Communications in Transportation Research 5 (2025) 100163

Contents lists available at ScienceDirect

Communications in Transportation Research

journal homepage: www.journals.elsevier.com/communications-in-transportation-research

Full Length Article

On the stochastic fundamental diagram: A general micro-macroscopic traffic flow modeling framework

Xiaohui Zhang, Jie Sun, Jian Sun *

Department of Traffic Engineering & Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University, Shanghai, 200092, China

ARTICLE INFO

Keywords:
Stochastic fundamental diagram
Brownian dynamic
Markov chain
Maximum entropy

ABSTRACT

The stochastic fundamental diagram (SFD), which describes the stochasticity of the macroscopic relations of traffic flow, plays a crucial role in understanding the uncertainty of traffic flow evolution and developing robust traffic control strategies. Although many efforts have been made to reproduce the SFD via various methods, few studies have focused on the analytical modeling of the SFD, particularly linking the macroscopic relations with microscopic behaviors. This study fills this gap by proposing a general micro-macroscopic modeling approach, which uses probabilistic leader–follower behavior to derive the macroscopic relations of a platoon and is referred to as the leader–follower conditional distribution-based stochastic traffic modeling (LFCD-STM) framework. Specifically, we first define a conditional probability distribution of speed for the leader–follower pair according to Brownian dynamics, which is proven to be a general representation of the longitudinal interaction and compatible with classical car-following models. As a result, we can describe the joint distribution of vehicle speeds of the platoon through Markov chain modeling and further derive the macroscopic relations (e.g., the mean flow–density relation and its variance) under equilibrium conditions. On the basis of this general micro-macroscopic framework, we utilize the maximum entropy approach to theoretically derive the SFD model, in which we provide a specific conditional distribution for longitudinal interaction and thus solve the analytical functions of the mean and variance of FD. The performance of the maximum entropy-based SFD model is thoroughly validated with the NGSIM I-80, US-101 and HighD datasets. The high consistency between the theoretical results and empirical results demonstrates the soundness of the LFCD-STM framework and the maximum entropy-based SFD model. Finally, the proposed SFD model has practical implications for promoting smoother driving behaviors to suppress stochasticity and improve traffic flow.

Nomenclature

Nomenclature	Notation	Parameter	Notation
Random variable		Variable	
Speed of vehicle i	V_i	Spacing between vehicle i and its leader	s_i
Equilibrium speed	V	Equilibrium spacing	s_e
Equilibrium flow	Q	Equilibrium density	k
		Total number of the vehicles in the study section	n
Function		Parameter	
Characteristic function for interaction	f_{int}	Average speed of the platoon	$\bar{v}$
Characteristic function for own speed	f_{own}	Maximum speed	v_{max}
Characteristic function for fluctuation	f_{flu}	Lagrange multiplier	$\lambda_1, \lambda_2, \lambda_3$
Normalization factor	z	Length of the study area	l
Brownian motion	$B(t)$	Regression parameters of the logarithm relationship	$\eta, \theta, \alpha, \beta$

(*continued on next column*)

(*continued*)

1. Introduction

The fundamental diagram (FD), which describes the equilibrium relationship among three macroscopic traffic flow properties, i.e., traffic flow (veh/h), traffic density (veh/km), and speed (km/h), has been considered the foundation of traffic flow theory and transportation engineering. Given that flow is the product of speed and density, FD usually refers to the flow–density relation. These relations have been widely employed in a diverse range of traffic research fields, such as traffic flow modeling and simulation (Daganzo, 1977; Treiber and Kesting, 2013),

* Corresponding author.
E-mail address: sunjian@tongji.edu.cn (J. Sun).

https://doi.org/10.1016/j.commtr.2025.100163
Received 5 October 2024; Received in revised form 10 November 2024; Accepted 20 November 2024

traffic state estimation (Seo et al., 2017), and traffic management and control (Papageorgiou et al., 2003).

Since the introduction of the seminal Greenshields model (Greenshields et al., 1935), extensive efforts have been made to enhance this oversimplified linear speed–density relationship either empirically or analytically (Edie, 1961; Greenberg, 1959; Ji et al., 2010; Keyvan-Ekbatani et al., 2012; Li and Zhang, 2011; Newell, 1961; Wang et al., 2011; Wu et al., 2011). These models can be categorized as single-regime models or multi-regime models in terms of whether the model works over the entire density range in one equation. However, significant issues exist for these models, where single-regime models have been criticized for not fitting the empirical data well (Ni, 2016) and multi-regime models have been criticized for their lack of mathematical tractability and inability to determine breakpoints in a scientific manner (Liu et al., 2019; Sun and Zhou, 2005). More importantly, in both single- and multi-regime models, the FD is modeled as a deterministic function, i.e., each density corresponds to a single deterministic value of speed, which is referred to as deterministic FD (DFD), whereas such a hypothesis of the deterministic function fails to accord with the widely scattered flow–density or speed–density data in real traffic. Although early research suggested that the scattering effect may be due to nonstationary traffic data (Cassidy, 1998), a growing number of studies have suggested that it is caused by the stochastic nature of traffic flow (Coifman, 2015; Daganzo, 2002; Kerner, 2009; Treiber et al., 2006), e.g., the uncertainty of driving behavior, dynamic traffic environments, heterogeneity of vehicle classes and driving styles that cannot be captured by DFD models.

Therefore, several pioneering studies have attempted to model the stochasticity in FD, which is referred to as SFD model, where the stochasticity is described as the variance of the FD. These SFD models can be generally categorized into three classes according to their modeling scales, namely, macroscopic, mesoscopic, and microscopic approaches. Macroscopic SFD models usually extend existing deterministic macroscopic relations with additional random terms or heterogeneous parameters to accommodate the uncertainty (Ahmed et al., 2021; Anupriya et al., 2023; Boel and Mihaylova, 2006; Cheng et al., 2024; Ngoduy, 2011; Treiber et al., 2006; Wagner et al., 1997; Zhang et al., 2018). Mesoscopic approaches comprise macroscopic and microscopic directions, where one harnesses the log-normal distribution of time headway with respect to speed and derives the probabilistic flow-speed relation on the basis of the reciprocal relationship between time headway and flow (Li and Chen, 2017; Wu and Liu, 2013), and the other adopts the stochastic differential equation method to describe speed transitions with a discrete speed spectrum and then derives the mean and variance functions of FD (Siqueira et al., 2016). Microscopic approaches aim to reproduce the uncertainty in macroscopic traffic flow via simulations by extending existing CF models with additional random terms or varying parameters to involve heterogeneity (Jabari et al., 2014; Nishinari et al., 2004). More details are presented in Section 2.

Although these approaches can replicate the characteristics of the SFD to some extent, one critical issue is that analytical solutions for the SFD, namely, the analytical mean and variance of the FD, are generally missing (except for the attempt of Siqueira et al. (2016)), although analytical tractability is highly important and in need of an explicit understanding and modeling of the uncertainty in traffic flow and thus the practical implications in suppressing traffic flow stochasticity and mitigating traffic congestion (Keyvan-Ekbatani et al., 2012; Qu et al., 2017). In particular, the recent development of connected and autonomous vehicles (CAVs) has created both challenges and opportunities for traffic systems, as mixed traffic flows consisting of conventional, connected, and autonomous vehicles may exhibit distinct characteristics with more heterogeneity and stochasticity, which further requires rigorous analysis. Moreover, CAVs can be used for developing effective control strategies to maximize capacity and minimize stochasticity accordingly on the basis of the analytical SFD. These findings further underline the importance and urgency of modeling SFD in an analytical manner.

More importantly, although the stochastic nature of the macroscopic traffic flow originates from the behavior of microscopic individuals (Jabari et al., 2014; Wu and Liu, 2013), how to link the microscopic behavior and their dynamic interactions with the SFD remains unclear. Although many DFD models have been derived from CF models (Pipes, 1967; Treiber and Kesting, 2013), these deterministic relations are only able to represent the mean FD and fail to capture the full range of traffic scenarios, which is not robust enough for real-world applications. While it is challenging to mathematically and simultaneously infer the mean and variance in SFD, as uncertainty at the microscopic level does not simply "average out" to yield a closed-form of stochasticity at the macro level, establishing an analytical link between SFD and microscopic driving behaviors is much needed. This link enables the optimization of macroscopic traffic flow with tailored control strategies using CAVs. In summary, the analytical model of the SFD, which gives explicit formulations for the mean and variance of the flow–density relation from a microscopic perspective, has not been established, and there is a great need to create this micro–macro link.

To fill this gap, we propose a general micro-macroscopic traffic flow modeling approach (referred to as the leader–follower conditional distribution-based stochastic traffic modeling (LFCD-STM) framework), which enables the derivation of stochastic macroscopic relations from probabilistic microscopic leader–follower interactions. Based on the maximum entropy approach, we further obtain the analytical formulations of the SFD. An overview of this study is illustrated in Fig. 1. Specifically, the major contributions of this study are as follows.

1) To capture the microscopic origin of stochasticity, we introduce the conditional distribution of follower speed given the leader's speed on the basis of Brownian dynamics, providing a general representation of the longitudinal interaction.
2) Applying the Markov chain, we bridge the micro-macroscopic gap by describing the joint distribution of speeds for the platoon with individual conditional distributions, based on which the general macroscopic relations are characterized under the equilibrium condition.
3) Using the maximum entropy approach, we present a particular conditional distribution of longitudinal interactions and obtain the analytical functions of the mean and variance of FD. We also validate this maximum entropy-based SFD model with the NGSIM and HighD datasets, where the results indicate high consistency with the observed data.

Moreover, we discuss several issues related to this methodology. Specifically, we first investigate the impact of traffic flow heterogeneity on the SFD model. Moreover, we examine the connections between the microscopic behavior and SFD and analyze the influence of different microscopic parameters on the SFD model. Finally, we highlight several applications in terms of FD estimation, traffic flow modeling, and robust traffic control strategies.

The rest of the paper is structured as follows: Section 2 reviews the related literature; Section 3 introduces the LFCD-STM framework; Section 4 derives the SFD model with the maximum entropy distribution; Section 5 calibrates and evaluates the SFD model using NGSIM and HighD datasets; Section 6 discusses related issues of the proposed model; and Section 7 summarizes the main findings of this work.

2. Literature review

In this section, we review representative studies of SFD modeling, which are divided into three classes according to the traffic flow modeling scales, i.e., macroscopic, mesoscopic, and microscopic approaches.

2.1. Macroscopic approaches

SFD modeling from the macroscopic perspective generally falls into two categories, namely, calibrating SFDs with empirical macroscopic

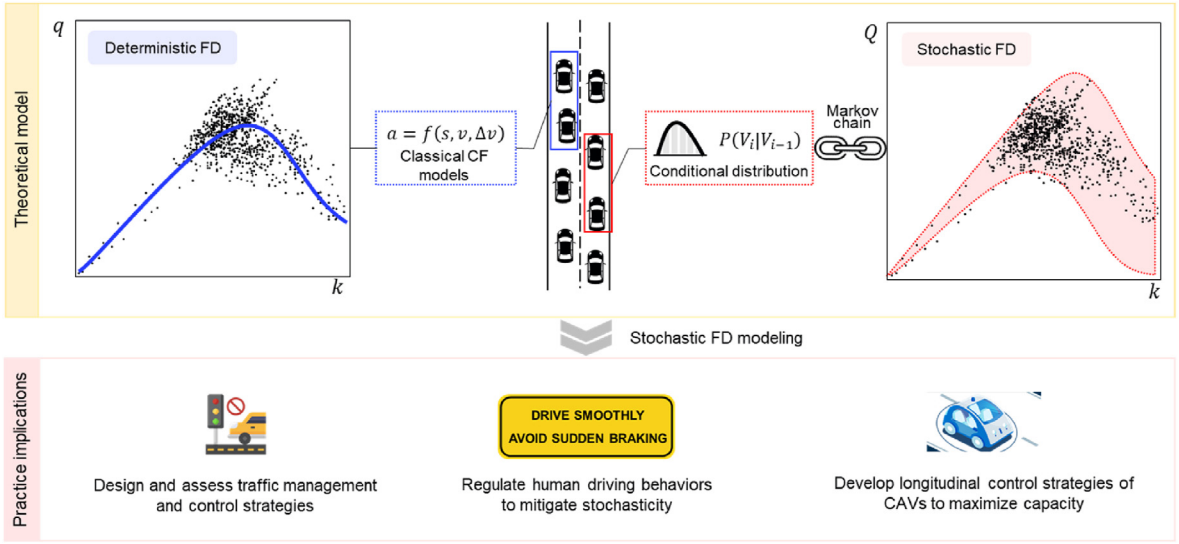


Fig. 1. Theoretical modeling of FD (DFD modeling from classical car-following (CF) models and stochastic fundamental diagram (SFD) modeling based on leader–follower conditional distribution-based (LFCD)) and practical implications.

traffic data and adding noise to the deterministic macroscopic model.

With respect to the former branch, by introducing a percentile parameter, a family of deterministic macroscopic relations was generated to cover the SFD via an optimization model that is based on the theorem of total probability (Qu et al., 2017; Wang et al., 2021). Additionally, Ni et al. (2018) and Bramich et al. (2023) presented a data-driven methodology that directly models the FD as a stochastic process, where the flow is assumed to follow a Gaussian distribution with parameters expressed as functions of the density. Bai et al. (2021) proposed a stochastic link-based FD, which explicitly considers speed heterogeneity to explore the effects of vehicle and driver diversity for a given density under identical environmental conditions. Using data from multiple sources with different temporal resolutions, Bai et al. (2024) estimated the empirical speed–density relationship with a novel procedure for determining a practical optimal dataset. A recent review on fitting empirical fundamental diagrams can be found in Bramich et al. (2022).

On the other hand, several studies have aimed to extend the deterministic macroscopic model to accommodate uncertainty. Boel and Mihaylova (2006) extended the cell transmission model by defining sending and receiving functions as random variables and specifying the dynamics of the average speed in each cell. This allowed them to describe the macroscopic traffic behavior of each cell and its interaction with neighboring cells and thus capture the randomness in the FD. Similarly, Ngoduy (2011) introduced a stochastic setting in the model parameters for a multiclass first-order model, enabling the simulation of wide scattering in the FD. Additionally, the SFD can be simulated by considering the fluctuations caused by the interactions between different vehicle classes with different parameter settings in a higher-order model (Treiber and Helbing, 1999). Moreover, Kerner (1998) obtained the range of speed for each density for the supposed synchronized flow on the basis of three-phase traffic flow models.

2.2. Mesoscopic approaches

For mesoscopic models, several studies have indicated that the scattering features in the flow–density diagram can be re-established by mesoscopic headway distributions (Chen et al., 2014; Li and Chen, 2017; Wu and Liu, 2013). Generally, suppose that the time headways with respect to speed follow certain log-normal distributions, the corresponding flow rates also follow the log-normal distribution under a given speed range, and the joint probability distributions of flow and density within a speed range can be obtained by projecting the probability distributions of headways within the same speed range.

An alternative mesoscopic model for the SFD utilizes the stochastic differential equation method, which features a discrete speed spectrum and corresponding transition dynamics among different speed states (Siqueira et al., 2016). To capture the inherent uncertainty of traffic flow, stochastic transition terms were introduced alongside traditional transition terms. The model was then simplified to a two-speed-state version, from which the analytic solution of the flow–density relation and its variance were derived to illustrate the essence.

2.3. Microscopic approaches

With respect to SFD modeling from a microscopic viewpoint, most studies in the traffic flow community focus on extending existing CF models with additional random terms or heterogeneous parameters. Based on Newell's simplified CF model, Jabari et al. (2014) developed a probabilistic stationary speed–density relation with varying parameters that represent the heterogeneity of vehicle classes. The optimal velocity model was also extended by incorporating multivehicle interactions to model wide scattering in FDs (Lenz et al., 1999). Moreover, the memory effect, which accounts for drivers' adaptation to surrounding traffic situations, has been incorporated into CF models, enabling the simulation of wide scattering (Treiber and Helbing, 2003). Additionally, Treiber et al. (2006) proposed a variance-driven adaptation mechanism for safety time gaps, describing the increase in safety time gaps in unstable local traffic dynamics. This mechanism also provides a plausible explanation for the widely scattered traffic flows via simulation.

In the field of physics, the cellular automaton (CA) model for freeway traffic has been intensively and extensively studied for a long time since the seminal model proposed by Nagel and Schreckenberg (Chowdhury et al., 2000; de Gier et al., 2019; Nagel and Schreckenberg, 1992; Sopasakis and Katsoulakis, 2006). The random effects introduced into the CA models, e.g., the speed of each vehicle is decreased by one with a certain probability, are deemed to play a vital role in reproducing the stop-and-go waves and the uncertainty in FD.

A summary of representative studies is provided in Table 1. In general, the majority of studies tend to model the SFD qualitatively via simulation, whereas analytical tractability is essential for the efficient prediction and optimization of traffic flow. Moreover, given that the stochasticity of the SFD is well recognized to stem from the microscopic behavior, we aim to analytically model the SFD by constructing a micro-macro link between the microscopic driving behavior and the

Table 1
Representative studies of SFD modeling.

Scale	Ref.	Principle	Underlying model
Macroscopic	Qu et al. (2017)	Calibrating FD model as stochastic process with empirical macroscopic traffic data	Nonlinear speed-density model
	Bramich et al. (2023)		Flow–density model with Gaussian distribution
	Bai et al. (2021)		Random-parameter speed–density model with speed heterogeneity
	Boel and Mihaylova (2006) Ngoduy (2011)	Adding random terms or heterogenous parameters to deterministic macroscopic models	Cell transmission model/LWR model
Mesoscopic	Li and Chen (2017) Wu and Liu (2013)	From the distribution of time headways with respect to speed to SFD via the reciprocal relationship	Distribution of time headways
	Siqueira et al. (2016)	Discrete speed spectrum with stochastic differential equations	Stochastic differential equations
Microscopic	Nagel and Schreckenberg (1992) Chowdhury et al. (2000)	Stochastic cellular automaton models for freeway traffic	Cellular automaton (CA)
	Treiber and Helbing (2003) Treiber et al. (2006)	Incorporating adaptation mechanism into deterministic CF models	Intelligent driver model (IDM)
	Jabari et al. (2014)	Extending deterministic CF models with random terms or heterogenous parameters	Newell's simplified CF model

macroscopic flow–density relation in this study.

3. General micro–macroscopic traffic flow modeling framework

To enable the analytical modeling of SFD, in this study, we develop a general micro–macroscopic traffic flow modeling framework, which links the probabilistic microscopic behavior with stochastic macroscopic relations leveraging the Markov property of the conditional distribution for each leader–follower pair and is referred to as the LFCD-STM framework. The structure of the LFCD-STM framework for estimating SFD is presented in Fig. 2. We first introduce a general representation for microscopic driving behavior, i.e., a conditional distribution of the follower's speed given the leader's speed with the Markov property. Then, a micro-macro link is created with the joint distribution of vehicle speeds in a platoon, which can also be defined as a Markov chain, based on which the flow–density relation and its variance are derived under equilibrium conditions. Notably, this framework provides a general method to model macroscopic traffic flow phenomena from a microscopic perspective, as we can further use the framework to model capacity drop (which has been performed in our following work).

3.1. Probabilistic modeling of microscopic behavior

Although deterministic functions are generally proposed in CF models, there is a growing consensus in the literature that microscopic variables (e.g., speed and acceleration) should be modeled as random variables (Branston, 1976; Hoogendoorn and Bovy, 1998; Jabari et al., 2014; Jabari and Liu, 2012; Mahnke and Kaupužs, 1999) because of the inconsistent nature of human drivers, dynamic traffic environments, and multiple sources of heterogeneity. More importantly, the states (e.g., speeds) of vehicles in the traffic flow are not only stochastic but also correlated. To capture the interactive behavior between the vehicle and its local traffic context, we define the conditional distribution $P(\text{STATE}_i|\text{ENV})$, where STATE_i is the state of vehicle i and ENV is the local environment information.

For the sake of concision, we restrict our focus to the CF process, which is a fundamental behavior at the microscopic level. We adopt the widely used hypothesis in modeling CF behavior, which assumes that the reaction of the following vehicle is mainly determined by the states of the leader and itself (Brackstone and McDonald, 1999). Furthermore, as speed is one of the principal states of a vehicle traveling on a road, we use it to represent the vehicle state as a discrete random variable. Therefore, the interaction of each leader–follower pair is represented by the conditional distribution of the follower's speed V_i given the leader's speed V_{i-1}, i.e., $P(V_i|V_{i-1})$, where $P(V_i|V_{i-1}) = P(V_i|V_{i-1},\dots,V_0)$, which can also be regarded as the Markov property along the vehicle index. In addition, this conditional distribution $P(V_i|V_{i-1})$ is assumed to be a function of s_i, the other influencing factor used in basic CF models, where s_i denotes the

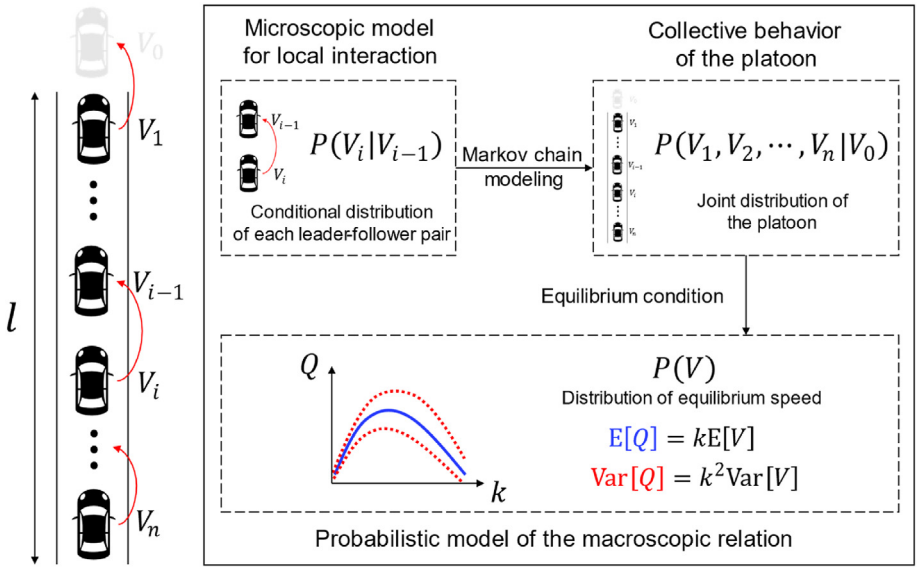


Fig. 2. LFCD-STM framework for analytical modeling of SFDs from a microscopic viewpoint. Note that as this study focuses on probabilistic behavior, we use a number of random variables. To avoid ambiguity, we distinguish random variables from deterministic variables by using upper case letters, including the speed of the i-th vehicle in the platoon V_i, the equilibrium speed of the platoon V, and the equilibrium flow of the platoon Q. Additionally, to better understand the concept of probabilistic modeling in this paper, random variables are treated as discrete random variables since the probability density function used to characterize a continuous random variable is more obscure. Note that as these random variables can be discretized with a very small step, there is no essential difference between the continuous and discrete types.

spacing between the leading vehicle $(i-1)$ and the subject vehicle i. Hence, $P(V_i|V_{i-1})$ is likely to be a stochastic process where the distribution differs as s_i changes. Note that it should be distinguished from the mesoscopic headway distribution, albeit it also uses the probability distribution, because the time headway distribution at the mesoscopic level ignores both the dimensions of time and the vehicle index along the platoon and aggregately describes the state of all vehicles rather than individual vehicles.

To deepen the understanding of the proposed formulation in the conditional distribution and demonstrate the generality of such a representation, we further show its relationship with classical CF models. First, the proposed conditional distribution has an equivalent dynamic CF-like model. According to Brownian dynamics (Ermak and McCammon, 1978; Kawasaki, 1973; Lemons and Gythiel, 1997; Schlick, 2010), the motion of a particle is described by a stochastic differential equation as Eq. (1):

$$\frac{\mathrm{d}V_i(t)}{\mathrm{d}t} = \nabla_{V_i}\log P(V_i(t)|V_{i-1}(t)) + \sqrt{2}B(t) \tag{1}$$

where $B(t)$ represents Brownian motion. It is well established in statistical mechanics that the dynamics will eventually converge to its stationary distribution $P(V_i|V_{i-1})$ as $t \to \infty$. Therefore, the underlying dynamic model of a conditional distribution $P(V_i|V_{i-1})$ can be easily derived with Eq. (1), which has the same form as classical CF models except for the random term. On the other hand, classical CF models can also be transformed to corresponding conditional distributions via the Brownian dynamics in Eq. (1). In Appendix A, we demonstrate this transformation in terms of a widely applied CF model, the optimal velocity relative velocity (OVRV) model (Liang and Peng, 1999; Milanes et al., 2014), as a showcase.

In addition to the direct connection and compatibility with classical CF models, the proposed conditional distribution $P(V_i|V_{i-1})$ has several additional advantages for micro–macro modeling. First, the form of the conditional distribution is a more general representation of microscopic behavior. Unlike the strong assumption of the CF rules in classical CF models, directly modeling microscopic behavior as a conditional distribution facilitates the derivation of the joint distribution macroscopically, where the essential properties of CF dynamics remain while the time evolution is eliminated, which is simple and effective for modeling macroscopic traffic flow. Moreover, the proposed conditional distribution can express and maintain the uncertainty for SFD modeling in an analytical way by accommodating uncertainty as a whole, which is challenging for classical CF models with additional random terms and time evolution. Additionally, this conditional distribution and the proposed LFCD-STM framework are promising to extend for complex nonequilibrium phenomena, e.g., capacity drop, with disturbance propagation, owing to its generality and simplicity.

3.2. Collective description of the platoon

On the basis of the probabilistic model of local interaction within each leader–follower pair, we describe the collective behavior of the whole platoon as a joint distribution, which reflects the interdependency of vehicles and characterizes the entire platoon as one. To show that the micro–macro link is applicable to any distribution of $P(V_i|V_{i-1})$, we use a generic function g to denote it, i.e., $P(V_i = v_i|V_{i-1} = v_{i-1}) = g(s_i, v_i, v_{i-1})$.

Considering a single-lane highway where vehicle i follows vehicle $(i-1)$, as shown in Fig. 2, the joint distribution of vehicle speeds in the system is $P(V_1, \ldots, V_n|V_0)$, where n is the total number of vehicles in the study section and V_0 is the speed of the first leading vehicle.

According to the definition of the conditional distribution, the joint distribution can be rearranged as

$$P(V_1, \ldots, V_n|V_0) = P(V_0, V_1, \ldots, V_n)/P(V_0) \tag{2}$$

Using the chain rule in probability theory and the Markov assumption of longitudinal interaction, we have

$$P(V_0, V_1, \ldots, V_n) = P(V_0)\prod_{i=1}^{n} P(V_i|V_{i-1}, \ldots, V_0) = P(V_0)\prod_{i=1}^{n} P(V_i|V_{i-1}) \tag{3}$$

Substituting Eq. (3) into Eq. (2), the joint distribution is given by

$$P(V_1, \ldots, V_n|V_0) = \prod_{i=1}^{n} P(V_i|V_{i-1}) \tag{4}$$

Assuming that all leader–follower pairs are homogenous and have the same conditional distribution, the joint distribution is derived as

$$P(V_1 = v_1, \ldots, V_n = v_n|V_0 = v_0) = \prod_{i=1}^{n} g(s_i, v_i, v_{i-1}) \tag{5}$$

3.3. Derivation of SFD under equilibrium

The FD is the macroscopic relationship under equilibrium conditions, which means that all the vehicles in the homogeneous platoon remain relatively steady with identical equilibrium states, e.g., equilibrium spacing s_e, where $s_1 = s_2 = \ldots = s_n = s_e$. However, as the speeds of vehicles are modeled as random variables, the equilibrium speed is also a random variable denoted as V. Thus, the corresponding equilibrium condition of speed needs further investigation in the context of probabilistic modeling.

Proposition 1. Under equilibrium conditions, each vehicle in the platoon must take the same speed value simultaneously, and their speeds are equal in distribution, which is also the distribution of equilibrium speed V, i.e., $V_0 \overset{d}{=} V_1 \overset{d}{=} \ldots \overset{d}{=} V_n \overset{d}{=} V$.

The proof of Proposition 1 is shown in Appendix B. Theoretically, for each given equilibrium speed value v_e, all vehicles in the platoon take this particular speed value simultaneously under the equilibrium condition.

The corresponding joint distribution under the equilibrium condition is

$$P(V_1 = v_e, \ldots, V_n = v_e|V_0 = v_e) = \prod_{i=1}^{n} g(s_e, v_e, v_e) = [g(s_e, v_e, v_e)]^n \tag{6}$$

Obviously, it collapses into the distribution of the equilibrium speed, as the equilibrium speed is the only random variable left on the right side of Eq. (6). With normalization, the complete distribution of the equilibrium speed is

$$P(V = v_e) = \frac{[g(s_e, v_e, v_e)]^n}{\sum_{v_e=0}^{v_{\max}} [g(s_e, v_e, v_e)]^n} \tag{7}$$

where $v_{\max}$ is the maximum speed of the platoon.

Remark 1. In general, the equilibrium condition for stochastic systems means that the probabilities of various states (i.e., the distribution) will remain constant, such as the equilibrium distribution of the Markov chain (Chung, 1960), which also accords with Proposition 1 that the distribution of speed is identical along the platoon.

According to the definition of density, that is, $k = n/l$, where k is the density and l is the length of the study area, and the reciprocal relationship between equilibrium spacing and density, that is, $s_e = 1/k$, we obtain the probabilistic relation between density k and equilibrium speed V:

$$P(V = v_e) = \frac{[g(1/k, v_e, v_e)]^{kl}}{\sum_{v_e=0}^{v_{\max}} [g(1/k, v_e, v_e)]^{kl}} = \varphi(k, v_e) \tag{8}$$

where φ denotes the simplified function. As traffic flow Q is the product

of density and speed, i.e., $Q = kV$, the expectation and variance of traffic flow Q can be obtained with respect to density k on the basis of the equilibrium speed distribution $P(V)$ in Eq. (8):

$$E[Q] = kE[V] = k\sum_{v_e=0}^{v_{max}} v_e P(V = v_e) \tag{9}$$

$$\mathrm{Var}[Q] = k^2\mathrm{Var}[V] = k^2\left[\sum_{v_e=0}^{v_{max}} {v_e}^2 P(V = v_e) - \left(\sum_{v_e=0}^{v_{max}} v_e P(V = v_e)\right)^2\right] \tag{10}$$

To verify the proposed general framework, we utilize the connection between classical CF models and conditional distributions and derive the SFD of the OVRV model as an example. The mean of the derived SFD is perfectly consistent with the deterministic FD obtained from the original OVRV model, which provides strong evidence for the effectiveness and generality of the proposed framework in modeling SFD. The variance of the derived SFD mainly depends on the characteristics of the microscopic behavior. More details are presented in Appendix A.

4. SFD modeling based on maximum entropy

To yield explicit expressions for SFD, in this section, we directly construct a conditional speed distribution using the well-known maximum entropy approach and then obtain the maximum entropy-based SFD model. Specifically, we first introduce the maximum entropy approach and then define the maximum entropy distribution of microscopic driving behavior. Finally, we derive the resulting SFD expressions on the basis of the general framework.

4.1. Introduction of the maximum entropy approach

Several approaches exist for formulating a specific conditional distribution $P(V_i|V_{i-1})$ to capture the leader–follower interaction. One approach is to make distributional assumptions, typically following the Gaussian distribution, and calibrate the parameters using maximum likelihood estimation. However, this so-called parametric density estimation (Bishop, 2006) imposes strong assumptions on the distribution, which may be not consistent with reality. In addition, as mentioned before, $P(V_i|V_{i-1})$ should be a function of the spacing s_i; thus, it is intractable to directly formulate the underlying distribution even with the assumption of a specific distribution family. Another approach corresponds to nonparametric density estimation, among which kernel density estimation is the most popular (Parzen, 1962). It places a symmetric density function such as the Gaussian, namely, the kernel, to interpolate between observed data to predict the density at the unobserved range. Nonetheless, beyond the appropriate choice of the kernel, a key challenge of kernel density estimation is the selection of the bandwidth (Li and Racine, 2011), especially in the presence of large datasets. Moreover, it is difficult to obtain a closed-form expression because of the nearly infinite sum in the kernel density estimator for large samples, which deviates from the aim of analytical modeling in this study.

Alternatively, the maximum entropy approach provides an efficient way to obtain a closed-form distribution without strong assumptions for the distribution. The basic idea is to use prior knowledge as constraints and then choose the distribution that is least biased, in the sense that the distribution does not depend on any assumption other than the information contained in our prior knowledge. Therefore, bias, which refers to assumptions not compelled by prior knowledge, is eliminated (Jaynes, 1957, 1982). The least biased distribution has the maximum entropy, which has been rigorously proven in information theory (Csiszar, 1991; Shore and Johnson, 1980). The maximum entropy approach has been applied to numerous problems across many disciplines, including collective behavior modeling in biological systems (Bialek et al., 2014), neurons (Schneidman et al., 2006), gene interaction networks (Lezon et al., 2006), text classification in natural language processing (Ratnaparkhi, 2011), and society and economics (Foley, 1994). A recent review of the maximum entropy approach is given in Golan and Harte (2022).

For easier understanding, the basic structure of a maximum entropy model is presented as:

$$\begin{aligned} &\max_p h(p) = -\sum_x p(x)\ln\ p(x) \\ &\text{s.t.} \sum_x p(x) f_j(x) = \overline{f}_j, \quad \text{for } j = 1, 2, \ldots \\ &\quad \sum_x p(x) = 1 \end{aligned} \tag{11}$$

where $p(x) = P(X = x)$ is the probability distribution of a discrete random variable X; $h(p)$ is the entropy of $p(x)$; and $\overline{f}_j$ is the j-th measured mean value of the characteristic function $f_j(x)$. The objective is to maximize the entropy, subject to the constraints of all $\overline{f}_j$ from prior knowledge, and $p(x)$ should be normalized to one as a distribution. The unique solution (Csiszar, 1991; Shore and Johnson, 1980) for the optimization problem above is as

$$p(x) = \frac{1}{z} e^{-\sum_j \lambda_j f_j(x)} \tag{12}$$

where z is the appropriate normalization factor and λ_j is the Lagrange multiplier that can be calibrated with maximum likelihood estimation. Since maximizing entropy subject to expectation constraints and maximizing likelihood given structural constraints on the distribution are proven to be convex duals of each other and thus equivalent (Grendar, 2001; Jaynes, 1982; Wainwright and Jordan, 2007). Note that the same procedure is also applicable for more complex problems in which the distribution is multivariable or conditional on other factors.

4.2. Maximum entropy-based conditional distribution

In this section, we apply the maximum entropy approach in the case of microscopic interaction represented by $P(V_i|V_{i-1})$. To capture the essence of the driving behavior, we consider three characteristic functions that measure the local interaction, the speed and speed fluctuation of the subject vehicle, following the studies of bird flocks (Bialek et al., 2012) and CF models (Brackstone and McDonald, 1999).

Specifically, the interaction term $f_{\mathrm{int}} = (v_i - v_{i-1})^2$ quantifies the speed difference between the leader and the follower, which describes the tendency of the follower to adjust its speed to follow the leader. The speed term $f_{\mathrm{own}} = v_i$ is the speed of the follower, standing for its own state, which dominates the distribution once the other terms are zero. The fluctuation term $f_{\mathrm{flu}} = (v_i - \overline{v})^2$ measures an individual vehicle's speed fluctuation compared with the mean value of the platoon $\overline{v}$, which is a parameter of the external traffic environment in which each leader–follower pair is involved.

Substituting the determined characteristic functions into Eq. (12), we have the conditional distribution:

$$P(V_i = v_i|V_{i-1} = v_{i-1}) = \frac{1}{z} e^{-\left(\lambda_1 f_{\mathrm{int}} + \lambda_2 f_{\mathrm{own}} + \lambda_3 f_{\mathrm{flu}}\right)} = \frac{1}{z} e^{-\lambda_1 (v_i - v_{i-1})^2 - \lambda_2 v_i - \lambda_3 (v_i - \overline{v})^2} \tag{13}$$

The expression of the normalization factor z is

$$z = \sum_{v_i=0}^{v_{max}} e^{-\lambda_1 (v_i - v_{i-1})^2 - \lambda_2 v_i - \lambda_3 (v_i - \overline{v})^2} \tag{14}$$

With $\mu_1 = \frac{-2\lambda_1 v_{i-1} + \lambda_2 - 2\lambda_3 \overline{v}}{2(\lambda_1 + \lambda_3)}$, $\mu_2 = -\frac{(-2\lambda_1 v_{i-1} + \lambda_2 - 2\lambda_3 \overline{v})^2}{4(\lambda_1 + \lambda_3)^2} + \frac{\lambda_1 {v_{i-1}}^2 + \lambda_3 \overline{v}^2}{\lambda_1 + \lambda_3}$, the summation in Eq. (14) is given below, where erf is the Gauss error function:

$$z = \frac{e^{-(\lambda_1+\lambda_3)\mu_2}}{\sqrt{\lambda_1+\lambda_3}} \frac{\sqrt{\pi}}{2} \left[\mathrm{erf}\left(\sqrt{\lambda_1+\lambda_3}\,(v_{\max}+\mu_1)\right) - \mathrm{erf}\left(\sqrt{\lambda_1+\lambda_3}\,\mu_1\right)\right] \tag{15}$$

where λ_1, λ_2 and λ_3 are parameters that can be further calibrated with maximum likelihood estimation (Bender and Orszag, 1999), which is formulated below, and the superscript (j) denotes the j-th sample of a leader–follower pair:

$$\max_{(\lambda_1,\lambda_2,\lambda_3)} \ln \prod_j P\left(V_i = v_i^{(j)} \middle| V_{i-1} = v_{i-1}^{(j)}\right) \tag{16}$$

Substituting Eq. (13) into Eq. (16), we have

$$\min_{(\lambda_1,\lambda_2,\lambda_3)} \sum_j \lambda_1 \left(v_i^{(j)} - v_{i-1}^{(j)}\right)^2 + \lambda_2 v_i^{(j)} + \lambda_3 \left(v_i^{(j)} - \bar{v}^{(j)}\right)^2 + \ln z^{(j)} \tag{17}$$

With the analytic expression of z, Eq. (17) can be solved as a standard unconstrained optimization problem to obtain $\lambda_1, \lambda_2, \lambda_3$ and thus the complete model of $P(V_i|V_{i-1})$.

Remark 2. The choice of appropriate characteristic functions is a challenging task in applying the maximum entropy approach (De Martino and De Martino, 2018; Golan and Harte, 2022). Generally, suitable constraints are constructed on the basis of domain knowledge. In this study, we carefully select three physically meaningful characteristic functions, as per the work in flocks of birds (Bialek et al., 2012, 2014). Moreover, the SFD model derived from the formulated maximum entropy distribution is validated with empirical datasets showing decent performance, as shown in Section 5. We can also derive an equivalent dynamic model, as shown in Eq. (18), following Brownian dynamics, which has a similar form to the CF model. All these results justify the effectiveness of the formulated maximum entropy distribution for microscopic behavior.

$$\frac{\mathrm{d}v_i(t)}{\mathrm{d}t} = -2\lambda_1(v_i - v_{i-1}) - \lambda_2 - 2\lambda_3(v_i - \bar{v}) + \sqrt{2}B(t) \tag{18}$$

4.3. Resulting SFD model

Based on the maximum entropy distribution for the probabilistic microscopic model, we can derive a concrete expression of SFD. Following the general framework, we insert Eq. (13) into Eq. (5), and the joint distribution is given below, where the normalization term is omitted to save space:

$$P(V_1 = v_1, \ldots, V_n = v_n | V_0 = v_0) \propto e^{-\left[\sum_{i=1}^{n} \lambda_1 (v_i - v_{i-1})^2 + \sum_{i=1}^{n} \lambda_2 v_i + \sum_{i=1}^{n} \lambda_3 (v_i - \bar{v})^2\right]} \tag{19}$$

Under equilibrium conditions, all vehicles take identical speed values simultaneously in each instance, which is also the value of the average speed of the platoon, and their speeds are equal in distribution to the equilibrium speed according to Proposition 1. Following Eqs. (6)–(8), we have the distribution of the equilibrium speed:

$$P(V = v_e) = \frac{e^{-\lambda_2 n v_e}}{\sum_{v_e=0}^{v_{\max}} e^{-\lambda_2 n v_e}} = \frac{\lambda_2 kl e^{-\lambda_2 kl v_e}}{1 - e^{-\lambda_2 kl v_{\max}}} \tag{20}$$

Remark 3. Note that the elimination of λ_1, λ_3 facilitates the theoretical analysis, while these terms are necessary for the microscopic probabilistic model and may be utilized in the stability analysis of the traffic system (Qian et al., 2017b), which is believed to be closely related to the capacity drop (Qian et al., 2017a).

By substituting Eq. (20) into Eqs. (9) and (10), the mean and variance of flow Q with respect to density k are

$$\mathrm{E}[Q] = \frac{1}{\lambda_2 l} - \frac{k v_{\max} e^{-\lambda_2 v_{\max} kl}}{1 - e^{-\lambda_2 v_{\max} kl}} \tag{21}$$

$$\mathrm{Var}[Q] = \frac{1}{{\lambda_2}^2 l^2} - \frac{k^2 {v_{\max}}^2}{e^{-\lambda_2 v_{\max} kl} + e^{\lambda_2 v_{\max} kl} - 2} \tag{22}$$

A sketch plot of the resulting FD is shown in Fig. 3, with a set of example parameters. The main feature of the flow–density relation, i.e., the well-known quadratic curve, is reproduced in Fig. 3a. We examine the properties of the model by computing the limits of the mean flow and its variance. When the density approaches zero, the flow expectation is zero, as shown in Eq. (23), which is natural, as the highway is almost empty. When the density approaches infinity, the flow expectation also needs to be zero, as shown in Eq. (24), as nearly all vehicles come to a complete stop, implying that $\lambda_2 \to \infty$ when $k \to +\infty$. This suggests that λ_2 is a function of spacing and density.

$$\lim_{k\to 0} \mathrm{E}[Q] = \frac{1}{\lambda_2 l} - \frac{1}{\lambda_2 l} = 0 \tag{23}$$

$$\lim_{k\to+\infty} \mathrm{E}[Q] = \frac{1}{\lambda_2 l} = 0 \tag{24}$$

Moreover, the variance of flow with respect to density in Fig. 3b reflects the uncertainty of traffic flow, which stems from the proposed probabilistic microscopic model. As observed in empirical data (Daganzo, 2002; Kerner and Rehborn, 1996), the flow variance is relatively low at a low density because vehicles can drive in free flow with a small speed variance, which results in a small flow variance according to $Q = kV$. On the other hand, under extremely congested conditions, the flow variance and speed variance also approach zero, as all vehicles are forced to stop. The limits of flow variance calculated in Eqs. (25) and (26) are consistent with the above discussion, assuming that $\lim_{k\to+\infty} \lambda_2 = \infty$.

$$\lim_{k\to 0} \mathrm{Var}[Q] = \frac{1}{{\lambda_2}^2 l^2} - \frac{1}{{\lambda_2}^2 l^2} = 0 \tag{25}$$

$$\lim_{k\to+\infty} \mathrm{Var}[Q] = \frac{1}{{\lambda_2}^2 l^2} = 0 \tag{26}$$

Furthermore, empirical evidence suggests that the maximum flow variance usually occurs shortly after the peak flow expectation (Ni, 2021), which can also be observed in Fig. 3.

5. Case study

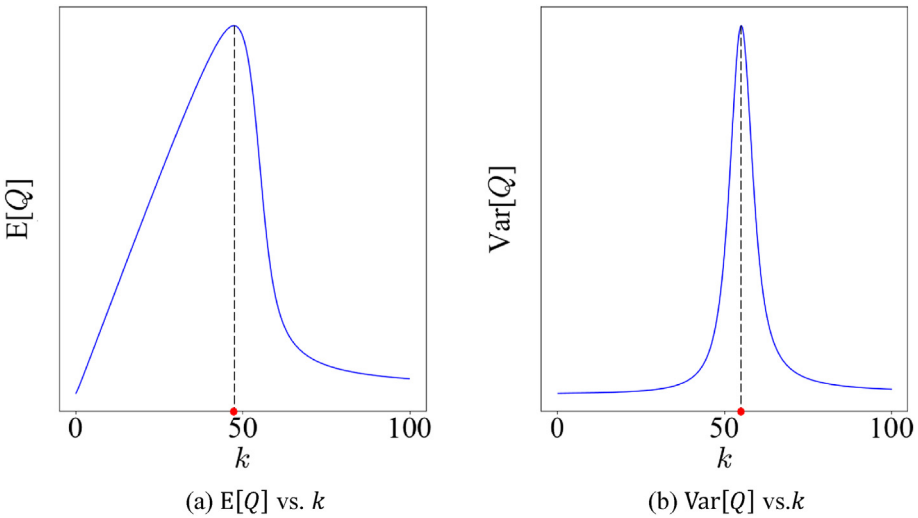
In this section, we use the NGSIM and HighD datasets to thoroughly examine the proposed maximum entropy-based SFD model and assess its ability to reproduce observed data. Specifically, we first extract CF pairs and calibrate the Lagrange multipliers in the maximum entropy distribution of the microscopic model. We then validate the estimated mean and variance of the flow–density relation derived from the microscopic model with empirical data.

5.1. Experiments on the NGSIM I-80 dataset

5.1.1. Data preparation

We first estimate the SFD model using the reconstructed NGSIM I-80 dataset, where the inconsistencies and noise in the raw data are eliminated via the reconstruction procedure (Montanino and Punzo, 2015). The data were collected along Interstate I-80 in the San Francisco Bay area during the period of 4:00 p.m.–4:15 p.m. on April 13, 2005. There are 6 lanes in the study area, including a left-most high-occupancy vehicle (HOV) lane, 4 regular lanes, and a shoulder lane with on-ramp and off-ramp. To avoid abnormal CF states caused by the HOV lane and merging and diverging behaviors, only the trajectory data on the central lanes (lanes 2, 3, and 4) are considered. A total of 481,602 CF samples were extracted for microscopic model calibration.

For the validation of the SFD model, we aggregate the trajectory data within lanes 2–4 upstream of the on-ramp bottleneck to capture



(a) E[Q] vs. k (b) Var[Q] vs.k

Fig. 3. SFD with example parameters ($l = 1, v_{\max} = 10$, and λ_2 is assumed to be proportional to ln$k$).

relatively stable traffic flow dynamics. The spatial interval for aggregation is 100 m, ranging from 129 to 229 m from local Y, and the time interval for aggregation is 10 steps (1 s). Notably, the aggregation interval should not be too large to be sensitive to traffic congestion; otherwise, the traffic congestion will be smoothed out, which fails to reproduce the complete traffic states. Moreover, the aggregation interval should not be too small to provide near-equilibrium traffic states and reasonable resolution along the density axis. Additionally, there should be enough data points for the calculation of the expectation and variance. In practice, multiple aggregation intervals are examined until the resulting plot satisfies all the criteria. The density is computed by dividing the average total number of vehicles per lane by the length of the spatial interval and time interval, and the flow is the product of the density and the average speed of all vehicles in the chosen spatial interval at each time interval. The empirical FD is shown in Fig. 4, which consists of 920 data points in total. In line with existing empirical studies (Lu et al., 2009; Ni, 2021), the flow expectation increases from zero as the density increases, reaches its maximum, and then begins to decrease. Moreover, the flow variance increases from zero with increasing density and peaks at a higher density (approximately 60 veh/km/lane) than the flow expectation (approximately 50 veh/km/lane).

5.1.2. Calibration of microscopic parameters

We calibrate the Lagrange multipliers λ_1, λ_2 and λ_3 in the microscopic probabilistic model via the maximum likelihood estimation formulated in Eq. (16). Two-thirds of the extracted CF data are randomly selected for calibration, and the rest are used for validation. Considering that the parameters are potentially dependent on spacing, as discussed before, we further divide the training data into subsets according to the different spacing intervals. Specifically, the spacing is discretized with a small step (0.5 m), and the training data in each spacing interval form a subset for calibration. Consequently, for each spacing, the estimated values of λ_1, λ_2, and λ_3 in Eq. (19) can be obtained as illustrated in Fig. 5.

As these Lagrange multipliers λ_1, λ_2, and λ_3 can be regarded as the corresponding gains for the interaction, own speed, and fluctuation terms, respectively, Fig. 5a reveals that with increasing spacing, λ_1 sharply decreases, which is expected because the interaction of the leader–follower pair deteriorates as the spacing increases. λ_2 in Fig. 5b decreases to a negative value when the spacing increases. This can be understood by recalling the CF-like dynamic model derived from the maximum entropy distribution of the microscopic model, as shown in Eq. (18), where $-\lambda_2$ is attributable to acceleration. Therefore, a decrease in positive values means a smaller deceleration, and a decrease in negative value means a greater acceleration, which both denote a greater desire to speed up as the spacing increases. λ_3 appears to be a piecewise function of

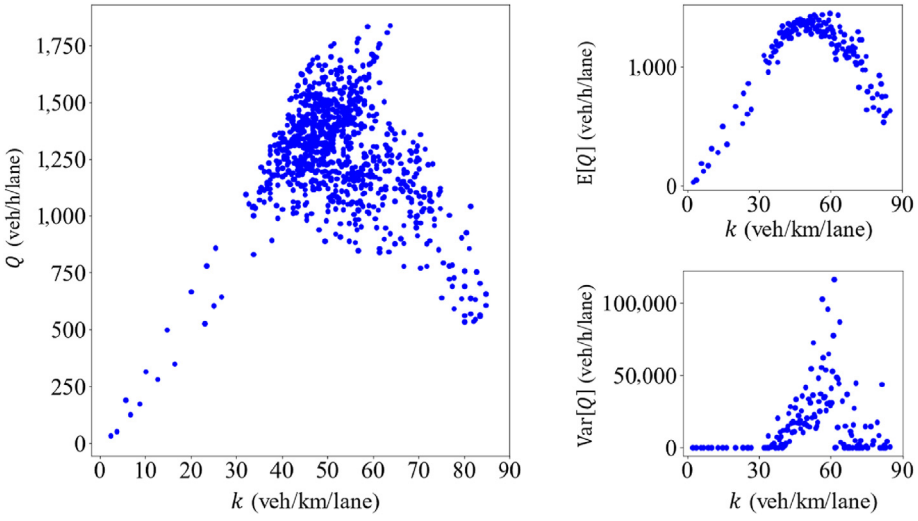
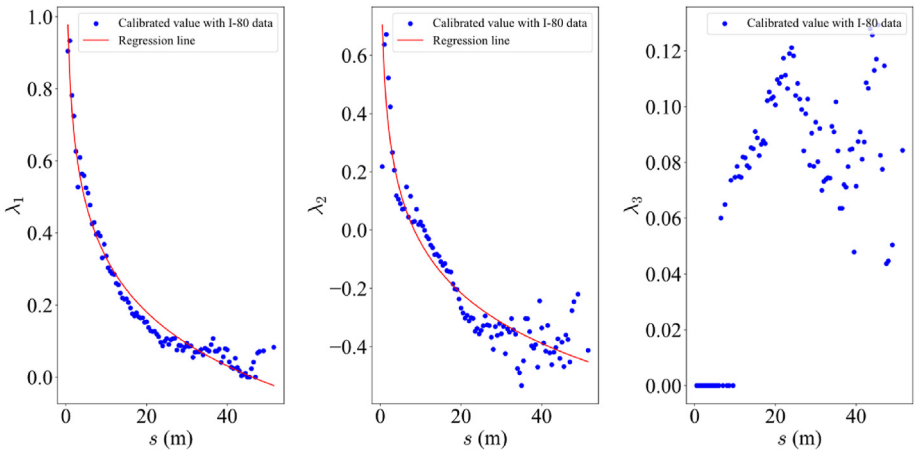


Fig. 4. Empirical FD using the reconstructed I-80 dataset.



(a) λ_1 vs. s (b) λ_2 vs. s (c) λ_3 vs. s

Fig. 5. Relationship between λ_i $(i = 1, 2, 3)$ and the spacing s in the I-80 dataset.

spacing, as shown in Fig. 5c, where for small spacings (less than 8 m), λ_3 remains zero, and for larger spacings, λ_3 fluctuates approximately 0.09. One possible reason is that the fluctuation term is caused by the external traffic environment, which is not dominant and has no impact with small spacings.

Fig. 5 clearly shows that a logarithmic relationship exists between λ_1, λ_2 and the spacing s, which can be used to approximate λ_1, λ_2, as formulated in Eqs. (27) and (28), where η, θ, α, and β are parameters that need to be calibrated and $\eta > 0$, where $\alpha > 0$. For simplicity, λ_3 is fixed to 0.09, which does not affect the results of the SFD.

$$\lambda_1 = -\eta \ln s + \theta \tag{27}$$

$$\lambda_2 = -\alpha \ln s + \beta \tag{28}$$

The parameters obtained via linear regression are presented in Table 2, which shows good fitting performance with high R^2 values (near or greater than 0.90). This validates the proposed functional form of λ_1, λ_2, which is also plotted in Fig. 5 for comparison. Moreover, λ_2 in Eq. (28) meets the previous assumption that $\lim_{k\to+\infty} \lambda_2 = \infty$, because $\lim_{k\to+\infty} \lambda_2 = \lim_{k\to+\infty} -\alpha \ln s + \beta = \lim_{k\to+\infty} \alpha \ln k + \beta = \infty$ is under equilibrium.

With the calibrated parameters, the probabilistic model $P(V_i|V_{i-1})$ is assessed with the validation dataset. Since the conditional distribution $P(V_i|V_{i-1})$ varies with the spacing and average speed of the subject lane, as implied in Eq. (13), the validation dataset is divided into multiple subsets according to the two factors and the speed of the leading vehicle. Additionally, given that the expectation is a commonly used measure to depict the probability distribution, it is adopted to validate the proposed conditional distribution. In each subset j with a given speed of leading vehicle $v_{i-1}^{(j)}$, the theoretical conditional expected speed of the subject vehicle, as formulated in Eq. (29), is compared with the empirical speed $\bar{v}_i^{(j)}$, which is the mean value of the speed of the subject vehicles in each subset.

Table 2
Calibration results of λ_1 and λ_2 in the I-80 dataset.

Parameter		Value (± standard deviation)	R^2
λ_1	η	0.235 ± 0.000	0.98
	θ	0.880 ± 0.001	
λ_2	α	0.283 ± 0.003	0.89
	β	0.779 ± 0.012	

$$\mathrm{E}\left[V_i \middle| V_{i-1} = v_{i-1}^{(j)}\right] = \sum_{v_i=0}^{v_{\max}} v_i P\left(V_i = v_i \middle| V_{i-1} = v_{i-1}^{(j)}\right) \tag{29}$$

The empirical and theoretical conditional expectations of speed for each subset are placed on the x-axis and y-axis of Fig. 6, respectively. As indicated by Fig. 6, the theoretical results are highly consistent with the empirical data, as most of the data points are close to the red line of $y = x$.

To further quantify the performance, the mean absolute percentage error (MAPE) is calculated:

$$\mathrm{MAPE}_{\exp} = \frac{1}{m} \sum_{j=1}^{m} \left| \frac{\mathrm{E}\left[V_i \middle| V_{i-1} = v_{i-1}^{(j)}\right] - \bar{v}_i^{(j)}}{\bar{v}_i^{(j)}} \right| \tag{30}$$

where m is the total number of subsets. The result of $\mathrm{MAPE}_{\exp}$ is 5.23%, which is superior to the best result of 6.29% in Bai et al. (2021), in which

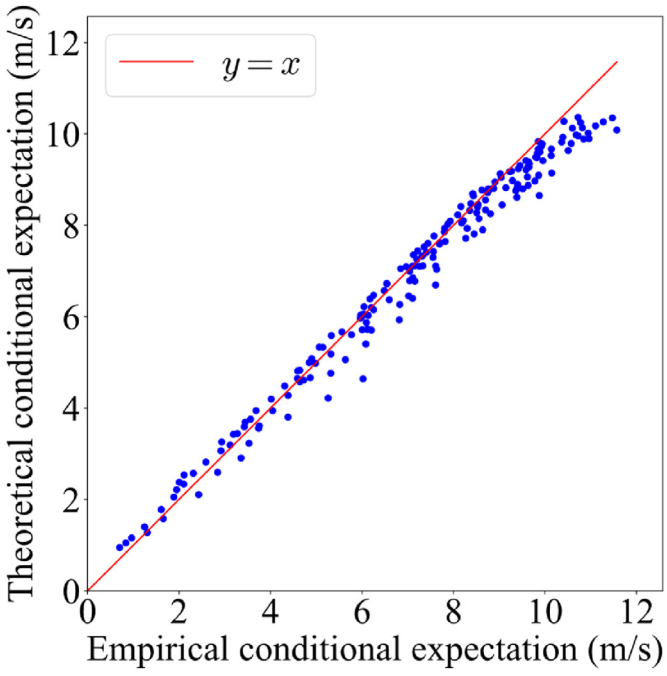


Fig. 6. Validation results of the maximum entropy distribution in the I-80 dataset.

a stochastic structure is proposed to consider the speed heterogeneity in FD and predict the mean speed under varying conditions. These results demonstrate the good performance of the proposed maximum entropy distribution of the microscopic model.

5.1.3. Estimation of SFD

On the basis of the previously derived SFD model (Eqs. (21) and (22)) and the function λ_2, the theoretical flow expectation and flow variance are formulated as

$$\mathrm{E}[Q] = \frac{1}{(\alpha \ln k + \beta) l} - \frac{k v_{\max} e^{-(\alpha \ln k + \beta) v_{\max} k l}}{1 - e^{-v_{\max} k l (\alpha \ln k + \beta)}} \tag{31}$$

$$\mathrm{Var}[Q] = \frac{1}{(\alpha \ln k + \beta)^2 l^2} - \frac{k^2 {v_{\max}}^2}{e^{-v_{\max} k l (\alpha \ln k + \beta)} + e^{v_{\max} k l (\alpha \ln k + \beta)} - 2} \tag{32}$$

There are a total of 4 parameters $(l, \alpha, \beta, v_{\max})$ in the above equations that need to be determined, where l is the length of aggregation, which is 100 m in this study; α and β are the calibrated parameters from the microscopic model; and $v_{\max}$ is the maximum speed of the study area. Given that the 85th percentile speed can cover most of the empirical velocities and represent a reasonable and realistic maximum speed, it is commonly utilized as the speed limit in practice (Forbes et al., 2012). We thus adopt the 85th percentile speed in the I-80 dataset as the maximum speed $v_{\max}$ to avoid anomalies, namely, $v_{\max} = 10.2$ m/s (36.6 km/h).

Consequently, the estimated SFD functions are presented in Fig. 7, which accords with the main features of the parabola-like observed data, and most of the empirical traffic data fall into the estimated standard deviation range. Additionally, we calculate the R^2 value for both flow expectation and variance to measure the goodness of fit. The overall R^2 value is 0.82, indicating that the observed FD has been reproduced effectively.

5.2. Experiments on the NGSIM US-101 dataset

Although the maximum entropy distribution for microscopic behavior and the micro-macro link is general, the function of $\lambda_2(s)$ in Eq. (28) is empirical and only observed in the I-80 dataset. To further examine the generality of the function of $\lambda_2(s)$ and the SFD model based on $\lambda_2(s)$, we apply the same pipeline on the US-101 dataset to calibrate our models and examine its ability to reproduce the observed flow-density data and its variance.

Similarly, we extract the vehicle trajectories from four lanes (lanes 1–4, excluding the lanes near ramps) in the US-101 dataset and obtain 759,756 CF samples in total to calibrate Lagrange multipliers via maximum likelihood estimation for different spacings. As shown in Fig. 8a, a logarithmic relationship between λ_2 and spacing s can be found, which validates the function of $\lambda_2(s)$ proposed on the basis of the I-80 dataset (to save space, we present only the results of $\lambda_2(s)$,; the results of λ_1 and λ_3 are also consistent with those in the I-80 dataset). The parameters α and β in the function of $\lambda_2(s)$ in Eq. (28) are then calibrated via linear regression, and the results are presented in Table 3.

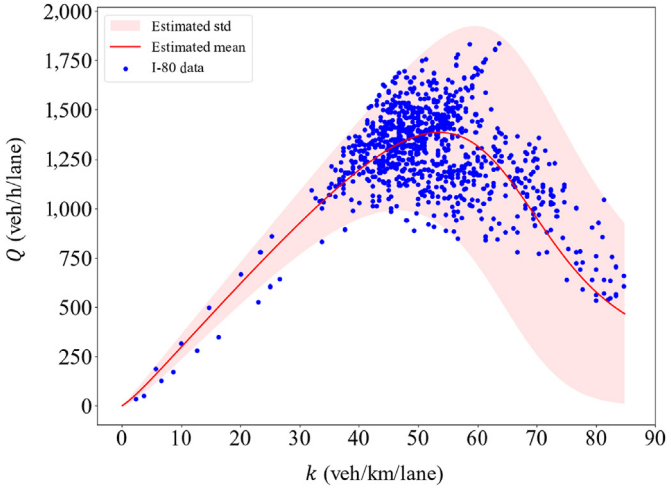


Fig. 7. Estimated SFD in the I-80 dataset.

Moreover, the empirical FD of US-101 is obtained by aggregating all the data within lanes 1–4 upstream of the bottleneck, ranging from 150 to 250 m from local Y. Same with previous settings, the spatial interval for aggregation is 100 m, and the time interval for aggregation is 10 steps (1 s). The 85th percentile speed in the US-101 dataset is used as the maximum speed, namely, $v_{\max} = 15.3$ m/s (54.9 km/h). With the calibrated α and β the flow expectation and variance with respect to density are estimated via Eqs. (31) and (32). The resulting SFD is illustrated in Fig. 8b. Compared with the observed data, the R^2 value is 0.85, which also demonstrates good consistency and thus further justifies the proposed SFD model.

5.3. Experiments on the HighD dataset

To further examine the performance of the proposed model, another naturalistic vehicle trajectory dataset recorded on German highways using drones, the HighD dataset (Krajewski et al., 2018), is utilized. The highD dataset has a relatively high proportion of trucks (approximately 23%), and the positioning error is typically less than 10 cm. There are 60 recordings with an average length of 17 min (16.5 h in total) covering a road segment of approximately 420 m in length, where most traffic states are free flow and thus quite different from those in the NGSIM dataset.

We choose the upper road (driving direction = 1) of the 36th recording, including high-density flow, to obtain a relatively complete FD result. Following the same procedure, we first extract 233,266 CF samples to calibrate $\lambda_1, \lambda_2, \lambda_3$ under different spacings. The relationship between λ_2 and the spacing s is presented in Fig. 9a, which still indicates a logarithmic relation. The coefficients in $\lambda_2(s)$, i.e., α and β in Eq. (28), are estimated, and the results are shown in Table 4.

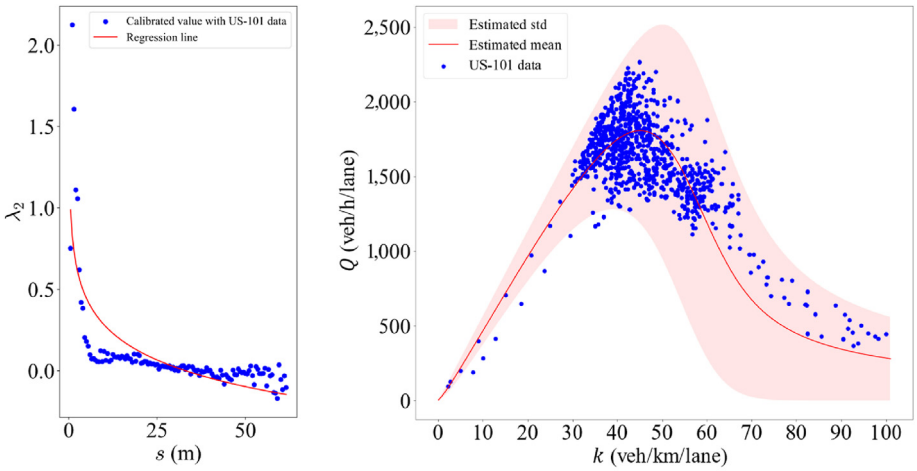
For the empirical FD, all vehicle data within the upper road (lanes 2–4) in the range of 60–160 m along the longitudinal axis (local X) are aggregated. The spatial interval and the time interval for aggregation are 100 m and 10 steps, respectively. The 85th percentile speed in the HighD dataset is adopted, i.e., $v_{\max} = 31.7$ m/s (114.2 km/h). With the aggregated spatial interval $l = 100$ m and the calibrated α and β , the SFD is estimated via Eqs. (31) and (32). As illustrated in Fig. 9b, the estimated FD is highly consistent with the empirical data, and the overall R^2 reaches 0.84, validating the capability of the proposed framework under the traffic conditions of HighD.

6. Discussion

The results above fully demonstrate the effectiveness of the proposed methodology, which analytically links the probabilistic microscopic model and the SFD. In this section, we further discuss several related issues of the SFD model. We first investigate the impact of traffic flow heterogeneity on the SFD model within different datasets. Additionally, we explore the physical meanings of the parameters and their potential influences on the SFD model. Finally, we present some potential applications of the SFD model.

6.1. Traffic flow heterogeneity

Heterogeneity has been regarded as one critical reason for the scattering pattern in the flow-density data. Previous studies have demonstrated the impact of heterogeneity in vehicle classes on the wide-scattering effect (Treiber and Helbing, 1999) and suggested that heterogeneous driving styles such as timid and aggressive drivers can also lead to scattering in FDs (Daganzo, 2002). Obviously, more heterogeneity



(a) λ_2 vs. s (b) Estimated SFD

Fig. 8. Results for the US-101 dataset.

Table 3
Parameters for the US-101 dataset.

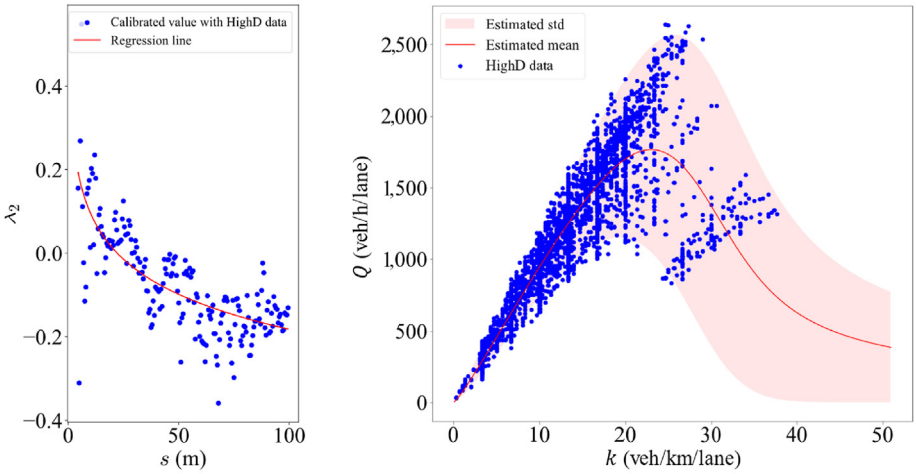
Parameter		Value	R^2 of the estimated SFD
λ_2	α	0.209 ± 0.004	0.85
	β	0.600 ± 0.017	
l		100 m	
v_{max}		15.3 m/s	

leads to wider scattering in the FD. Therefore, the observed flow–density variance of the HighD dataset is apparently larger than that of the NGSIM dataset because of the high proportion of heavy load vehicles in the HighD dataset (23% compared with 5% and 3.5% in the I-80 and US-101 datasets, respectively). In addition, more fluctuations are found in the calibrated λ_2 in the HighD dataset.

Although various heterogeneities exist in different datasets, the proposed SFD model can reproduce all the observed data with high R^2 values greater than 0.8, indicating the ability to address heterogeneous traffic flows. However, this theoretical model considers only the uncertainty of driving behavior and assumes homogeneous driver–vehicle units with identical conditional distributions for local interactions instead of the explicit consideration of heterogeneity within drivers and vehicle types. This suggests that the stochastic model might capture multiple sources of heterogeneity implicitly and that the heterogeneous system might be mapped onto an effective homogeneous system (Venkata Ramana and Jabari, 2020). While the proposed framework is sufficient for modeling the uncertainty of macroscopic traffic flow, an extension to explicitly accommodate heterogeneity (e.g., different Lagrange multipliers $\lambda_1^i, \lambda_2^i, \lambda_3^i$ for each vehicle i) can be implemented in future work to improve the SFD model.

Table 4
Parameters for the HighD dataset.

Parameter		Value	R^2 of the estimated SFD
λ_2	α	0.137 ± 0.008	0.84
	β	0.493 ± 0.031	
l		100 m	
v_{max}		31.7 m/s	



(a) λ_2 vs s (b) Estimated SFD

Fig. 9. Results for the HighD dataset.

6.2. Analysis of parameters

In this study, the analytic functions of the mean flow density and its variance are given in Eqs. (31) and (32) and validated via empirical data. These functions contain four parameters, including l, v_{max}, α, and β, where l and v_{max} are rather fixed in a certain traffic environment, whereas α and β denote the relationship between the gain of the own speed term λ_2 and the spacing s in the microscopic model and thus describe the microscopic behavior. Therefore, to study the relationship between microscopic behavior and the SFD, we demonstrate how the macroscopic relations vary by changing the microscopic parameters α and β and provide a potential interpretation for these parameters.

We first implement the sensitivity analysis of SFD in terms of parameter α while fixing other parameters. In Fig. 10, the direction of the arrow indicates that α increases from 0.262 to 0.292 with a step of 0.01. The results of the flow expectation in Fig. 10a illustrate that both the capacity and the critical density decrease as α increases. Moreover, the flow variances increase steadily in the free flow branch, as presented in Fig. 10b, whereas the maximum flow variances in the congested branch decrease with decreasing capacity.

To gain intuitions for interpreting parameter α, we recall the equivalent dynamic model of the maximum entropy distribution for local interaction in Eq. (18), while the white noise term is ignored here for the purpose of simplicity:

$$a_i = -2\lambda_1(v_i - v_{i-1}) - \lambda_2 - 2\lambda_3(v_i - \bar{v}) \tag{33}$$

where a_i denotes the acceleration of the subject vehicle.

Considering the situation of negligible speed difference ($v_i - v_{i-1} \approx 0$) and small spacing ($\lambda_3 \approx 0$), the acceleration is governed by $-\lambda_2$. The partial derivative of acceleration in terms of the spacing can be approximated as Eq. (34), where the minimum distance is set to $s_{min} = 1$ m and $\alpha > 0$:

$$\frac{\partial a_i}{\partial s_i} \approx \frac{\partial(-\lambda_2)}{\partial s_i} = \frac{\partial(\alpha \ln s_i - \beta)}{\partial s_i} = \frac{\alpha}{s_i} \leq \frac{\alpha}{s_{min}} = \alpha \tag{34}$$

Given that the acceleration is highly sensitive to the distance under this case, α can be interpreted as the maximum spacing gain, where the calibrated values of α (0.283, 0.209, 0.137 s^{-2} for the I-80, US-101 and HighD datasets, respectively) accord with the feasible range of 0–0.3 s^{-2} in existing findings (de Souza and Stern, 2021). The relatively lower value of the maximum spacing gain in the HighD dataset could be attributed to the greater spacing under higher speeds on the German highway, with an average spacing of 47 m compared with 14 and 19 m in the I-80 and US-101 datasets, respectively.

Under this interpretation, a larger α indicates greater sensitivity to the spacing, which leads to greater traffic flow instability, as implied in the general string stability criterion (Sun et al., 2018). Hence, the phase transition from free flow to congested flow is triggered earlier, leading to a decrease in the critical density and capacity, as shown in Fig. 10a, and an increase in the flow variance in the free flow branch, as shown in Fig. 10b.

Similarly, the sensitivity analysis is implemented for parameters β from 0.755 to 0.785 with a step of 0.01. From Fig. 11, similar trends can be found for increasing α. A larger β results in an earlier phase transition and decreasing capacity, as shown in Fig. 11a, and larger fluctuations in the free-flow branch, as shown in Fig. 11b. A possible interpretation of β is that the model will not brake harder than β, which refers to the comfortable deceleration when the acceleration is dominated by $-\lambda_2$ and $s_{min} = 1$ m:

$$a_i \approx -\lambda_2 = \alpha \ln s_i - \beta \geq \alpha \ln s_{min} - \beta = -\beta \tag{35}$$

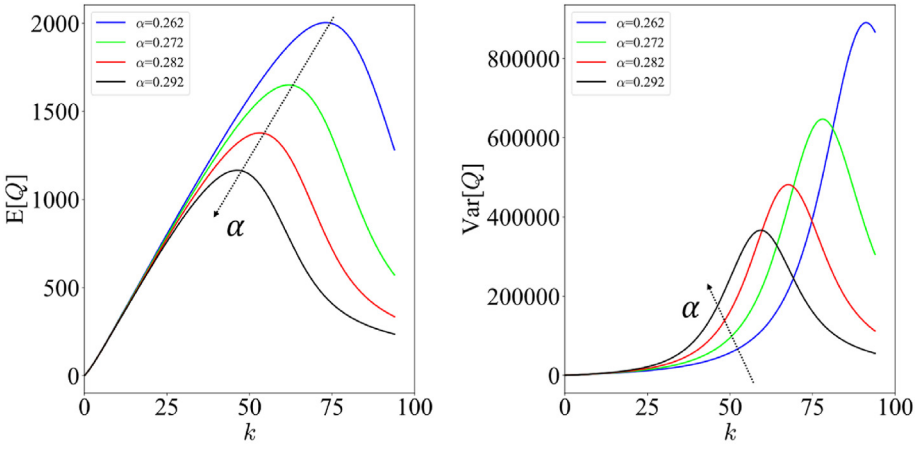
Thus, the results of the SFD with a larger β can be explained by the fact that harder brake behaviors facilitate the propagation of traffic oscillation, leading to lower capacity and greater stochasticity. In this study, the calibration results of β are 0.779, 0.600, and 0.493 m/s^2 for the I-80, US-101, and HighD datasets, respectively, which are rational values for comfortable deceleration. The smaller β in the HighD dataset may be due to the higher speed on the German highway, whereas the on-ramp and off-ramp bottlenecks in the NGSIM dataset lead to harder brakes with more stop-and-gos.

In summary, larger α and β values have devastating effects on traffic flow, including decreasing capacity, advancing phase transition, and increasing fluctuation. Notably, even a slight change such as 0.01 leads to a large difference, especially for α, as the macroscopic relation is more sensitive to α than to β. Furthermore, according to the simplified dynamic model, we can infer that α and β may reflect the maximum spacing gain and comfortable deceleration, respectively, which is consistent with the results of the parameter sensitivity analysis.

6.3. Applications and practice implications

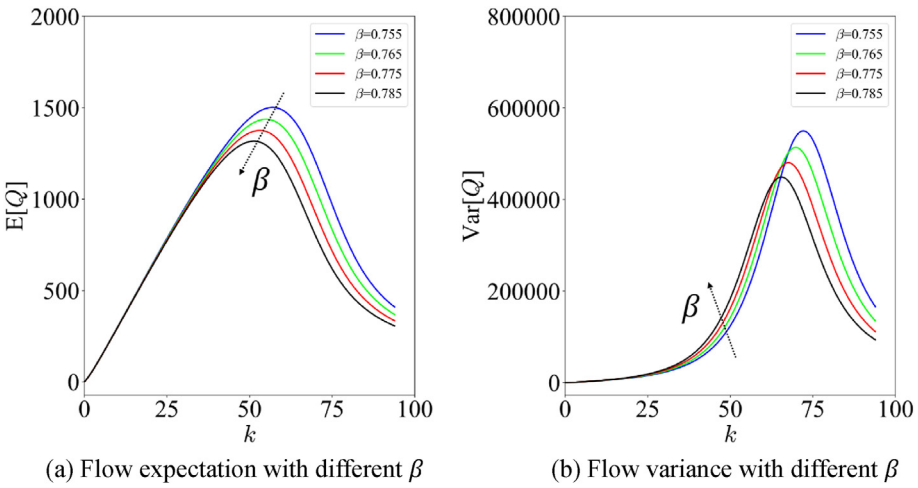
In addition to the proposed general LFCD-STM framework for bridging the micro–macro gap for SFD modeling, which can be applied once the microscopic behavior is described by a conditional distribution, the analytical maximum entropy-based SFD model has several potential applications in FD estimation, traffic flow modeling, and control practices.

First, as advanced monitoring technologies facilitate the accurate



(a) Flow expectation with different α (b) Flow variance with different α

Fig. 10. SFD for different values of α (increasing in the direction of the arrow).



(a) Flow expectation with different β (b) Flow variance with different β

Fig. 11. SFD for different values of β (increasing in the direction of the arrow).

collection of high-resolution trajectories of individual vehicles (Zhu et al., 2019), the proposed SFD model can thus be applied for real-time FD estimation to further aid in transport policy decision making. For example, with probe vehicle data from mobile technology and connected vehicles, our SFD model with only four physically meaningful parameters can be easily calibrated with an extended computational algorithm for FD estimation using trajectories of probe vehicles (Seo et al., 2019).

Second, as the SFD model proposed in this study enables predictable macroscopic characteristics that converge to known microscopic behavior, by combining our analytic model of the SFD and the conservation law ((which leads to the LWR model) (Lighthill and Whitham, 1955; Richards, 1956)) or other principles ((high-order models such as the Payne–Witham model) (Payne, 1971; Whitham, 1999)), more realistic macroscopic traffic flow dynamics can be easily simulated with the explicit consideration of uncertainty in traffic flow.

Furthermore, the SFD model can be utilized to devise and evaluate robust traffic control strategies. As the DFD cannot capture the complete traffic state in real traffic, traffic control strategies such as ramp metering (Papageorgiou and Kotsialos, 2002) may easily yield incorrect results, whereas SFD model-based control strategies are able to consider the majority of real-world scenarios and are thus more reliable. For example, the 85th percentile is usually used for traffic engineering (e.g., to determine the speed limit), and we can utilize the 85th percentile-based flow–density curve to model and assess traffic control strategies in an attempt to address 85% of the scenarios.

More importantly, such a micro-macroscopic-based SFD model is highly important for regulating microscopic driving behaviors to ease congestion. In particular, reducing the stochasticity of traffic flow can largely improve the performance of highway systems (Keyvan-Ekbatani et al., 2012; Qu et al., 2017), which can be achieved by reducing the maximum spacing gain and comfortable deceleration, as suggested in the parameter sensitivity analysis. To this end, concise messages such as "Drive Smoothly and Avoid Sudden Braking" on the variable message board can encourage drivers to adjust their speed gradually and brake gently. Moreover, educating new drivers in driving schools on the benefits of smooth acceleration and deceleration can help them understand how these practices contribute to a more efficient traffic system and thus foster such behaviors.

With the development of CAVs, the proposed micro-macroscopic based SFD model is also essential for analyzing mixed traffic flows and developing longitudinal control strategies to minimize stochasticity and enhance overall traffic. A direct yet effective way of controlling CAVs is to adjust the maximum spacing gain and the comfortable deceleration in the CF model according to the analysis of parameters in the proposed SFD model. Alternatively, the objective with an analytical function of the traffic flow variance can be embedded in the reward function, and the optimal policy can be learned through maximizing the overall reward via reinforcement learning (Sutton and Barto, 1998).

7. Conclusions

Despite the profound importance of FDs, analytical modeling of SFDs from a microscopic perspective that can accurately capture their macroscopic characteristics is still lacking. To fill this gap, this study develops a theoretical micro-macro framework, the LFCD-STM framework, to model SFDs by linking the probabilistic microscopic behavior with the equilibrium flow–density relation of traffic flow. A specific SFD model with explicit expressions of the mean and variance is then derived via the maximum entropy distribution for leader–follower behavior.

Specifically, we introduce a general conditional distribution of the follower's speed given the leader's speed to describe the longitudinal interaction among vehicles. On the basis of Markov chain modeling, the joint distribution of vehicle speeds representing the collective behavior of the platoon was derived. We thus obtained the distribution of the equilibrium speed under equilibrium conditions and derived general formulations of the mean flow–density relation and its variance. We then employed the maximum entropy distribution for the longitudinal interaction and obtained the explicit functions of the mean and variance of FD, which were calibrated and evaluated with the NGSIM and HighD datasets. The results indicate the high consistency of the theoretical results with the empirical data, which verifies the feasibility of the LFCD-STM framework and the maximum entropy-based SFD model.

The analytical SFD model involving microscopic parameters also enables us to explore the evolution of the macroscopic relation by analyzing the microscopic behavior, as the parameters are conjectured to denote the maximum spacing gain and the comfortable deceleration. This suggests that a larger maximum spacing gain or a greater comfortable deceleration would result in deteriorated traffic flow with decreasing capacity, advancing phase transition, and increasing fluctuation.

The proposed methodology can be further applied for estimating real-time FD, modeling macroscopic traffic flow, developing robust traffic control strategies, regulating human driving behaviors and guiding the development of CAV controllers for traffic flow optimization. Moreover, SFD models based on other conditional distributions of speed are possible because of the general micro-macro framework.

Several future directions based on this work can be explored. First, we can further extend the framework while considering perturbation propagation along the platoon to model the capacity drop phenomenon. This has been done in our following work. In addition, the assumption of a homogeneous speed distribution can be relaxed with different

parameters for distinct types of drivers and mixed traffic flows, including traditional and connected and automated vehicles. Moreover, incorporating the stability of CF behavior and/or lane-changing and analyzing their impacts on the macroscopic relations would be a promising direction.

CRediT authorship contribution statement

Xiaohui Zhang: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jie Sun:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Jian Sun:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Replication and data sharing

The code for the proposed model is available at https://github.com/tjzxh/Stochastic_fundamental_diagram_modeling.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is partially funded by the National Natural Science Foundation of China (52125208) and the Fundamental Research Funds for the Central Universities.

Appendix A. Connection between the conditional distribution and classical car-following models

Here, we demonstrate the transformation to the conditional distribution from a widely applied CF model, the optimal velocity relative velocity (OVRV) model (Liang and Peng, 1999; Milanes et al., 2014), as a showcase for its simplicity. The OVRV model is defined below, where ω_1 and ω_2 are the gains on the effective gap term and the relative velocity term, respectively; s_0 and t_h are the jam distance and desired time gap, respectively; $v_i(t)$ and $v_{i-1}(t)$ are the deterministic velocities of vehicle i and $i-1$ and $s_i(t)$ are the spacings in between.

$$\frac{\mathrm{d}v_i(t)}{\mathrm{d}t} = \omega_1(s_i(t) - s_0 - t_h v_i(t)) + \omega_2(v_{i-1}(t) - v_i(t)) \tag{A1}$$

where ω_1 and ω_2 are the gains on the effective gap term and the relative velocity term, respectively; s_0 and t_h are the jam distance and desired time gap, respectively; $v_i(t)$ and $v_{i-1}(t)$ are the deterministic velocities of vehicle i and $i-1$ and $s_i(t)$ are the spacings in between. To match Eq. (1), the Brownian motion term is added to the right side of Eq. (A1), which makes the velocity a random variable, as shown in Eq. (A2):

$$\frac{\mathrm{d}V_i(t)}{\mathrm{d}t} = \omega_1(s_i(t) - s_0 - t_h V_i(t)) + \omega_2(V_{i-1}(t) - V_i(t)) + \sqrt{2}B(t) \tag{A2}$$

Then, we have

$$\nabla \log P_{\mathrm{OVRV}}(V_i(t)|V_{i-1}(t)) = \omega_1(s_i(t) - s_0 - t_h V_i(t)) + \omega_2(V_{i-1}(t) - V_i(t)) \tag{A3}$$

Integrating both sides of Eq. (A3), the stationary distribution is derived and simplified as Eq. (A4):

$$P_{\mathrm{OVRV}}(V_i = v_i|V_{i-1} = v_{i-1}) = e^{\int \omega_1(s_i - s_0 - t_h v_i) + \omega_2(v_{i-1} - v_i)\mathrm{d}v_i} \propto e^{-\frac{1}{2}(\omega_1 t_h - \omega_2)\left(v_i - \frac{\omega_1(s_i - s_0) - \omega_2 v_{i-1}}{\omega_1 t_h - \omega_2}\right)^2} \tag{A4}$$

The stationary distribution of the OVRV model clearly follows a normal distribution:

$$P_{\mathrm{OVRV}}(V_i = v_i|V_{i-1} = v_{i-1}) \propto \mathcal{N}\left(\mu = \frac{\omega_1(s_i - s_0) - \omega_2 v_{i-1}}{\omega_1 t_h - \omega_2}, \sigma^2 = \frac{1}{\omega_1 t_h - \omega_2}\right) \tag{A5}$$

To justify the equivalence between the stationary distribution in Eq. (A5) and the stochastic OVRV model in Eq. (A2), we obtain sample sequences from Eq. (A2) via a commonly used discrete-time method as

$$v_i(t+1) = v_i(t) + \delta[\omega_1(s_i(t) - s_0 - t_h v_i(t)) + \omega_2(v_{i-1}(t) - v_i(t))] + \sqrt{2\delta} z_i \tag{A6}$$

where δ is a time step fixed to 0.1 in this study, each z_i is an independent draw from the standard normal distribution, the initial value $v_i(t=0)$ is arbitrarily fixed to 1 m/s, and other influencing factors, such as $s_i(t)$, and $v_{i-1}(t)$, are fixed. The histogram of the resulting sequences $(v_i(t))_{t=0}^{8000}$ from Eq. (A6) is compared with the conditional distribution in Eq. (A5), as shown in Fig. A1. Here, we only present a comparison under the conditions of $V_{i-1} = \{5\ \mathrm{m/s}, 10\ \mathrm{m/s}, 15\ \mathrm{m/s}, 20\ \mathrm{m/s}, 25\ \mathrm{m/s}\}$, and $s_i = 25$ m as examples, and the parameters of the OVRV model are $\omega_1 = \omega_2 = 0.5$, $s_0 = 8$ m, and $t_h = 1$ s. From Fig. A1, samples from the OVRV model are highly consistent with the derived conditional distribution, which validates the transformation between classical CF models and their stationary conditional distributions.

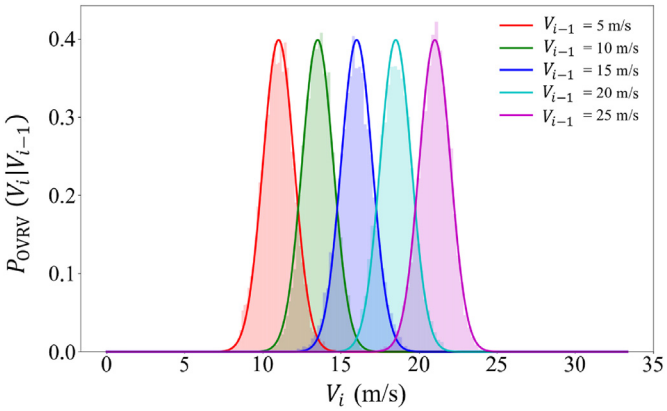


Fig. A1. Comparison between the OVRV model (histogram) and its conditional distribution (density plot).

Given the connection between classical CF models and the conditional distribution $P(V_i|V_{i-1})$, the corresponding SFD can also be derived for classical CF models through the LFCD-STM framework. Therefore, to examine the proposed general framework, we compare the derived SFD of classical CF models with their corresponding deterministic FD, which is supposedly the mean of the derived SFD.

Specifically, we still use the OVRV model for consistency. According to the equilibrium condition, the deterministic FD of the OVRV model can be derived with $v_{i-1}(t) = v_i(t) = v_e$, $s_i(t) = s_e = 1/k$:

$$s_e = s_0 + t_h v_e \tag{A7}$$

$$q = kv = \frac{1 - ks_0}{t_h} \tag{A8}$$

On the other hand, we can obtain the SFD of the OVRV model by substituting the conditional distribution of the OVRV model $P_{\mathrm{OVRV}}(V_i|V_{i-1})$, as shown in Eq. (A5), into Eqs. (4)–(10):

$$\mathrm{E}[Q] = \frac{1 - ks_0}{t_h} \tag{A9}$$

$$\mathrm{Var}[Q] = \frac{(\omega_1 t_h - \omega_2)k}{{\omega_1}^2 {t_h}^2 l} \tag{A10}$$

The mean of the derived SFD is perfectly consistent with the deterministic FD obtained from the original OVRV model, which provides strong evidence for the effectiveness and generality of the LFCD-STM framework in modeling SFD. The variance of the derived SFD mainly depends on the characteristics of the microscopic behavior, i.e., the acceleration gains with respect to the gap and relative speed and the desired time gap. The length of the study area l is also introduced, which is reasonable, as a larger length for aggregation generally results in a lower variance.

Appendix B. Proof of proposition 1

Proof by contradiction. We assume that any two speed distributions of vehicles are different in equilibrium. For the sake of concision, we define that the speed can only take two values, i.e., the sample space is $\{v_a, v_b\}$. By our assumption, $P(V_i = v_a) \neq P(V_j = v_a)$ and $P(V_i = v_b) \neq P(V_j = v_b)$ for any vehicle i and j, where $i \neq j$. With many experiments (τ times and $\tau \to \infty$) in equilibrium states, the expected frequencies of v_a for any vehicle i and j are $\tau P(V_i = v_a)$ and $\tau P(V_j = v_a)$, respectively. Since the equilibrium condition requires identical speeds for all the vehicles in the platoon, the frequency of v_a for each vehicle in the sampling equilibrium state should also be identical. Hence, we have $\tau P(V_i = v_a) = \tau P(V_j = v_a)$, leading to $P(V_i = v_a) = P(V_j = v_a)$. A contradiction. The speed of each vehicle in the platoon must be equal in distribution under equilibrium conditions, which is also the distribution of the equilibrium speed.

References

Ahmed, A., Ngoduy, D., Adnan, M., Baig, M.A.U., 2021. On the fundamental diagram and driving behavior modeling of heterogeneous traffic flow using UAV-based data. Transp. Res. Part A Policy Pract. 148, 100–115.

Anupriya, Bansal, P., Graham, D.J., 2023. Congestion in cities: can road capacity expansions provide a solution? Transp. Res. Part A Policy Pract. 174, 103726.

Bai, L., Wong, S.C., Xu, P., Chow, A.H.F., Lam, W.H.K., 2021. Calibration of stochastic link-based fundamental diagram with explicit consideration of speed heterogeneity. Transp. Res. Part B Methodol. 150, 524–539.

Bai, L., Wong, W., Xu, P., Liu, P., Chow, A.H.F., Lam, W.H.K., et al., 2024. Fusion of multi-resolution data for estimating speed-density relationships. Transp. Res. Part C Emerg. Technol. 165, 104742.

Bender, C., Orszag, S., 1999. Advanced mathematical methods for scientists and engineers I: asymptotic methods and perturbation theory. https://doi.org/10.1007/978-1-4757-3069-2.

Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M., et al., 2012. Statistical mechanics for natural flocks of birds. Proc. Natl. Acad. Sci. 109, 4786–4791.

Bialek, W., Cavagna, A., Giardina, I., Mora, T., Pohl, O., Silvestri, E., et al., 2014. Social interactions dominate speed control in poising natural flocks near criticality. Proc. Natl. Acad. Sci. 111, 7212–7217.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning, first ed. Springer, New York, USA.

Boel, R., Mihaylova, L., 2006. A compositional stochastic model for real time freeway traffic simulation. Transp. Res. Part B Methodol. 40, 319–334.

Brackstone, M., McDonald, M., 1999. Car-following: a historical review. Transp. Res. Part F Psychol. Behav. 2, 181–196.

Bramich, D.M., Menendez, M., Ambuhl, L., 2022. Fitting empirical fundamental diagrams of road traffic: a comprehensive review and comparison of models using an extensive data set. IEEE Trans. Intell. Transport. Syst. 23, 14104–14127.

Bramich, D.M., Menéndez, M., Ambühl, L., 2023. FitFun: a modelling framework for successfully capturing the functional form and noise of observed traffic flow–density–speed relationships. Transp. Res. Part C: Emerg. Technol. 151, 104068.
Branston, D., 1976. Models of single lane time headway distributions. Transp. Sci. 10, 125–148.
Cassidy, M.J., 1998. Bivariate relations in nearly stationary highway traffic. Transp. Res. Part B Methodol. 32, 49–59.
Chen, X.M., Li, Z., Li, L., Shi, Q., 2014. Characterising scattering features in flow–density plots using a stochastic platoon model. Transp. A Transp. Sci. 10, 820–848.
Cheng, Q., Lin, Y., Zhou, X., Liu, Z., 2024. Analytical formulation for explaining the variations in traffic states: a fundamental diagram modeling perspective with stochastic parameters. Eur. J. Oper. Res. 312, 182–197.
Chowdhury, D., Santen, L., Schadschneider, A., 2000. Statistical physics of vehicular traffic and some related systems. Phys. Rep. 329, 199–329.
Chung, K.L., 1960. Markov Chains with Stationary Transition Probabilities. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-49686-8.
Coifman, B., 2015. Empirical flow-density and speed-spacing relationships: evidence of vehicle length dependency. Transp. Res. Part B Methodol. 78, 54–65.
Csiszar, I., 1991. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. Ann. Stat. 19, 2032–2066.
Daganzo, C.F., 1977. Fundamentals of Transportation and Traffic Operations. Emerald Group Publishing Limited, Huddersfield, USA.
Daganzo, C.F., 2002. A behavioral theory of multi-lane traffic flow. Part I: long homogeneous freeway sections. Transp. Res. Part B Methodol. 36, 131–158.
de Gier, J., Schadschneider, A., Schmidt, J., Schütz, G.M., 2019. Kardar-Parisi-Zhang universality of the Nagel-Schreckenberg model. Phys. Rev. E 100, 052111.
De Martino, A., De Martino, D., 2018. An introduction to the maximum entropy approach and its application to inference problems in biology. Heliyon 4, e00596.
de Souza, F., Stern, R., 2021. Calibrating microscopic car-following models for adaptive cruise control vehicles: multiobjective approach. J. Transp. Eng. Part A Syst. 147, 04020150.
Edie, L.C., 1961. Car-following and steady-state theory for noncongested traffic. Oper. Res. 9, 66–76.
Ermak, D.L., McCammon, J.A., 1978. Brownian dynamics with hydrodynamic interactions. J. Chem. Phys. 69, 1352–1360.
Foley, D.K., 1994. A statistical equilibrium theory of markets. J. Econ. Theory 62, 321–345.
Forbes, G., Gardner, T., McGee, H., Srinivasan, R., 2012. Methods and Practices for Setting Speed Limits: an Informational Report. Federal Highway Administration, United States. Office of Safety. https://safety.fhwa.dot.gov/speedmgt/ref_mats/fhwasa12004/fhwasa12004.pdf.
Golan, A., Harte, J., 2022. Information theory: a foundation for complexity science. Proc Natl Acad Sci 119, e2119089119.
Greenberg, H., 1959. An analysis of traffic flow. Oper. Res. 7, 79–85.
Greenshields, B., Bibbins, J., Miller, H., 1935. A Study of Traffic Capacity, vol. 14. Highway Research Board Proceedings, pp. 448–477.
Grendar, M., 2001. MiniMax entropy and maximum likelihood: complementarity of tasks, identity of solutions. AIP Conf. Proc. 49–60.
Hoogendoorn, S.P., Bovy, P.H.L., 1998. New estimation technique for vehicle-type-specific headway distributions. Transp. Res. Rec. 1646, 18–28.
Jabari, S.E., Liu, H.X., 2012. A stochastic model of traffic flow: theoretical foundations. Transp. Res. Part B Methodol. 46, 156–174.
Jabari, S.E., Zheng, J., Liu, H.X., 2014. A probabilistic stationary speed–density relation based on Newell's simplified car-following model. Transp. Res. Part B Methodol. 68, 205–223.
Jaynes, E.T., 1957. Information theory and statistical mechanics. Phys. Rev. 106, 620–630.
Jaynes, E.T., 1982. On the rationale of maximum-entropy methods. Proc. IEEE 70, 939–952.
Ji, Y., Daamen, W., Hoogendoorn, S., Hoogendoorn-Lanser, S., Qian, X., 2010. Investigating the shape of the macroscopic fundamental diagram using simulation data. Transp. Res. Rec. 2161, 40–48.
Kawasaki, K., 1973. Simple derivations of generalized linear and nonlinear Langevin equations. J. Phys. Math. Gen. 6, 1289–1295.
Kerner, B.S., 1998. Experimental features of self-organization in traffic flow. Phys. Rev. Lett. 81, 3797–3800.
Kerner, B.S., 2009. Introduction to Modern Traffic Flow Theory and Control: the Long Road to Three-phase Traffic Theory. Springer Science & Business Media, Berlin, Germany.
Kerner, B., Rehborn, H., 1996. Experimental properties of complexity in traffic flow. Phys. Rev. E 53, R4275.
Keyvan-Ekbatani, M., Kouvelas, A., Papamichail, I., Papageorgiou, M., 2012. Exploiting the fundamental diagram of urban networks for feedback-based gating. Transp. Res. Part B Methodol. 46, 1393–1403.
Krajewski, R., Bock, J., Kloeker, L., Eckstein, L., 2018. The highD dataset: a drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 2118–2125.
Lemons, D.S., Gythiel, A., 1997. Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146, 530–533 (1908)]. Am. J. Phys. 65, 1079–1081.
Lenz, H., Wagner, C.K., Sollacher, R., 1999. Multi-anticipative car-following model. Eur. Phys. J. B 7, 331–335.
Lezon, T.R., Banavar, J.R., Cieplak, M., Maritan, A., Fedoroff, N.V., 2006. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. Proc. Natl. Acad. Sci. 103, 19033–19038.
Li, L., Chen, X., 2017. Vehicle headway modeling and its inferences in macroscopic/microscopic traffic flow theory: a survey. Transp. Res. Part C Emerging Technol. 76, 170–188.
Li, Qi, Racine, J.Scott, 2011. Nonparametric Econometrics : Theory and Practice. Princeton University Press, Princeton, USA.
Li, J., Zhang, H.M., 2011. Fundamental diagram of traffic flow. Transp. Res. Rec. 2260, 50–59.
Liang, C.Y., Peng, H., 1999. Optimal adaptive cruise control with guaranteed string stability. Veh. Syst. Dyn. 32, 313–330.
Lighthill, M., Whitham, G., 1955. On kinematic waves II. A theory of traffic flow on long crowded roads. Series A. Math. Phys. Sci. 229, 317–345.
Liu, X., Xu, J., Li, M., Wei, L., Ru, H., 2019. General-logistic-based speed-density relationship model incorporating the effect of heavy vehicles. Math. Probl Eng. 2019, 6039846.
Lu, X.Y., Varaiya, P., Horowitz, R., 2009. Fundamental diagram modelling and analysis based NGSIM data. IFAC Proc 42, 367–374.
Mahnke, R., Kaupužs, J., 1999. Stochastic theory of freeway traffic. Phys. Rev. E 59, 117–125.
Milanes, V., Shladover, S.E., Spring, J., Nowakowski, C., Kawazoe, H., Nakamura, M., 2014. Cooperative adaptive cruise control in real traffic situations. IEEE Trans. Intell. Transport. Syst. 15, 296–305.
Montanino, M., Punzo, V., 2015. Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns. Transp. Res. Part B Methodol. 80, 82–106.
Nagel, K., Schreckenberg, M., 1992. A cellular automaton model for freeway traffic. J. Phys. I 2, 2221–2229.
Newell, G.F., 1961. Nonlinear effects in the dynamics of car following. Oper. Res. 9, 209–229.
Ngoduy, D., 2011. Multiclass first-order traffic model using stochastic fundamental diagrams. Transportmetrica 7, 111–125.
Ni, D., 2016. Traffic flow theory. Traffic Flow Theory: Characteristics, Experimental Methods, and Numerical Techniques. Elsevier, Amsterdam, the Netherlands.
Ni, D., 2021. Field theory for some traffic phenomena and fundamental diagram. Transp. Res. Rec. 2675, 1195–1208.
Ni, D., Hsieh, H.K., Jiang, T., 2018. Modeling phase diagrams as stochastic processes with application in vehicular traffic flow. Appl. Math. Model. 53, 106–117.
Nishinari, K., Fukui, M., Schadschneider, A., 2004. A stochastic cellular automaton model for traffic flow with multiple metastable states. J. Phys. Math. Gen. 37, 3101–3110.
Papageorgiou, M., Kotsialos, A., 2002. Freeway ramp metering: an overview. IEEE Trans. Intell. Transp. Syst. 3, 271–281.
Papageorgiou, M., Kiakaki, C., Dinopoulou, V., Kotsialos, A., Wang, Y., 2003. Review of road traffic control strategies. Proc. IEEE 91, 2043–2067.
Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Stat. 33, 1065–1076.
Payne, H.J., 1971. Models of freeway traffic and control. Math. Models Public Syst. 1, 51–61.
Pipes, L.A., 1967. Car following models and the fundamental diagram of road traffic. Transp. Res. 1, 21–29.
Qian, W.L., Siqueira, A.F., Machado, R.F., Lin, K., Grant, T.W., 2017a. Dynamical capacity drop in a nonlinear stochastic traffic model. Transp. Res. Part B Methodol. 105, 328–339.
Qian, W.L., Wang, B., Lin, K., Machado, R.F., Hama, Y., 2017b. A mesoscopic approach on stability and phase transition between different traffic flow states. Int. J. Non-Linear Mech. 89, 59–68.
Qu, X., Zhang, J., Wang, S., 2017. On the stochastic fundamental diagram for freeway traffic: model development, analytical properties, validation, and extensive applications. Transp. Res. Part B Methodol. 104, 256–271.
Ratnaparkhi, A., 2011. Maximum entropy models for natural language processing. In: Encyclopedia of Machine Learning, pp. 647–651.
Richards, P.I., 1956. Shock waves on the highway. Oper. Res. 4, 42–51.
Schlick, T., 2010. Molecular Modeling and Simulation: an Interdisciplinary Guide. Springer, New York, USA.
Schneidman, E., Berry, M. J. 2nd, Segev, R., Bialek, W., 2006. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440, 1007–1012.
Seo, T., Bayen, A.M., Kusakabe, T., Asakura, Y., 2017. Traffic state estimation on highway: a comprehensive survey. Annu. Rev. Control 43, 128–151.
Seo, T., Kawasaki, Y., Kusakabe, T., Asakura, Y., 2019. Fundamental diagram estimation by using trajectories of probe vehicles. Transp. Res. Part B Methodol. 122, 40–56.
Shore, J.E., Johnson, R.W., 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans. Inf. Theory 26, 26–37.
Siqueira, A.F., Peixoto, C.J.T., Wu, C., Qian, W.L., 2016. Effect of stochastic transition in the fundamental diagram of traffic flow. Transp. Res. Part B Methodol. 87, 1–13.

Sopasakis, A., Katsoulakis, M.A., 2006. Stochastic modeling and simulation of traffic flow: asymmetric single exclusion process with Arrhenius look-ahead dynamics. SIAM J. Appl. Math. 66, 921–944.
Sun, L., Zhou, J., 2005. Development of multiregime speed–density relationships by cluster analysis. Transp. Res. Rec. 1934, 64–71.
Sun, J., Zheng, Z., Sun, J., 2018. Stability analysis methods and their applicability to car-following models in conventional and connected environments. Transp. Res. Part B Methodol. 109, 212–237.
Sutton, R.S., Barto, A.G., 1998. Reinforcement learning: an introduction. IEEE Trans. Neural Netw. 9, 1054.
Treiber, M., Helbing, D., 1999. Macroscopic simulation of widely scattered synchronized traffic states. J. Phys. Math. Gen. 32, L17–L23.
Treiber, M., Helbing, D., 2003. Memory effects in microscopic traffic models and wide scattering in flow-density data. Phys. Rev. E 68, 046119.
Treiber, M., Kesting, A., 2013. Traffic flow dynamics. Traffic Flow Dynamics: Data, Models and Simulation. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32460-4.
Treiber, M., Kesting, A., Helbing, D., 2006. Understanding widely scattered traffic flows, the capacity drop, and platoons as effects of variance-driven time gaps. Phys. Rev. E 74, 016123.
Venkata Ramana, A.S., Jabari, S.E., 2020. Traffic flow with multiple quenched disorders. Phys. Rev. E 101, 052127.
Wagner, P., Nagel, K., Wolf, D.E., 1997. Realistic multi-lane traffic rules for cellular automata. Phys. A 234, 687–698.
Wainwright, M.J., Jordan, M.I., 2007. Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. 1, 1–305.
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.
Wang, S., Chen, X., Qu, X., 2021. Model on empirically calibrating stochastic traffic flow fundamental diagram. Commun. Transp. Res. 1, 100015.
Whitham, G.B., 1999. Linear and Nonlinear Waves. John Wiley & Sons, Hoboken, USA.
Wu, X., Liu, H.X., 2013. The uncertainty of drivers' gap selection and its impact on the fundamental diagram. Procedia. Soc. Behav. Sci. 80, 901–921.
Wu, X., Liu, H.X., Geroliminis, N., 2011. An empirical analysis on the arterial fundamental diagram. Transp. Res. Part B Methodol. 45, 255–266.
Zhang, J., Qu, X., Wang, S., 2018. Reproducible generation of experimental data sample for calibrating traffic flow fundamental diagram. Transp. Res. Part A Policy Pract. 111, 41–52.
Zhu, L., Yu, F.R., Wang, Y., Ning, B., Tang, T., 2019. Big data analytics in intelligent transportation systems: a survey. IEEE Trans. Intell. Transp. Syst. 20, 383–398.

Xiaohui Zhang received her B.S. and M.S. degrees in transportation engineering from Tongji University in 2020. She is currently pursuing her Ph.D. degree at Tongji University. Her main research interests include traffic flow theory and modeling and emerging technology for intelligent traffic system integrated machine learning.

Jie Sun received his Ph.D. degree in transportation engineering from Tongji University in 2019. He is currently working as an Associate Professor at Tongji University. He worked as a Postdoctoral Research Fellow at the University of Queensland and The Hong Kong University of Science and Technology, respectively. His main research interests include traffic flow theory and modeling, traffic simulation, and connected and automated vehicles.

Jian Sun received his Ph.D. degree from Tongji University in 2006. Subsequently, he was at Tongji University as a Lecturer and then promoted to the position of Professor in 2011, where he is currently a Professor with the College of Transportation Engineering and the Dean of the Department of Traffic Engineering. His main research interests include traffic flow theory, traffic simulation, connected and automated vehicles, and intelligent transportation systems.