transportation systems continue to evolve in the digital era, VDF research must move beyond generic problem statements toward prioritized and technically grounded solutions. Our proposed roadmap emphasizes (1) robust calibration using multisource data, (2) the integration of reliability metrics and stochasticity for scalable planning, and (3) AI-driven adaptivity for real-time applications. Collectively, these pathways define a forward-looking, physics-informed, and data-enhanced research agenda that is both actionable and innovative.

Appendix A

Table A1 Notations used in this study

Symbol	Definition
v	Speed
$\bar{v}$	Mean speed (km/h)
v_{f}	Free-flow speed (km/h)
v_{CO}	Cutoff speed
v_{c}	Critical speed
k	Density
k_{c}	Critical density (veh·km^{-1}·ln^{-1})
k_{j}	Jam density (veh·km^{-1}·ln^{-1})
w	Backward wave speed (km/h)
V	Hourly volume
C	Roadway hourly capacity
C_{p}	Lower critical capacity (onset of congestion threshold), below this, travel time grows only slightly with volume
C_{u}	Upper capacity threshold, beyond this, traffic enters the oversaturated region with severe queuing
μ	Constant discharge rate/link's service rate (veh·h^{-1}·ln^{-1})
D	Demand volume (during congestion) (veh·ln^{-1})
V/C	Volume-to-capacity ratio
D/C	Demand-to-capacity ratio
V^{H}	Human-driven hourly volume
V^{A}	Autonomous vehicle hourly volume
P	Congestion duration (h)
T	Analysis flow period
T_{f}	Flow period (the analysis period, e.g., 15 min, 1 h)
tt	Average travel time per unit distance
t_{f}	Free-flow travel time
t_{p}	Base (or free-flow) travel time, corresponding to low-volume conditions
t_{u}	Travel time at the end of the capacity interval
t_0	Start time of congestion period
t_1	The time of the highest inflow rate
t_2	The time of equal inflow and outflow rates during the congestion period
t_3	End time of congestion period
t_{s}	Start time of analysis period
t_{e}	End time of analysis period
$q(t)$	Time-dependent flow rate
$k(t)$	Time-dependent density
$Q(t)$	Time-dependent queue length
$w(t)$	Time-dependent waiting time
$tt(t)$	Time-dependent travel time
$\lambda(t)$	Time-dependent inflow rate
$\mu(t)$	Time-dependent outflow rate
$A(t)$	Cumulative arrival counts
$D(t)$	Cumulative departure counts
$\lambda(t_1)$	The highest inflow rates during the congestion period
$\overline{w}$	Average delay during the congestion period
g	Effective green time
g/C	As the proportion of the cycle which is effectively green for the phase under consideration
x	The degree of saturation, traffic intensity

Table A1 (Continued)

Symbol	Definition
ρ	Utilization rate as a queueing system
γ	Inflow demand curvature parameter
$\alpha, \beta, J, J_{\mathrm{A}}$	Parameters in VDFs
$g(m)$	Conversion factor in polynomial arrival queue models
d	Minimum gap (m)
τ	Reaction time (s)
f_{p}	Constant in elasticity function for mapping congestion duration to the magnitude of speed reduction
f_{d}	Constant in elasticity function for mapping D/C ratio to congestion duration
n	Elasticity coefficient of congestion duration in response to D/C charges, i.e., oversaturation-to-duration elasticity
s	Elasticity coefficient of speed reduction magnitude in response to congestion duration changes, i.e., duration-to-speed reduction elasticity

Appendix B

Summary of the key VDFs and related publications

Table B1 Most representative leading VDFs with the publication papers

Author	Paper title	Function	Remark
Webster (1958)	Traffic signal settings	$d=\dfrac{C(1-\lambda)^2}{2(1-\lambda x)}+\dfrac{x^2}{2V(1-x)}-0.65\left(\dfrac{C}{V^2}\right)^{\frac{1}{3}}x^{(2+5\lambda)}$	Queueing model-based, $\lambda=g/C$ ratio
Chicago Area Transportation Study (CATS) (1960)	Data projections	$tt=t_{\mathrm{f}}\cdot 2^{\left(\frac{V}{C}\right)}$	—
Smock (1962)	An iterative assignment approach to capacity restraint on arterial networks	$tt=t_{\mathrm{f}}\cdot e^{\left(\frac{V}{C}-1\right)}$	—
BPR (1964)	Traffic Assignment Manual	$tt=t_{\mathrm{f}}\left[1+\alpha\left(\dfrac{V}{C}\right)^{\beta}\right]$	—
Davidson (1966)	A flow-travel time relationship for use in transportation planning	$tt=t_{\mathrm{f}}\left[1+j\cdot\dfrac{V}{C-V}\right]$	Single queueing with random arrival rates and exponentially distributed service rates
Davidson (1978)	The theoretical basis of a flow-travel time relationship for use in transportation planning	$tt=t_{\mathrm{f}}\left[1+J\cdot\dfrac{V}{C-V}\right], J=\dfrac{k+1}{2k}$	Queue system with random arrivals and an Erlang service distribution
Newell (1982)	Applications of queueing theory	$tt=t_{\mathrm{f}}\left[1+\dfrac{\gamma}{36\mu\cdot t_0}\cdot\left(\dfrac{D}{\mu}\right)^{3}\right]$	Fluid-based queue with quadratic arrival rates and fixed inflow curvature parameter γ
Small (1983)	The incidence of congestion tolls on urban highways	$tt=\begin{cases} t_f, V\le C \\ t_{\mathrm{f}}+\dfrac{1}{2}P\cdot\left(\dfrac{V}{C}-1\right), V>C\end{cases}$	Peak hour-based
Spiess (1990)	Conical volume-delay functions	$tt=t_{\mathrm{f}}\cdot\left[2+\sqrt{\alpha^2\left(1-\dfrac{V}{C}\right)^2+\beta^2}-\alpha\left(1-\dfrac{V}{C}\right)-\beta\right]$	Well-defined VDF properties proposed
Akçelik (1991)	Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time function	$tt=t_{\mathrm{f}}+0.25T\cdot\left[\left(\dfrac{V}{C}-1\right)+\sqrt{\left(\dfrac{V}{C}-1\right)^2+\dfrac{8JV}{TC^2}}\right]$	Time-dependent form, derived by using the coordinate transformation technique
Huntsinger and Rouphail (2011)	Bottleneck and queuing analysis: Calibrating volume-delay functions of travel demand models	$tt=\begin{cases} t_{\mathrm{f}}\left[1+\alpha\left(\dfrac{V}{C}\right)^{\beta}\right], V\le C \\ t_{\mathrm{f}}\left[1+\alpha\left(\dfrac{D}{C}\right)^{\beta}\right], V>C\end{cases}$ $D=$demand at capacity + queue	Queued demand-based

Table B1 (Continued)

Author	Paper title	Function	Remark
Moses et al. (2013)	Development of speed models for improving travel forecasting and highway performance evaluation	$tt = \begin{cases} t_f\left[1+\alpha(C)^{\beta}\right], V \le C \\ t_f\left[1+\alpha\left(\frac{D}{C}\right)^{\beta}\right], V > C \end{cases}$ $D = C + (C - V)$	—
Bliemer et al. (2014)	Quasidynamic traffic assignment with residual point queues incorporating a first order node model	$tt = \sum_{a\varepsilon A} \frac{L_a}{v_f^a} \cdot \delta_{ap} + \tau_p^{\text{queue}},\ \forall a\varepsilon A$	Focus on incorporating hard capacity constraints and residual queues (including point queues) to effectively simulate congestion.
Kucharski and Drabicki (2017)	Estimating macroscopic volume delay functions with the traffic density derived from measured speeds and flows	$tt = t_f\left[1+\alpha\left(\frac{k}{k_c}\right)^{\beta}\right]$	Density as a proxy of V/C
Lazar et al. (2020)	Routing for traffic networks with mixed autonomy	$tt = t_f\left[1+\alpha\left(\frac{V^H+V^A}{C}\right)^{\beta}\right]$	Mixed flow-based
Cheng et al. (2022)	Estimating key traffic state parameters through parsimonious spatial queue models	$tt = t_f\left[1+\frac{\gamma \cdot g(m)}{\mu \cdot t_f} \cdot \left(\frac{D}{\mu}\right)^{4}\right]$	Fluid-based queue with cubic arrival rates and fixed inflow curvature parameter γ
Zhou et al. (2022)	A meso-to-macro cross-resolution performance approach for connecting polynomial arrival queue model to volume-delay function with inflow demand-to-capacity ratio	$tt = t_f\left[1+\alpha\left(\frac{D}{C}\right)^{\beta}\right]$ $\alpha = \theta f_p \cdot f_d^{\,s}$ $\beta = ns$ $\theta = \overline{w}/w_{t_2}$	Fluid based queue with a family of inflow curvature parameter values, derived from two elasticity forms and D/C ratio

Acknowledgements

The author Xuesong (Simon) Zhou would like to extend sincere thanks to Dr. David Levinson at the University of Sydney for many insightful comments. The author Xuesong (Simon) Zhou gratefully acknowledges the guidance from Prof. David Boyce, the feedback from Dr. Baloka Belezamo at Arizona Department of Transportation, Dr. Livshits, Dr. Arup Dutta, Dr. Wang Zhang, and Dr. Haidong Zhu from the Maricopa Association of Governments (MAG).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Akçelik, R., 1996. Relating flow, density, speed and travel time models for uninterrupted and interrupted traffic. *Traffic Eng Control*, **37**, 511–516.

Akçelik, R., 1991. Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time function. *Aust Road Res*, **21**, 3.

Akçelik, R., 1988. The highway capacity manual delay formula for signalized intersections. *Ite J Inst Transp Eng*, **58**, 23–27.

Akçelik, R., 1980. Time-dependent expressions for delay, stop rate and queue length at traffic signals. Ausralian Road Research Board, Vermont South, Australia, Internal Report AIR 367-1, 1–24.

Akçelik, R., Rouphail, N. M., 1993. Estimation of delays at traffic signals for variable demand conditions. *Transp Res Part B Methodol*, **27**, 109–131.

Anupriya, Graham, D.J., Bansal, P., Hörcher, D., Anderson, R., 2023. Optimal congestion control strategies for near-capacity urban metros: Informing intervention via fundamental diagrams. *Phys Stat Mech Its Appl*, **609**, 128390.

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., Zhang, K., 2016. POLARIS: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. *Transp Res Part C Emerg Technol*, **64**, 101–116.

Azizi, L., Beagan, D. 2022. Inclusion of Reliability in the Volume Delay Function. No. TRBAM-22-04192.

Azizi, L., Beagan, D. 2023. Testing a New Volume Delay Function. In: International Conference on Transportation and Development 2023, 25-33.

Bagherian, M., Hickman, M., Tavassoli, A., 2017. Considering the impact of precipitation on the accuracy of delay-function parameters, In: Proceedings from the 39th Australasian Transport Research Forum, 1–14.

Balmer, M., Meister, K., Rieser, M., Nagel, K., Axhausen, K.W., 2008. Agent-based simulation of travel demand: Structure and computational performance of MATSim-T. In: 2nd TRB Conference on Innovations in Travel Modeling, 504.

Banks, J. H., 1991. Two-capacity phenomenon at freeway bottlenecks: A basis for ramp metering? *Transp Res Rec*, 1320, 83–90.

Barceló, J., Casas, J., 2005. Dynamic network simulation with AIMSUN. *Simulation Approaches in Transportation Analysis*. New York, USA: Springer-Verlag, 57–98.

Bayram, V., Yaman, H., 2018. Shelter location and evacuation route assignment under uncertainty: A benders decomposition approach. *Transp Sci*, **52**, 416–436.

Beckmann, M. J., McGuire, C. B., Winsten, C. B., 1956. *Studies in the Economics of Transportation*. New Haven, USA: Yale University Press, 226.

Behrisch, M., Bieker, L., Erdmann, J., Krajzewicz, D., 2011. SUMO-Simulation of urban mobility. https://www. eclipse. org/sumo

Belezamo, B., 2020. Data-driven Methods for Characterizing Transportation System Performances Under Congested Conditions: A Phoenix Study. Ph.D. dissertation. Arizona, USA: Arizona State University.

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H., Mishalani, R., 1998. DynaMIT: a simulation-based system for traffic prediction. In: DACCORD short term forecasting workshop, 12, 1–12.

Benedek, C. M., Rilett, L. R., 1998. Equitable traffic assignment with envi-

ronmental cost functions. *J Transp Eng*, **124**, 16–22.
Berman, O., Larson, R. C., Parkan, C., 1987. The stochastic queue p-Median problem. *Transp Sci*, **21**, 207–216.
Bertsekas, D.P., Gallager, R.G., 2021. *Data Networks*. Amsterdam, the Netherlands: Athena Scientific.
Bliemer, M. C. J., Raadsen, M. P. H., 2020. Static traffic assignment with residual queues and spillback. *Transp Res Part B Methodol*, **132**, 303–319.
Bliemer, M. C. J., Raadsen, M. P. H., Brederode, L. J. N., Bell, M. G. H., Wismans, L. J. J., Smith, M. J., 2017. Genetics of traffic assignment models for strategic transport planning. *Transp Rev*, **37**, 56–78.
Bliemer, M. C. J., Raadsen, M. P. H., Smits, E. S., Zhou, B., Bell, M. G. H., 2014. Quasi-dynamic traffic assignment with residual point queues incorporating a first order node model. *Transp Res Part B Methodol*, **68**, 363–384.
Boyce, D. E., Janson, B. N., Eash, R. W., 1981. The effect on equilibrium trip assignment of different link congestion functions. *Transp Res Part A Gen*, **15**, 223–232.
Boyles, S., Ukkusuri, S.V., Waller, S.T., Kockelman, K.M., 2008. A comparison of static and dynamic traffic assignment under tolls in the Dallas-Fort Worth region. In: Transportation Research Board Conference Proceedings, 114–117.
Branston, D., 1976. Link capacity functions: A review. *Transp Res*, **10**, 223–236.
Brederode, L., Pel, A., Wismans, L., de Romph, E., Hoogendoorn, S., 2019. Static Traffic Assignment with Queuing: Model properties and applications. *Transp A Transp Sci*, **15**, 179–214.
Bureau of Public Roads (BPR), 1964. *Traffic Assignment Manual for Application with a Large, High-Speed Computer*. U.S. Department of Commerce, Urban Planning Division, Washington, D.C.
Caliper, 2009. TransModeler traffic simulation software-version 2.5 user's guide.
Campbell, E. W., 1968. The Transportation System: An Evaluation of Alternative Land Use and Transportation Systems in the Chicago Area. https://onlinepubs.trb.org/Onlinepubs/hrr/1968/238/238-006.pdf
Carey, M., 2004. Link travel times I: Desirable properties. *Netw Spatial Econ*, **4**, 257–268.
Carey, M., Humphreys, P., McHugh, M., McIvor, R., 2014. Extending travel-time based models for dynamic network loading and assignment, to achieve adherence to first-in-first-out and link capacities. *Transp Res Part B Methodol*, **65**, 90–104.
Cassidy, M. J., Bertini, R. L., 1999. Some traffic features at freeway bottlenecks. *Transp Res Part B Methodol* **33**, 25–42.
Chang, J.S., Mackett, R. L., 2006. A bi-level model of the relationship between transport and residential location. T*ransp Res Part B Methodol* **40**, 123–146.
Charnes, A., Cooper, W. W., 1959. Chance-constrained programming. *Manag Sci*, **6**, 73–79.
Charnes, A., Cooper, W. W., Symonds, G. H., 1958. Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Manag Sci*, **4**, 235–263.
Chen, A., Zhou, Z., Chootinan, P., Ryu, S., Yang, C., Wong, S. C., 2011. Transport network design problem under uncertainty: A review and new developments. *Transp Rev*, **31**, 743–768.
Chen, Q., Li, X., Ouyang, Y., 2011. Joint inventory-location problem under the risk of probabilistic facility disruptions. *Transp Res Part B Methodol*, **45**, 991–1003.
Cheng, Q., Liu, Z., Guo, J., Wu, X., Pendyala, R., Belezamo, B., et al., 2022. Estimating key traffic state parameters through parsimonious spatial queue models. *Transp Res Part C Emerg Technol*, **137**, 103596.
Cheng, Q., Liu, Z., Lin, Y., Zhou, X., 2021. An s-shaped three-parameter (S3) traffic stream model with consistent car following relationship. *Transp Res Part B Methodol*, **153**, 246–271.
Cheng, J., Li, G., Chen, X., 2019. Developing a travel time estimation method of freeway based on floating car using random forests. *J Adv Transp*, **2019**, 8582761.
Chicago Area Transportation Study (CATS), 1960. Data projections. . Chicago Area Transportation Study, Chicago, IL.
Chiu, Y. C., Zhou, L., Song, H., 2010. Development and calibration of the Anisotropic Mesoscopic Simulation model for uninterrupted flow facilities. *Transp Res Part B Methodol*, **44**, 152–174.
Correa, J. R., Schulz, A. S., Stier-Moses, N. E., 2004. Selfish routing in capacitated networks. *Math Oper Res*, **29**, 961–976.
Crainic, T. G., Florian, M., Léal, J. E., 1990. A model for the strategic planning of national freight transportation by rail. *Transp Sci*, **24**, 1–24.
Cui, Z., Ke, R., Pu, Z., Ma, X., Wang, Y. (2020). Learning traffic as a graph: A gated graph wavelet recurrent neural network for network-scale traffic prediction. Transp Res C: Emerg Technol, 115, 102620.
Daganzo, C. F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transp Res Part B Methodol*, **28**, 269–287.
Daganzo, C. F., Geroliminis, N., 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. *Transp Res Part B Methodol*, **42**, 771–781.
Davidson, K.B., 1978. The theoretical basis of a flow-travel time relationship for use in transportation planning. In: Australian Road Research Board (ARRB) Conference, 32–35.
Davidson, K.B., 1966. A flow travel time relationship for use in transportation planning. In: Australian Road Research Board (ARRB) Conference, 94–183.
Deshpande, N., Park, H. J., 2025. Physics-informed deep learning with Kalman filter mixture for traffic state prediction. *Int J Transp Sci Technol*, **17**, 161–174.
Dion, F., Rakha, H., Kang, Y. S., 2004. Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. *Transp Res Part B Methodol*, **38**, 99–122.
Doll, A., Abbasi, M., Zhao, M., Zhou, X. S., 2024. Oversaturated intersections: A real-world assessment of polynomial fluid queue models. *Phys A Stat Mech Appl*, **651**, 129864.
Dong, J., Mahmassani, H. S., 2009. Flow breakdown and travel time reliability. *Transp Res Rec J Transp Res Board*, **2124**, 203–212.
Dowling, R., Skabardonis, A., 1993. Improving average travel speeds estimated by planning models. *Transp Res Rec*, **1366**, 68–74.
Dowling, R. G., Singh, R., Cheng, W., 1998. Accuracy and performance of improved speed-flow curves. *Transp Res Rec J Transp Res Board*, **1646**, 9–17.
Dowling, R., Margiotta, R., Cohen, H., Skabardonis, A., Elias, A., 2011. Methodology to evaluate active transportation and demand management strategies. *Procedia Soc Behav Sci*, **16**, 751–761.
Duan, Y., L V, Y., Wang, F. Y., 2016. Travel time prediction with LSTM neural network. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), 1053–1058.
Duffin, R. J., 1947. Nonlinear networks. *IIa. Bull Amer Math Soc*, **53**, 963–971.
Engelson, L., van Amelsfort, D., 2015. The role of volume-delay functions in forecasting and evaluating congestion charging schemes: The Stockholm case. *Transp Plan Technol*, **38**, 684–707.
Fambro, D. B., Rouphail, N. M., 1997. Generalized delay model for signalized intersections and arterial streets. *Transp Res Rec J Transp Res Board*, **1572**, 112–121.
Fellendorf, M., Vortisch, P., 2010. Microscopic traffic flow simulator VISSIM. In: *Fundamentals of Traffic Simulation*. New York: Springer New York, 63–93.
Fisk, C. S., 1991. Link travel time functions for traffic assignment. *Transp Res Part B Methodol*, **25**, 103–113.
Fosgerau, M., Small, K. A., 2012. Marginal congestion cost on a dynamic expressway network. *J Transp Econ Policy*, **46**, 431–450.
Gayah, V. V., Daganzo, C. F., 2011. Clockwise hysteresis loops in the Macroscopic Fundamental Diagram: An effect of network instability. *Transp Res Part B Methodol*, **45**, 643–655.
Geroliminis, N., Sun, J., 2011. Properties of a well-defined macroscopic fundamental diagram for urban traffic. *Transp Res Part B Methodol*, **45**, 605–617.

Genser, A., Kouvelas, A., 2022. Dynamic optimal congestion pricing in multi-region urban networks by application of a Multi-Layer-Neural network. *Transp Res Part C Emerg Technol*, **134**, 103485.

Guerrieri, M., 2024. A theoretical model for evaluating the impact of connected and autonomous vehicles on the operational performance of turbo round abouts. *Int J Transp Sci Technol*, **14**, 202–218.

Guo, L., Huang, S., Sadek, A. W., 2013. An evaluation of environmental benefits of time-dependent green routing in the greater buffalo–Niagara region. *J Intell Transp Syst*, **17**, 18–30.

Hadi, M., Zhou, X., Hale, D., 2022. Multiresolution modeling for traffic analysis: Guidebook. (No. FHWA-HRT-22-055). United States. Federal Highway Administration.

Hall, F.L., Agyemang-Duah, K., 1991. Freeway capacity drop and the definition of capacity. *Transp Res Rec*, **1320**, 91–98.

Hajbabaie, A., Tajalli, M., Bardaka, E., 2024. Effects of connectivity and automation on saturation headway and capacity at signalized intersections. *Transp Res Rec J Transp Res Board*, **2678**, 31–46.

Han, Y., Wang, M., Li, L., Roncoli, C., Gao, J., Liu, P., 2022. A physics-informed reinforcement learning-based strategy for local and coordinated ramp metering. *Transp Res Part C Emerg Technol*, **137**, 103584.

Han, Y., Wu, J., Ding, F., Li, Z., Liu, P., Leclercq, L., 2025. Capacity drop at active bottlenecks: An empirical study based on trajectory data. *Transportation Research Part B: Methodological*, **196**, 103218.

Harks, T., Klimm, M., 2012. On the existence of pure Nash equilibria in weighted congestion games. *Math Oper Res*, **37**, 419–436.

Highway capacity manual, 2000. Highway capacity manual. *Washington DC*, **2**, 1.

Hörcher, D., Graham, D. J., Anderson, R. J., 2017. Crowding cost estimation with large scale smart card and vehicle location data. *Transp Res Part B Methodol*, **95**, 105–125.

Horowitz, A., Creasey, T., Pendyala, R., Chen, M., 2014. *Analytical Travel Forecasting Approaches for Project-Level Planning and Design (No. Project 08-83)*. Washington, USA: Transportation Research Board.

Horowitz, A., 1991. Delay/Volume relations for travel forecasting based upon the 1985 Highway Capacity Manual. Milwaukee: Department of Civil Engineering and Mechanics University of Wisconsin–Milwaukee. Federal Highway Administration U.S. Department of Transportation, 1-71.

Huang, Y., Ye, Y., Sun, J., Tian, Y., 2023a. Characterizing the impact of autonomous vehicles on macroscopic fundamental diagrams. *IEEE Trans Intell Transp Syst*, **24**, 6530–6541.

Huang, A. J., Agarwal, S., 2023b. On the limitations of physics-informed deep learning: Illustrations using first-order hyperbolic conservation law-based traffic flow models. *IEEE Open J Intell Transp Syst*, **4**, 279–293.

Huntsinger, L. F., Rouphail, N. M., 2011. Bottleneck and queuing analysis: Calibrating volume–delay functions of travel demand models. *Transp Res Rec J Transp Res Board*, **2255**, 117–124.

Huo, J., Wu, X., Lyu, C., Zhang, W., Liu, Z., 2022. Quantify the road link performance and capacity using deep learning models. *IEEE Trans Intell Transp Syst*, **23**, 18581–18591.

Hurdle, V. F., 1984. Signalized intersection delay models—a primer for the uninitiated. *Transp Res Rec*, **971**, 96–105.

Jia, A., Zhou, X., Li, M., Rouphail, N. M., Williams, B. M., 2011. Incorporating stochastic road capacity into day-to-day traffic simulation and traveler learning framework: Model development and case study. *Transp Res Rec J Transp Res Board*, **2254**, 112–121.

Ka, E., Xue, J., Leclercq, L., Ukkusuri, S. V., 2024. A physics-informed machine learning for generalized bathtub model in large-scale urban networks. *Transp Res Part C Emerg Technol*, **164**, 104661.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. *Nat Rev Phys*, **3**, 422–440.

Kerner, B. S., 2000. Theory of breakdown phenomenon at highway bottlenecks. *Transp Res Rec J Transp Res Board*, **1710**, 136–144.

Ke, J., Zheng, H., Yang, H., Chen, X., 2017. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transp Res Part C Emerg Technol*, **85**, 591–608.

Kim, J., Mahmassani, H. S., Dong, J., 2010. Likelihood and duration of flow breakdown: Modeling the effect of weather. *Transp Res Rec J Transp Res Board*, **2188**, 19–28.

Kim, Y. G., Mahmassani, H. S., 1987. Link performance functions for urban freeways with asymmetric car-truck interactions. *Transp Res Rec*, **1120**, 32–39.

Kimber, R.M., Hollis, E.M., 1979. Traffic queues and delays at road junctions. https://api.semanticscholar.org/CorpusID:106923193

Kosun, C., Tayfur, G., Celik, H. M., 2016. Soft computing and regression modelling approaches for link-capacity functions. *Neural Netw World*, **26**, 129–140.

Kucharski, R., Drabicki, A., 2017. Estimating macroscopic volume delay functions with the traffic density derived from measured speeds and flows. *J Adv Transp*, **2017**, 4629792.

Lai, Y. R., Barkan, C. P. L., 2009. Enhanced parametric railway capacity evaluation tool. *Transp Res Rec J Transp Res Board*, **2117**, 33–40.

Lam, W.H.K., Cheung, C., 2000. Pedestrian speed/flow relationships for walking facilities in Hong Kong. *J Transp Eng*, **126**, 343–349.

Lam, W.H.K., Lee, J.Y.S., Cheung, C.Y., 2002. A study of the bi-directional pedestrian flow characteristics at Hong Kong signalized crosswalk facilities. *Transportation*, **29**, 169–192.

Larsson, T., Patriksson, M., 1995. An augmented lagrangean dual algorithm for link capacity side constrained traffic assignment problems. *Transp Res Part B Methodol*, **29**, 433–455.

Lawson, T. W., Lovell, D. J., Daganzo, C. F., 1997. Using input-output diagram to determine spatial and temporal extents of a queue upstream of a bottleneck. *Transp Res Rec J Transp Res Board*, **1572**, 140–147.

Lazar, D. A., Coogan, S., Pedarsani, R., 2020. Routing for traffic networks with mixed autonomy. *IEEE Trans Autom Control*, **66**, 2664–2676.

Leclercq, L., Chiabaut, N., Trinquier, B., 2014. Macroscopic fundamental diagrams: A cross-comparison of estimation methods. *Transportation Research Part B: Methodological*, **62**, 1–12.

Levin, M. W., Boyles, S. D., 2015. Effects of autonomous vehicle ownership on trip, mode, and route choice. *Transp Res Rec J Transp Res Board*, **2493**, 29–38.

Li, X., 2022. Trade-off between safety, mobility and stability in automated vehicle following control: An analytical method. *Transp Res Part B Methodol*, **166**, 1–18.

Li, Y., Zheng, J., Qin, L., Li, H., 2025. Highway capacity of mixed traffic flow with autonomous vehicles: A review. *Phys A Stat Mech Appl*, **671**, 130653.

Lianeas, T., Nikolova, E., Stier-Moses, N. E., 2018. Risk-averse selfish routing. *Math Oper Res*, **44**, 38–57.

Liu, J., Jiang, R., Zhao, J., Shen, W., 2023. A quantile-regression physics-informed deep learning for car-following model. *Transp Res Part C Emerg Technol*, **154**, 104275.

Liu, T., Meidani, H., 2024. End-to-end heterogeneous graph neural networks for traffic assignment. *Transp Res Part C Emerg Technol*, **165**, 104695.

Long, J., Gao, Z., Szeto, W. Y., 2011. Discretised link travel time models based on cumulative flows: Formulations and properties. *Transp Res Part B Methodol*, **45**, 232–254.

Long, K., Sheng, Z., Shi, H., Li, X., Chen, S., Ahn, S., 2025. Physical enhanced residual learning (PERL) framework for vehicle trajectory prediction. *Commun Transp Res*, **5**, 100166.

Long, K., Shi, X., Li, X., 2024. Physics-informed neural network for cross-dynamics vehicle trajectory stitching. *Transp Res Part E Logist Transp Rev*, **192**, 103799.

Lorenz, M. R., Elefteriadou, L., 2000. A probabilistic approach to defining freeway capacity and breakdown. https://api.semanticscholar.org/CorpusID:55398699

Lu, J., Li, C., Wu, X. B., Zhou, X. S., 2023. Physics-informed neural networks for integrated traffic state and queue profile estimation: A differentiable programming approach on layered computational graphs. *Transp Res Part C Emerg Technol*, **153**, 104224.

Ma, X., Lo, H. K., 2012. Modeling transport management and land use over time. *Transp Res Part B Methodol*, **46**, 687–709.

Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transp Res Part C Emerg Technol*, **54**, 187–197.

Mahmassani, H. S., 2016. 50th anniversary invited article—Autonomous vehicles and connected vehicle systems: Flow and operations considerations. *Transp Sci*, **50**, 1140–1162.

Mahmassani, H. S., Hou, T., Saberi, M., 2013. Connecting networkwide travel time reliability and the network fundamental diagram of traffic flow. *Transp Res Rec J Transp Res Board*, **2391**, 80–91.

Mesbah, M., Sarvi, M., Ouveysi, I., Currie, G., 2011. Optimization of transit priority in the transportation network using a decomposition methodology. *Transp Res Part C Emerg Technol*, **19**, 363–373.

Miandoabchi, E., Farahani, R. Z., Dullaert, W., Szeto, W. Y., 2012. Hybrid evolutionary metaheuristics for concurrent multi-objective design of urban road and public transit networks. *Netw Spatial Econ*, **12**, 441–480.

Mo, Z., Shi, R., Di, X., 2021. A physics-informed deep learning paradigm for car-following models. *Transp Res Part C Emerg Technol*, **130**, 103240.

Montanino, M., Monteil, J., Punzo, V., 2021. From homogeneous to heterogeneous traffic flows: L_p String stability under uncertain model parameters. *Transp Res Part B Methodol*, **146**, 136–154.

Moses, R., Mtoi, E., Ruegg, S., McBean, H., 2013. Development of speed models for improving travel forecasting and highway performance evaluation. BDK83-977-14.

Mosher Jr, W. W. 1963. A capacity-restraint algorithm for assigning flow to a transport network. Highw Res Rec 6, 41–70.

Müller, S., Schiller, C., 2015. Improvement of the volume-delay function by incorporating the impact of trucks on traffic flow. *Transp Plan Technol*, **38**, 878–888.

Nagel, K., Beckman, R.J., Barrett, C.L., 1999. TRANSIMS for urban planning. In: 6th international conference on computers in urban planning and urban management, Venice, Italy.

Nagurney, A., Qiang, Q., 2012. Fragile networks: Identifying vulnerabilities and synergies in an uncertain age. *Int Trans Oper Res*, **19**, 123–160.

Newell, G. F., 1982. *Applications of Queueing Theory*. Dordrecht, the Netherlands: Springer .

Newell, G. F., 1968a. Queues with time-dependent arrival rates I: The transition through saturation. *J Appl Probab*, **5**, 436–451.

Newell, G. F., 1968b. Queues with time-dependent arrival rates: II. *The maximum queue and the return to equilibrium. J Appl Probab*, **5**, 579–590.

Ni, D., Leonard, J. D., Jia, C., Wang, J., 2016. Vehicle longitudinal control and traffic stream modeling. *Transp Sci*, **50**, 1016–1031.

Nie, Y., Zhang, H. M., Lee, D. H. (2004). Models and algorithms for the traffic assignment problem with link capacity constraints. Transportation Research Part B: Methodological, 38(4), 285-312.

Ordóñez, F., Stier-Moses, N. E., 2010. Wardrop equilibria with risk-averse users. *Transp Sci*, **44**, 63–86.

Pan, Y., Guo, J., Chen, Y., 2022. Calibration of dynamic volume-delay functions: A rolling horizon-based parsimonious modeling perspective. *Transp Res Rec J Transp Res Board*, **2676**, 606–620.

Pan, Y. A., Guo, J., Chen, Y., Cheng, Q., Li, W., Liu, Y., 2024. A fundamental diagram based hybrid framework for traffic flow estimation and prediction by combining a Markovian model with deep learning. *Expert Syst Appl*, **238**, 122219.

Pan, Y. A., Li, F., Li, A., Niu, Z., Liu, Z., 2025. Urban intersection traffic flow prediction: A physics-guided stepwise framework utilizing spatio-temporal graph neural network algorithms. *Multimodal Transp*, **4**, 100207.

Pan, Y., Cheng, Q., Li, A., Zhang, J., Guo, J., Chen, Y., 2025. Analysis of congestion key parameters, dynamic discharge process, and capacity estimation at urban freeway bottlenecks: A case study in Beijing, China. *Transp Lett*, **17**, 984–1003.

Patriksson, M., 2015. *The Traffic Assignment Problem: Models and Methods*. New York, USA: Dover Publications.

Peeta, S., Ziliaskopoulos, A. K., 2001. Foundations of dynamic traffic assignment: The past, the present and the future. *Netw Spatial Econ*, **1**, 233–265.

Petersen, E. R., 1974. Over-the-road transit time for a single track railway. *Transp Sci*, **8**, 65–74.

Powell, W. B., Sheffi, Y., 1982. The convergence of equilibrium algorithms with predetermined step sizes. *Transp Sci*, **16**, 45–55.

Prokopy, J.C., Richard, B.R., 1975. Parametric analyis of railway line capacity. FRA-OPPD-75-1.

Qian, Z.S., Li, J., Li, X., Zhang, M., Wang, H., 2017. Modeling heterogeneous traffic flow: A pragmatic approach. *Transp Res Part B Methodol*, **99**, 183–204.

Raadsen, M. P. H., Bliemer, M. C. J., 2019. Steady-state link travel time methods: Formulation, derivation, classification, and unification. *Transp Res Part B Methodol*, **122**, 167–191.

Raadsen, M. P. H., Bliemer, M. C. J., Bell, M. G. H., 2020. Aggregation, disaggregation and decomposition methods in traffic assignment: Historical perspectives and new trends. *Transp Res Part B Methodol*, **139**, 199–223.

Rakha, H., Zhang, W., 2005. Consistency of shock-wave and queuing theory. In: 84th Annual Meeting of the Transportation Research Board, 219–226.

Ran, B., Boyce, D. E., 1996. A link-based variational inequality formulation of ideal dynamic user-optimal route choice problem. *Transp Res Part C Emerg Technol*, **4**, 1–12.

Ran, B., Hall, R. W., Boyce, D. E., 1996. A link-based variational inequality model for dynamic departure time/route choice. *Transp Res Part B Methodol*, **30**, 31–46.

Roughgarden, T., Tardos, É., 2002. How bad is selfish routing? *J ACM*, **49**, 236–259.

Rouphail, N., Tarko, A., Li, J., 1992. Traffic Flow at Signalized Intersections. https://www.fhwa.dot.gov/publications/research/operations/tft/chap9.pdf

Rowan, D., He, H., Hui, F., Yasir, A., Mohammed, Q., 2025. A systematic review of machine learning-based microscopic traffic flow models and simulations. *Commun Transp Res*, **5**, 100164.

Samaranayake, S., Chand, S., Sinha, A., Dixit, V., 2024. Impact of connected and automated vehicles on the travel time reliability of an urban network. *Int J Transp Sci Technol*, **13**, 171–185.

Saric, A., Albinovic, S., Dzebo, S., Pozder, M., 2018. Volume-delay functions: A review. In: Advanced Technologies, Systems, and Applications III, 3–12.

Schwarz, J. A., Selinka, G., Stolletz, R., 2016. Performance analysis of time-dependent queueing systems: Survey and classification. *Omega*, **63**, 170–189.

Sengupta, A., Guler, S. I., 2025. Deep learning–based spatial translation of traffic prediction using newell's theory. *J Transp Eng Part A Syst*, **151**, 04025041.

Sheffi, Y., 1984. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Englewood Cliffs, USA: Prentice-Hall.

Shi, X., Li, X., 2021. Constructing a fundamental diagram for traffic flow with automated vehicles: Methodology and demonstration. *Transp Res Part B Methodol*, **150**, 279–292.

Shi, R., Mo, Z., Huang, K., Di, X., Du, Q., 2022. A physics-informed deep learning paradigm for traffic state and fundamental diagram estimation. *IEEE Trans Intell Transp Syst*, **23**, 11688–11698.

Skabardonis, A., Dowling, R., 1997. Improved speed-flow relationships for planning applications. *Transp Res Rec J Transp Res Board*, **1572**, 18–23.

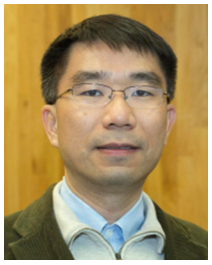
Small, K. A., 1983. The incidence of congestion tolls on urban highways. *J Urban Econ*, **13**, 90–111.

Small, K. A., Verhoef, E. T., Lindsey, R., 2007. *The Economics of Urban Transportation*. London, UK: Routledge.
Smock, R., 1962. An iterative assignment approach to capacity restraint on arterial networks. *Highw Res Board Bull*, **347**, 60−66.
Sogin, S. L., Lai, Y. R., Dick, C. T., Barkan, C. P., 2016. Analyzing the transition from single- to double-track railway lines with nonlinear regression analysis. *Proc Inst Mech Eng Part F J Rail Rapid Transit*, **230**, 1877−1889.
Soltman, T. J. 1966. *Selected Bibliography for Transportation Planning* (No. 97 pp). Washington DC: Pittsburgh Area Transportation Study, 97.
Song, L., Fan, W., 2023. Intersection capacity adjustments considering different market penetration rates of connected and automated vehicles. *Transp Plan Technol*, **46**, 286−303.
Spiess, H., 1990. Conical volume-delay functions. *Transp Sci*, **24**, 153−158.
Srinivasan, D., Choy, M. C., Cheu, R. L., 2006. Neural networks for real-time traffic signal control. *IEEE Trans Intell Transp Syst*, **7**, 261−272.
Szeto, W. Y., Jiang, Y., Wang, D. Z. W., Sumalee, A., 2015. A sustainable road network design problem with land use transportation interaction over time. *Netw Spatial Econ*, **15**, 791−822.
Tak, S., Kim, S., Jang, K., Yeo, H., 2014. Real-time travel time prediction using multi-level k-nearest neighbor algorithm and data fusion method. In: Computing in Civil and Building Engineering (2014), 1861−1868.
Talebpour, A., Mahmassani, H. S., 2016. Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transp Res Part C Emerg Technol*, **71**, 143−163.
Tang, Y., Jin, L., Ozbay, K., 2024. Physics-informed machine learning for calibrating macroscopic traffic flow models. *Transp Sci*, **58**, 1389−1402.
Tarko, A. P., Perez-Cartagena, R. I., 2005. Variability of peak hour factor at intersections. *Transp Res Rec J Transp Res Board*, **1920**, 125−130.
Tisato, P., 1991. Suggestions for an improved Davidson travel time function. *Australian road research*, **21**, 85−100.
Van, O., Fosgerau, M., 2009. Workers' marginal costs of commuting. *J Urban Econ*, **65**, 38−47.
Vickrey, W.S., 1969. Congestion theory and transport investment. *Am Econ Rev*, **59**, 251−260.
Wang, D., Zhang, J., Cao, W., Li, J., Zheng, Y., 2018. When will you arrive? estimating travel time based on deep neural networks. Proc AAAI Conf Artif Intell 32. https://doi.org/10.1609/aaai.v32i1.11877
Wang, H., Li, J., Chen, Q. Y., Ni, D., 2011. Logistic modeling of the equilibrium speed–density relationship. *Transp Res Part A Policy Pract*, **45**, 554−566.
Webster, F.V., 1958. Traffic signal settings. https://trid.trb.org/View/113579
Wei, Y., Avcı, C., Liu, J., Belezamo, B., Aydın, N., Li, P. et al., 2017. Dynamic programming-based multi-vehicle longitudinal trajectory optimization with simplified car following models. *Transp Res Part B Methodol*, **106**, 102−129.
Wei, Z., Cheng, Y., Zhu, J., Huang, Q., 2019. Genipin-crosslinked ovotransferrin particle-stabilized Pickering emulsions as delivery vehicles for hesperidin. *Food Hydrocoll*, **94**, 561−573.
Wong, W., Wong, S. C., 2016. Network topological effects on the macroscopic Bureau of Public Roads function. *Transp A Transp Sci*, **12**, 272−296.
Wu, C. H., Ho, J. M., Lee, D. T., 2004. Travel-time prediction with support vector regression. *IEEE Trans Intell Transp Syst*, **5**, 276−281.
Wu, X., Guo, J., Xian, K., Zhou, X., 2018. Hierarchical travel demand estimation using multiple data sources: A forward and backward propagation algorithmic framework on a layered computational graph. *Transp Res Part C Emerg Technol*, **96**, 321−346.
Wu, X., Dutta, A., Zhang, W., Zhu, H., Livshits, V., Zhou, X., 2021. *Characterization and Calibration of Volume-to-Capacity Ratio in Volume Delay Functions on Freeways Based on a Queue Analysis Approach*. In: Transportation Research Board 100th Annual Meeting.
Xiong, H., Davis, G. A., 2009. Field evaluation of model-based estimation of arterial link travel times. *Transp Res Rec J Transp Res Board*, **2130**, 149−157.
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.
Yang, H., Wang, X., 2011. Managing network mobility with tradable credits. *Transp Res Part B Methodol*, **45**, 580−594.
Yang, H., Yagar, S., 1995. Traffic assignment and signal control in saturated road networks. *Transp Res Part A Policy Pract*, **29**, 125−139.
Yeon, J., Hernandez, S., Elefteriadou, L., 2009. Differences in freeway capacity by day of the week, time of day, and segment type. *J Transp Eng*, **135**, 416−426.
Yperman, I., Logghe, S., Immers, B., 2005. The link transmission model: An efficient implementation of the kinematic wave theory in traffic networks, In: Proceedings of the 10th EWGT Meeting, 122−127.
Zhang, L., Levinson, D., 2004. Some properties of flows at freeway bottlenecks. *Transp Res Rec J Transp Res Board*, **1883**, 122−131.
Zhang, X., Waller, S. T., 2018. Link performance functions for high occupancy vehicle lanes of freeways. *Transport*, **33**, 657−668.
Zhang, X., Sun, J., Sun, J., 2025. On the stochastic fundamental diagram: A general micro-macroscopic traffic flow modeling framework. *Commun Transp Res*, **5**, 100163.
Zhao, Z., Liang, Y., 2023. A deep inverse reinforcement learning approach to route choice modeling with context-dependent rewards. *Transp Res Part C Emerg Technol*, **149**, 104079.
Zhou, X., Hadi, M., Hale, D., 2021. Multiresolution modeling for traffic analysis state-of-practice and gap analysis report. FHWA-HRT-21-082.
Zhou, X. S., Cheng, Q., Wu, X., Li, P., Belezamo, B., Lu, J., Abbasi, M., 2022. A meso-to-macro cross-resolution performance approach for connecting polynomial arrival queue model to volume-delay function with inflow demand-to-capacity ratio. *Multimodal Transp*, **1**, 100017.
Zhou, D., Gayah, V. V., 2023. Improving deep reinforcement learning-based perimeter metering control methods with domain control knowledge. *Transp Res Rec J Transp Res Board*, **2677**, 384−405.

Yuyan (Annie) Pan received the Ph.D. degree in transportation engineering from Beijing University of Technology in 2023. She is currently a Postdoctoral Scholar in the Department of Civil and Environmental Engineering at The Pennsylvania State University. Her research focuses on connected and automated vehicles, electric vehicle charging, traffic flow theory, and data-driven modeling. She has published more than 20 peer-reviewed papers in journals including *Transportation Research Part C*, and serves as a reviewer for journals such as *Transportation Research Part B*, *Part C*, and *Part E*. Her work integrates fundamental traffic theory with AI-based methods to enhance traffic system efficiency and resilience.

Xianbiao (XB) Hu received the Ph.D. degree from the University of Arizona, Tucson, AZ, USA, in 2013. He is currently an Associate Professor in the Department of Civil and Environmental Engineering, The Pennsylvania State University. His current research interests include smart mobility systems, connected and automated vehicles, electric vehicles, mobility behavior management, transportation big data analytics, and traffic flow and system modeling. He serves as an Associate Editor for *IEEE Transactions on Intelligent Transportation Systems*, an Associate Editor of the *Journal of Intelligent Transportation Systems*, and a Handling Editor of *Transportation Research Record* and an Associate Editor for *The International Journal of Transportation Science and Technology*.

George List received the B.S.E.E. degree from Carnegie Mellon University, Pittsburgh, PA, USA, in 1971; the M.E.E. degree from the University of Delaware, Newark, DE, USA, in 1976; and the Ph.D. degree in civil engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 1984. He is currently a Professor in the Department of Civil, Construction and Environmental Engineering at North Carolina State University, Raleigh, NC, USA. He is best known for his work on the modeling, simulation, and optimization of transport systems and networks. His professional affiliations include ASCE (Fellow), TRB, IEEE, ITE, and INFORMS.



Xuesong (Simon) Zhou received the Ph.D. degree in civil engineering from the University of Maryland, College Park, MD, USA, in 2004. He is currently a Professor in the School of Sustainable Engineering and the Built Environment at Arizona State University, Tempe, AZ, USA. His research focuses on dynamic traffic assignment, traffic estimation and prediction, large-scale routing, and rail scheduling. He previously served as an Associate Editor for *Transportation Research Part C* and is currently the Executive Co-Editor-in-Chief of *Urban Rail Transit*, and a Senior Associate Editor for *Transportation Research Part B* and an Associate Editor for *Transportation Science*. He has also chaired the INFORMS Rail Application Section (2016 and 2025) and currently serves as Chair of the Transportation Research Board Committee on Transportation Network Analysis (AEP13).