Transportation Research Part C 110 (2020) 1–19

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Hierarchical ramp metering in freeways: An aggregated modeling and control approach

Yu Han[a,*], Mohsen Ramezani[c], Andreas Hegyi[b], Yufei Yuan[b], Serge Hoogendoorn[b]

[a] School of Transportation, Southeast University, Nanjing, China
[b] Department of Transport and Planning, Delft University of Technology, the Netherlands
[c] The University of Sydney, School of Civil Engineering, Sydney, Australia

ARTICLE INFO

Keywords:
Network fundamental diagram (NFD)
Heterogeneity
Capacity drop
Parsimonious traffic modeling
Highway congestion management

ABSTRACT

This paper develops a model-based hierarchical control method for coordinated ramp metering on freeway networks with multiple bottlenecks and on- and off-ramps. The controller consists of two levels where at the upper level, a Model Predictive Control (MPC) approach is developed to optimize total network travel time by manipulating total inflow from on-ramps to the freeway network. The lower level controller distributes the optimal total inflows to each on-ramp of the freeway based on local traffic state feedback. The control method is based on a parsimonious aggregated traffic model that relates the freeway total outflow to the number of vehicles on the freeway sections.

Studies on aggregated traffic modeling of networks have shown the existence of a well-defined and low-scatter Macroscopic Fundamental Diagram (MFD) for urban networks. The MFD links network aggregated flow and density (accumulation). However, the MFD of freeway networks typically exhibits high scatter and hysteresis loops that challenge the control performance of MFD-based controllers for freeways. This paper addresses these challenges by modelling the effect of density heterogeneity along the freeway and capacity drop on characteristics of freeway MFD using field traffic data. In addition, we introduce a model to predict the evolution of density heterogeneity that is essential to reproduce the dynamics of freeway MFD accurately. The proposed model is integrated as the prediction model of the MPC in the hierarchical control method.

The proposed coordinated ramp metering method shows desirable performance to reduce the vehicles total time spent and eliminate congestion. The control approach is compared with other coordinated ramp metering controllers based on the MPC framework with different traffic prediction models (e.g. CTM and METANET). The outcomes of numerical experiments highlight that the MFD-based hierarchical controller (i) is better able to overcome the modeling mismatch between the prediction model and the plant (process model) in the MPC framework and (ii) requires less computation effort than other nonlinear controllers.

1. Introduction

Ramp metering control is a common operation measure in freeways to regulate flows from on-ramps. There is an extensive literature on ramp metering control. Early works attempted to alleviate local traffic problems using local ramp metering (Papageorgiou et al., 1991). It has been identified that local ramp metering approaches are not always efficient to ameliorate global

* Corresponding author.
E-mail address: yuhan@seu.edu.cn (Y. Han).

https://doi.org/10.1016/j.trc.2019.09.023
Received 1 February 2019; Received in revised form 26 September 2019; Accepted 30 September 2019

traffic conditions of the overall traffic network (PPapageorgiou and Kotsialosageorgiou and Kotsialos, 2000). Since then, various advanced coordinated ramp metering methods have been developed (e.g. (Chow, 2015)), and a noticeable method among them is the Model Predictive Control (MPC) approach that is based on traffic flow models. MPC framework contains a prediction model that predicts the evolution of traffic dynamics and estimates the optimal control scheme for the time period in which the relevant traffic dynamics occur. This feature enables the controller to take advantage of potentially larger future gains at a current (smaller) cost, so as to avoid myopic control actions. Studies of Kotsialos and Papageorgiou (2004), Hegyi et al. (2005a), Papamichail et al. (2010) presented different optimal approaches of ramp metering that were based on the METANET model. Gomes and Horowitz (2006) presented an optimal freeway ramp metering approach that was based on the asymmetric cell transmission model, which can be efficiently solved by linear programming. Han et al. (2015) presented a linear quadratic MPC approach for ramp metering based on an extended cell transmission model (CTM) which took the capacity drop into account. For a detailed description of METANET and CTM, readers are referred to Kotsialos and Papageorgiou (2004), Daganzo (1994). The aforementioned studies used same models as the prediction model and the process model which represents the reality. However in field applications, there is inherently significant mismatch between the dynamics of the prediction model and the process (plant). Hence, developing a model-based traffic control method that demonstrates robustness to modeling mismatch in the MPC framework will increase the confidence that the proposed approach will also accurately work in practice, where there is always model mismatch. This is a challenging task and is not sufficiently studied in the literature.

This paper seeks potentials in parsimonious macroscopic traffic models for ramp metering to be incorporated in the MPC approach as the prediction model to address the modeling mismatch difficulty. Among them, the Macroscopic Fundamental Diagram (MFD) enable us to develop a hierarchical model-based control method. The MFD provides an efficient tool for expressing aggregated dynamics of traffic networks that links the accumulation (weighted sum of links densities and lengths) and the production (weighted sum of links flows and lengths) of the network (Daganzo, 2007; Geroliminis and Daganzo, 2008). Although a unimodal and low-scatter MFD was observed in a homogeneous region of a city (i.e. with low spatial density heterogeneity), existence of a well-defined MFD in general cases specifically in freeway networks is still an open question, which might undermine the benefits of MFD in modeling traffic dynamics and traffic control applications (Daganzo et al., 2011; Buisson and Ladier, 2009). For example, studies of Mazloumian et al. (2010), Geroliminis and Sun (2011a), Gayah and Daganzo (2011) have shown that urban networks with heterogeneous distribution of density exhibit network flows smaller than those that approximately meet homogeneous conditions. Thus, recent studies focus more on modeling and controlling dynamics of heterogeneity in urban networks (Ramezani et al., 2015; Yildirimoglu et al., 2015). Nevertheless, the MFD-based controllers require less detailed information of traffic states compared to microscopic, mesoscopic, and non-parsimonious macroscopic model-based traffic controllers. The MFD substantially reduces the complexity of traffic models because the packets of vehicular traffic are considered as a single continuum entity (Laval et al., 2017). Furthermore, MFD-based controllers need less computation effort that is beneficial for real-time applications.

The MFD has been utilized to introduce city-scale traffic control strategies, e.g. Perimeter control (Geroliminis et al., 2013), to decrease delays in large-scale urban networks. In the literature, different control approaches have been used to solve the perimeter control problems. For example, classical feedback control approaches have been implemented by Keyvan-Ekbatani et al. (2012, 2015), Kouvelas et al. (2017), Ampountolas et al. (2017), and the model predictive control (MPC) approach has been used to solve the optimal control problems in (Geroliminis et al., 2013; Haddad et al., 2013; Ramezani et al., 2015; Yang et al., 2017). Moreover, methods of MFD estimation are studied in (Leclercq et al., 2014; Laval and Castrillón, 2015; Saberi et al., 2014). Recently, Haddad (2015) designed a robust perimeter controller to systematically take into account uncertainties in MFD-based controllers; Ramezani and Nourinejad (2017) developed a city-scale dispatching of ride-sourcing (e.g. taxi) systems based on the MFD dynamics.

Although MFD-based controllers have been investigated in urban networks, they have not been scrutinized for freeway traffic control. It has been found that a well-defined and low-scatter MFD does not always exist for freeways (Daganzo, 2011). Cassidy et al. (2011) showed that a well-defined and low-scatter freeway MFD can be observed only when traffic is in stable regime (either congested or non-congested) in all lanes and on all links. Studies of Ji et al. (2010), Saberi and Mahmassani (2012), Geroliminis and Sun (2011b) have shown that freeway MFD exhibits hysteresis pattern during the onset and offset of congestion, which represents an adverse effect on the freeway overall performance. In these studies, the heterogeneity of density distribution was perceived as a main factor that causes the hysteresis pattern. Knoop and Hoogendoorn (2013) defined the spatial spread of density as a function of the weighted variance of the densities in all sections, and fitted a third-order polynomial functional form to field traffic data obtained from a Dutch freeway. It was found that by including the density heterogeneity, the accuracy of the MFD improves significantly. In Geroliminis and Sun (2011b), it was found that non-equilibrium states in individual detectors' measurements in freeways, which include transition flows and the capacity drop, also influence the production of the MFD. The traffic patterns in freeway networks have distinct characteristics, e.g., route choice and congestion propagation, that challenge the traffic congestion modeling and management. Thus, understanding and modeling the dynamics of density heterogeneity is crucial for developing MFD-based freeway traffic control schemes. This paper presents a systematic analysis to quantify the effect of determining traffic state factors on MFD and offers a tractable coordinated freeway congestion control method.

The objectives of this paper are twofold related to modeling and control aspects. Regarding the modeling aspect, we introduce a model to capture the evolution of density heterogeneity which is essential to reproduce the dynamics of freeway MFD. Note that the density heterogeneity in this paper refers to the longitudinal heterogeneity along the freeway rather than the lateral heterogeneity across lanes. By analyzing field data, we model the density heterogeneity as (i) a linear function of the average density when no congestion occurs in the freeways and (ii) a non-linear function when congestion occurs. The presented model captures this phenomenon and *predicts* the density heterogeneity. Furthermore, we present a quantitative measure to represent the effect of capacity drop on the MFD, which is fitted to and tested with field data. Regarding the control aspect, we present a hierarchical control

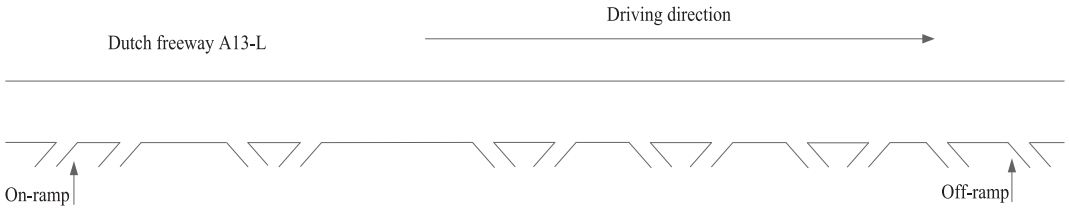

Fig. 1. The sketch of Dutch freeway A13-L. The freeway stretch is about 16 km in length and includes 6 on-ramps and 6 off-ramps.

approach for coordinated freeway ramp metering. At the upper level, an MPC approach is developed to optimize total inflows from on-ramps to the freeway stretch, where the prediction model is formulated based on the proposed MFD model. At the lower level, the controller distributes the optimal total inflow to each on-ramp of the freeway based on local traffic state feedback. The presented controller is tested with different models (e.g. CTM and METANET) as the process model and compared with other optimal control approaches. The results demonstrate that the MFD-based controller achieves desirable performance to reduce vehicles total time spent. It is of particular significance that the simulation results show that the proposed controller performs better with an inherent model mismatch.

The remaining of the paper is set up as follows. The modeling of the freeway MFD is presented in Section 2. In Section 2.1, we explore the relation between network production and network outflow through field data. Sections 2.2 and 2.3 present the density heterogeneity model and the capacity drop model respectively. The accuracy of the MFD is tested with field data in Section 2.4. Section 3 describes the control design, where the upper level MPC controller and the lower level controller are presented. In Section 4, numerical experiments to test the performance of the proposed controller are presented. A summary and the discussion of future work conclude the paper in Section 5.

2. Modeling of the freeway MFD

The MFD considered in this paper represents the aggregated traffic behavior of a freeway stretch, which has one mainstream origin and destination, and several on-ramps and off-ramps with corresponding origins and destinations demands. The field data of the Dutch freeway A13-L are used for empirical analysis. As shown in Fig. 1, the freeway stretch is about 16 km in length including 6 on-ramps and 6 off-ramps. 38 detector stations are placed at the freeway mainstream, while the distances between the detector range from 300 m to 600 m. The data consist of detected flows and time mean speeds, respectively denoted as q_l (veh/h) and v_l (km/h), of lane l at each detector station during one minute, from March and April, 2013. The data cover all weekdays in the two months, from 6:00 to 24:00. Bottlenecks include recurrent bottlenecks at merge locations (originated from on-ramps) and non-recurrent bottlenecks such as off-ramps blockade and moving jams.

The freeway stretch is divided into segments, where we consider that the middle of two neighboring detectors is the boundary of a segment. There are multiple lanes in a freeway segment, and the number of lanes in segment i is represented as λ_i. The detectors measure vehicle counts and the average speed of vehicles crossing the loop for every lane. Speeds over lanes in each segment i are harmonically averaged to approximate the space mean speed v_i,

$$v_i = \left(\frac{\sum_l q_l \frac{1}{v_l}}{\sum_l q_l} \right)^{-1}. \tag{1}$$

The detected flows over lanes in each segment are also aggregated, and the aggregated flow of segment i is denoted as q_i. q_i and v_i are aggregated for every 5 min. Then, the average density of each segment, i.e. ρ_i, is estimated as the ratio of q_i and v_i based on the fundamental relation. The MFD links the weighted averaged flow, $\bar{q}$, and the weighted average density, $\bar{\rho}$, which are estimated as,

$$\bar{q} = \frac{\sum_{i=1}^{I} q_i l_i}{\sum_{i=1}^{I} l_i}, \tag{2}$$

$$\bar{\rho} = \frac{\sum_{i=1}^{I} \rho_i l_i}{\sum_{i=1}^{I} l_i}, \tag{3}$$

where l_i is the length of segment i and I is the number of segments in the freeway stretch. Note that the denominators of (2) and (3) are the total length of the freeway stretch, and the numerators of (2) and (3) are respectively the so-called production (represented by P) and accumulation (represented by A) of the freeway stretch. The MFD links P and A, or interchangeably $\bar{q}$ and $\bar{\rho}$.

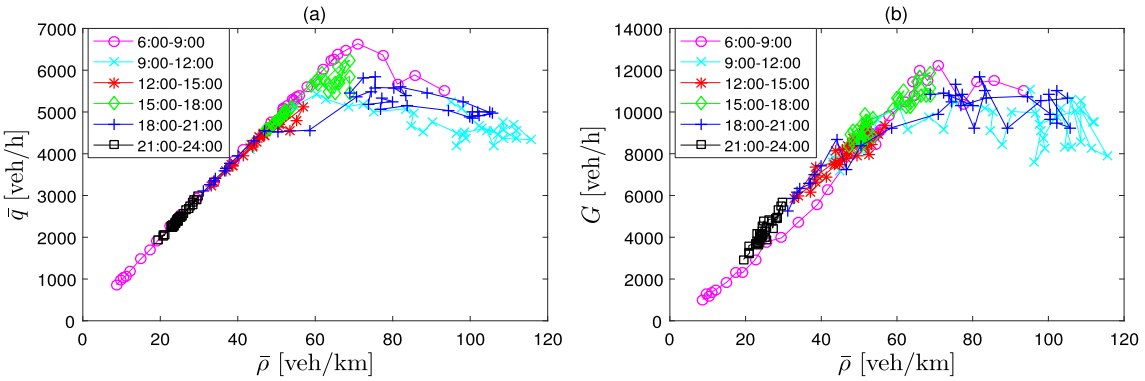

Fig. 2. Aggregated loop detector data of April 25, 2013: (a) weighted average flow versus weighted average density; (b) total outflow versus weighted average density. The shapes in (a) and (b) are unimodal, with less scatter in the left part and high scatter in the right part. Note that in (b), some data points from 6:00 to 9:00 have lower values of the total outflow compared to the traffic condition from 21:00 to 24:00 due to longer average trip lengths, see Fig. 3.

Fig. 2(a) shows the freeway MFD which is obtained from loop detector data on April 25, 2013. It can been seen that the freeway MFD is unimodal, with less scatter in the left part and high scatter in the right part. The left part of the MFD represents the aggregated traffic states of the freeway when no congestion occurs, which often happens during off-peak hours. The right part of the MFD represents the aggregated traffic states of the freeway when congestion occurs, which often happens during peak hours.

For traffic control purposes on freeways, we are interested in the total outflow (network exit function) of the freeway stretch (represented by G, see Fig. 2(b)) which is essential to predict traffic dynamics through mass (the number of vehicles) conservation equations. The total outflow of the freeway stretch includes the flows that leave the freeway stretch at the downstream end and the off-ramps. The relation between the production and the total outflow of the freeway stretch is investigated in the next section.

2.1. Relationship between network production and outflow

The analysis of field data of Yokohama in Geroliminis and Daganzo (2008) shows that the ratio of the production and the trip completion rate (network outflow) is relatively constant over time. Note that the ratio is time-varying intrinsically however, assuming a time-invariant ratio between the network production and trip completion rate has been shown to be useful for modeling and control purposes. The ratio can be perceived as the vehicular average trip length (represented by $\bar{x}$) in the network (Leclercq et al., 2015; Batista et al., 2019). This relationship can be readily adapted to the freeway system. For example, suppose a pair of on-ramp and off-ramp are close to each other and are located in the middle of a freeway stretch. We consider two situations that occur in the freeway stretch: Situation 1, no traffic is using either the off-ramp or the on-ramp, and all the vehicles are going out of the freeway stretch at the downstream end; Situation 2, all of the vehicles at the upstream of the off-ramp are going out the freeway stretch through the off-ramp and the same amount of vehicles are entering the freeway stretch from the on-ramp and going out at the downstream end. For these two situations, the following relationships hold: $A_1 \approx A_2$, $P_1 \approx P_2$, $\bar{x}_1 \approx 2\bar{x}_2$, and $G_1 = \frac{1}{2}G_2$. For the same accumulation, the productions of the two situations are the same, but the total outflow of situation 2 is two times higher than that of situation 1 while the average trip length in situation 2 is half of that in situation 1. Thus, the total outflow is not only a function of the accumulation, but also relevant to the average trip length.

Data from the Dutch freeway A13-L are explored to investigate the relation between network production and outflow. Since there is no detector placed at off-ramps, off-ramp flows are estimated as the flow difference of detectors directly upstream and downstream of the off-ramps. The ratio between P and G is shown as a box plot in Fig. 3, where each box represents data of a time slot of 5 min that are obtained from all weekdays in March and April, 2013. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers in each time slot extend to the most extreme data point.

Fig. 3 depicts the temporal variation of the average trip length. Traffic volumes at late evening and early morning are low and less travelers exit the freeway stretch through off-ramps, which result in longer average trip lengths. At peak hours (around 9:00 and 18:00) the average trip lengths are slightly shorter because some travelers might exit the freeway stretch through off-ramps once congestion occurs downstream. If the average trip length is constant over time, then it is physically equivalent to use G and P as the performance indicator in the MFD. The MFD which links the total outflow and the weighted average density is shown in Fig. 2(b). It can be seen that the scatter of two MFDs have similar features.

Although the average trip length is relatively and inherently time-variant (in the range of 8.5 ± 1 km during 7:30 to 21:00), to derive the model to capture the evolution of vehicles accumulation on the freeway and the on-ramps and develop a tractable control method, we assume that the average trip length is time-invariant. Note that, in all the numerical case studies in Section 4, the process models (plants) have time-varying trip length that is a function of time-varying O-D. Yet, the MFD-based controller by assuming a reasonable time-invariant trip length demonstrates that this assumption is not limiting and the proposed model is applicable for the control purposes. Note that if the average trip length is significantly time-variant, to accurately predict the total outflow of the

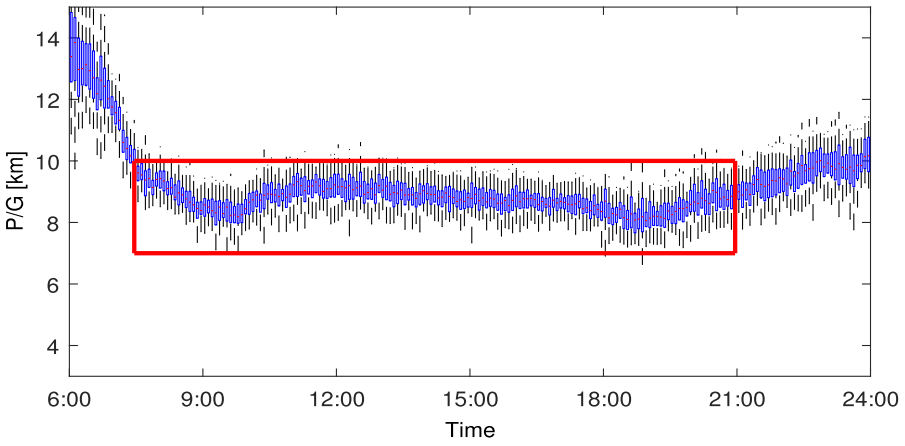

Fig. 3. The whisker plot of P/G aggregated for all of the weekdays during March and April 2013. On each box, the central red mark indicates the median, and the bottom and top blue edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points. In the rectangle area, the values of P/G are relatively time-invariant, which indicates that the average trip lengths are nearly time-invariant. The average trip lengths from 6:00 to 7:30 are longer, which results in lower values of total outflows, see Fig. 2(b). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

freeway stretch, the dynamics of the average trip length also needs to be predicted. The average trip length can be predicted based on dynamic OD matrix, and this is a future research direction. A general assumption for the rest of this paper is that the average trip length is time-invariant.

2.2. Modeling the density heterogeneity

A major factor of MFD scatter, which has been identified in the literature (e.g. Mazloumian et al. (2010)), is the spatial density heterogeneity. In this paper, we use the standard deviation of density per lane to represent the density heterogeneity, σ. This is formulated as,

$$\sigma(t) = \sqrt{\frac{1}{I}\sum_{i=1}^{I}\left(\frac{\rho_i(t)}{\lambda_i} - \frac{\sum_{i=1}^{I}\frac{\rho_i(t)}{\lambda_i}}{I}\right)^2}\ (\frac{\text{veh}}{\text{km}\cdot\text{lane}}), \tag{4}$$

where λ_i is the number of lanes in segment i. Empirical studies of freeway networks show that for the same values of accumulations, higher density heterogeneity results in lower network production (Geroliminis and Sun, 2011b; Knoop and Hoogendoorn, 2013).

To demonstrate the significance of the density heterogeneity in MFD modeling, Fig. 4 shows the aggregated loop detector data of

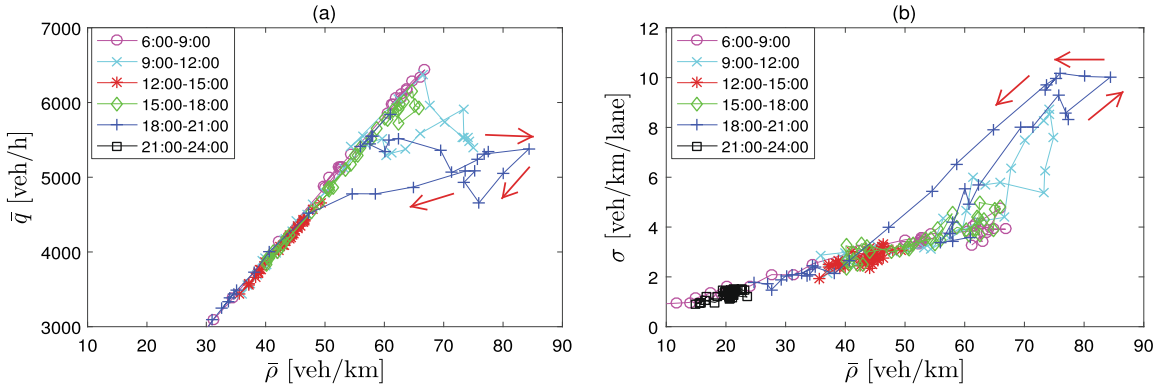

Fig. 4. Aggregated loop detector data of April 16, 2013: (a) weighted average flow versus weighted average density; (b) density heterogeneity versus weighted average density. Red arrows in the figures represent the direction of time in the hysteresis loop. Note that all the hysteresis patterns within the day and during the two months are considered in the modeling. While only the PM period is shown here with red arrows to keep the figure graphically appealing. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

April 16, 2013, in which (a) represents the weighted average flow as a function of the average density and (b) represents the heterogeneity as a function of the average density. In the figure, red arrows represent the direction of time in the hysteresis loops. In the density-heterogeneity diagram, the hysteresis loop is in a counter-clockwise shape, which means that density distribution is more heterogeneous in congestion unloading phase than congestion loading phase. Besides the density heterogeneity, the shape of MFD is also related to the extents of traffic load. For example, the maximum average density in Fig. 2(a) is as high as 120 veh/km while in Fig. 4(a) it is below 90 veh/km.

Fig. 4(a) and (b) are perceived as the combination of two parts, (i) data points on a straight line and (ii) a more complex pattern of hysteresis loops. For the same value of the average density, the points on the straight line have lower values of density heterogeneity and higher values of average flow than other points. The points can be divided into two groups, where one is from the morning peak (within the time period 9:00–12:00) and the other is from the evening peak (within the time period from 18:00–21:00). In this example, generally, data from the morning peak have lower values of the density heterogeneity and higher values of average flow than data from the evening peak (for same values of the average density). The points from the evening peak exhibit hysteresis pattern, which in Fig. 4(a) is in a clockwise shape and in Fig. 4(b) is in a counter-clockwise shape. All these observations indicate that higher density heterogeneity results in lower network production.

Ramezani et al. (2015) developed a traffic congestion control strategy to reduce the heterogeneity in urban networks. They assumed that a well-defined relationship between the average occupancy and the variance of occupancy exists in a homogeneous sub-region of the urban network. However in a heterogeneous freeway network, the density heterogeneity and the average density are strongly correlated with a complex pattern, see Fig. 4(b). Hence, characteristics of the dynamics of density heterogeneity, which may have a significant influence to the control performance, are needed to be further investigated and properly modeled.

The relation between σ and $\bar{\rho}$ can be perceived as the combination of two parts represented by a linear function and a non-linear function. If no congestion occurs in the freeway, the data points fall into the linear function which is represented as,

$$\sigma(t) = s_1 \cdot \bar{\rho}(t), \quad \text{if} \quad \forall\, i \in [1, 2, \ldots I],\ \rho_i(t) \leqslant \rho_i^{\mathrm{cr}} \tag{5}$$

where i is the index of segments of the freeway, ρ_i^{cr} is the critical density of Fundamental Diagram (FD) of segment i, and s_1 is the positive model parameter. It is assumed that if the density of every segment in the freeway is lower than the critical density, then no congestion occurs in the freeway. Thus if no congestion occurs in the freeway, the density heterogeneity increases linearly with the freeway average density. Once congestion occurs, the density heterogeneity evolves in a non-linear way which can be represented by high-order polynomial functions.

The value of the average density when congestion occurs in the freeway is not always the same because it depends on different traffic situations, for example, the activation of different bottlenecks, the blockade of different off-ramps, and the occurrence of moving jams under different traffic loadings. Thus, the threshold of the model switching to the non-linear function is not constant. If no congestion is detected, the model is represented by the linear function (5) and once congestion is detected, the model switches to the non-linear function (8). Note that depending on different traffic situations, the scatter points of the density heterogeneity are widely dispersed in σ-$\bar{\rho}$ plane. Thus, we propose a non-linear function that includes a time varying parameter, $\bar{\rho}^{\mathrm{cr}}(t)$, to represent the maximum average density that can be represented by the linear function under the traffic situation at time t. $\bar{\rho}^{\mathrm{cr}}(t)$ is formulated as,

$$\bar{\rho}^{\mathrm{cr}}(t) = \frac{1}{L} \sum_{i=1}^{I} \widetilde{\rho}_i(t) l_i, \tag{6}$$

where L is the total length of the freeway stretch, and

$$\widetilde{\rho}_i(t) = \begin{cases} \rho_i(t) & \text{if } \rho_i(t) \leqslant \rho_i^{\mathrm{cr}} \\ \rho_i^{\mathrm{cr}} & \text{if } \rho_i(t) > \rho_i^{\mathrm{cr}}. \end{cases} \tag{7}$$

Ultimately, the non-linear function is formulated as,

$$\sigma(t) = s_2 \cdot (\bar{\rho}(t) - \bar{\rho}^{\mathrm{cr}}(t))^3 + s_3 \cdot (\bar{\rho}(t) - \bar{\rho}^{\mathrm{cr}}(t))^2 + s_4 \cdot (\bar{\rho}(t) - \bar{\rho}^{\mathrm{cr}}(t)) + s_5 \quad \text{if} \quad \exists\, i \in [1, 2, \ldots I],\ \rho_i(t) > \rho_i^{\mathrm{cr}} \tag{8}$$

where s_2, s_3, s_4, and s_5 are model parameters.

The proposed heterogeneity model is thoroughly investigated in the Section 2.4 with different training and test datasets. For graphical representation of estimation and prediction of the model, we test the data from April 16, 2013. The time varying parameter $\bar{\rho}^{\mathrm{cr}}(t)$ is estimated at every time instance based on (6). The parameters s_1 to s_5 estimation is formulated as a non-linear optimization problem that aims to minimize the root mean square error (RMSE) between the model estimations and real training data. The estimation results are shown in Fig. 5(a), which corresponds to the data points from 18:00 to 21:00 in the counter-clockwise hysteresis loop in Fig. 4(b). The comparison between ground truth and model estimation is shown in Fig. 5(a). It can be seen that the model well reproduces the scatters and hysteresis of the density heterogeneity. Fig. 5(a) also shows higher discrepancies between the real data and model estimation around density range 70–80 [veh/km] that is associated with the critical density of production and outflow MFDs, see Fig. 2. The reason is that the extent of heterogeneity in the uncongested regime (the increasing part of MFD where the density is lower than the critical density) is very low and easier to estimate. The extent of heterogeneity is more difficult to estimate when the average density is round the critical density as traffic situation in the network is more complex.

To use the model for *prediction* in the MPC approach, $\bar{\rho}^{\mathrm{cr}}$ is assumed to be constant during the prediction horizon. The reason is that the dynamics of $\bar{\rho}^{\mathrm{cr}}$ are difficult to capture since the state of individual segments, ρ_i, is not forecast by the MFD. The accuracy of the prediction with this assumption is tested and the results are presented in Fig. 5(b), where blue lines represent the evolution of σ of

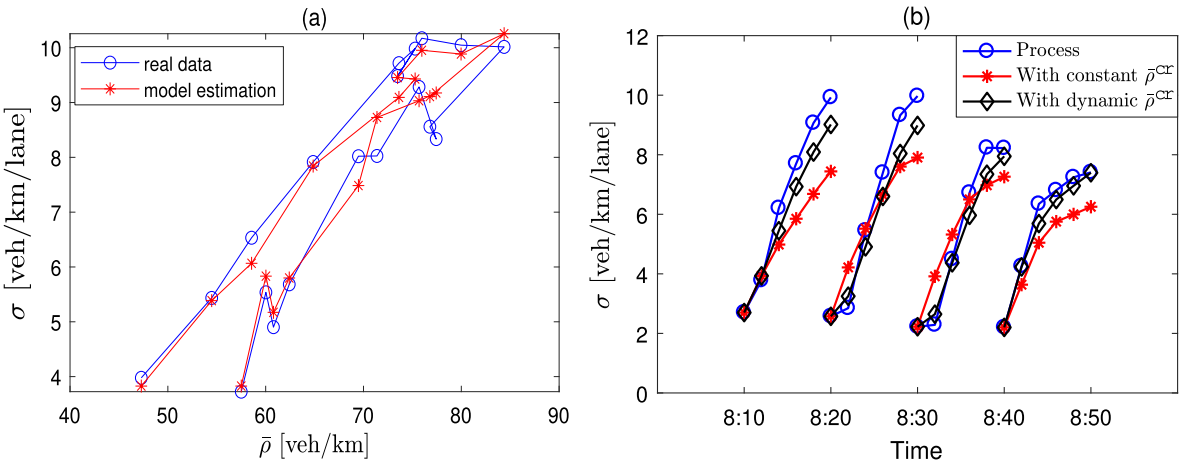

Fig. 5. (a) The estimation results of the density heterogeneity model. The red points are the estimated density heterogeneity at each time instance. The blue poins are the ground truth. Data are obtained from April 16, 2013. (b) The accuracy of the prediction of σ. Blue lines represents the evolution of σ in the process. Black lines represent the value of σ estimated by (8) which has a dynamic value of $\bar{\rho}^{cr}$. Red lines represent the values of σ predicted by the prediction model in the MPC approach, which assumes a constant value of $\bar{\rho}^{cr}$ in the prediction horizon. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the process model (plant, i.e. extended CTM) at certain time instances. In the prediction model of the MPC, it is assumed that $\bar{\rho}^{cr}$ is constant over the prediction horizon, where the predictions of σ are shown with the red lines. Moreover, black lines represent the values of σ predicted by (8), in which the values of $\sigma(t)$ are calculated based on dynamic values of $\bar{\rho}^{cr}$. Evidently, the model with dynamic $\bar{\rho}^{cr}$ models density heterogeneity, i.e. σ, more accurately, while the prediction model of the controller also captures the trend of increasing σ accurately.

2.3. The effect of capacity drop on freeway MFD

In addition to the density heterogeneity, it has been found that the freeway MFD is also affected by the non-equilibrium states in individual detectors' measurements, which include transition flows and capacity drop (Geroliminis and Sun, 2011b; Daganzo, 2011; Cassidy et al., 2011). In this Section, we explore the effect of the capacity drop on the weighted average flow of a freeway stretch through a synthetic case. The simulation model of the synthetic case is an extended cell transmission model (CTM) that is capable of reproducing the capacity drop and propagation of jam waves (Han et al., 2016). The extended CTM reproduces different extent of the capacity drop by assuming that the outflow of a jam is lower if the density of the jam head is higher. A simulation test is conducted on a synthetic 3-lane freeway stretch which is 16 km in length and divided into 40 cells. Four on-ramps are placed on the freeway stretch, which are potential bottlenecks. The layout of the synthetic freeway stretch is shown in Fig. 6.

To demonstrate the influence of capacity drop on freeway MFD, we consider two scenarios where $\bar{\rho}$ and σ are equal or similar as much as possible, but with different extents of the capacity drop. According to the mechanism of the model that the extent of capacity drop depends on the densities of the jam heads, we create two jams with different jam heads densities, which are shown in Fig. 7(a) and (b) respectively. In scenario 1, the congestion is triggered at on-ramp 2 and the density at the head of the congestion is about 35 [veh/km/lane]. In scenario 2, the congestion is triggered at on-ramp 3, and the density at the head of the congestion is about 50 [veh/km/lane]. The relationship between σ and $\bar{\rho}$, and $\bar{q}$ and $\bar{\rho}$ of the two scenarios are plotted in Fig. 7(c) and (d) respectively. It can be seen that when the average density is around 80 [veh/km], two scenarios have equal density heterogeneity. However, the average flow of scenario 2 is significantly lower (5.5%) than the one of scenario 1, as a result of higher capacity drop.

It is not trivial to quantify the influence of capacity drop on the average flow of MFD. Empirical studies suggested that there is a negative relation between the speed in the jam and the capacity drop (Yuan et al., 2015). It was found that if the speed in the jam is lower, the queue discharge rate is also lower. Thus, the speed in the jam is considered as a factor to quantify the effect of the capacity

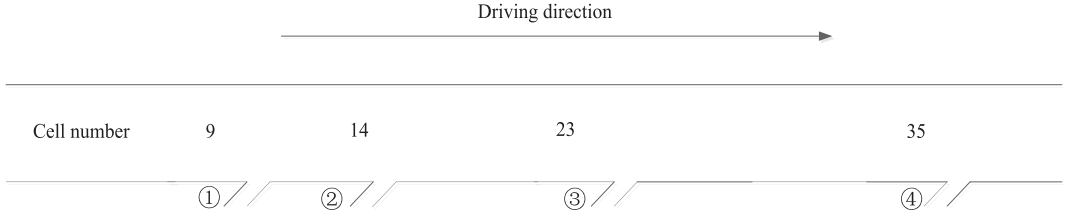

Fig. 6. The layout of the synthetic freeway stretch that is used to model the effect of capacity drop on MFD. Cell number are marked at the upstream of each on-ramp. Circled numbers represent the order of the on-ramps.

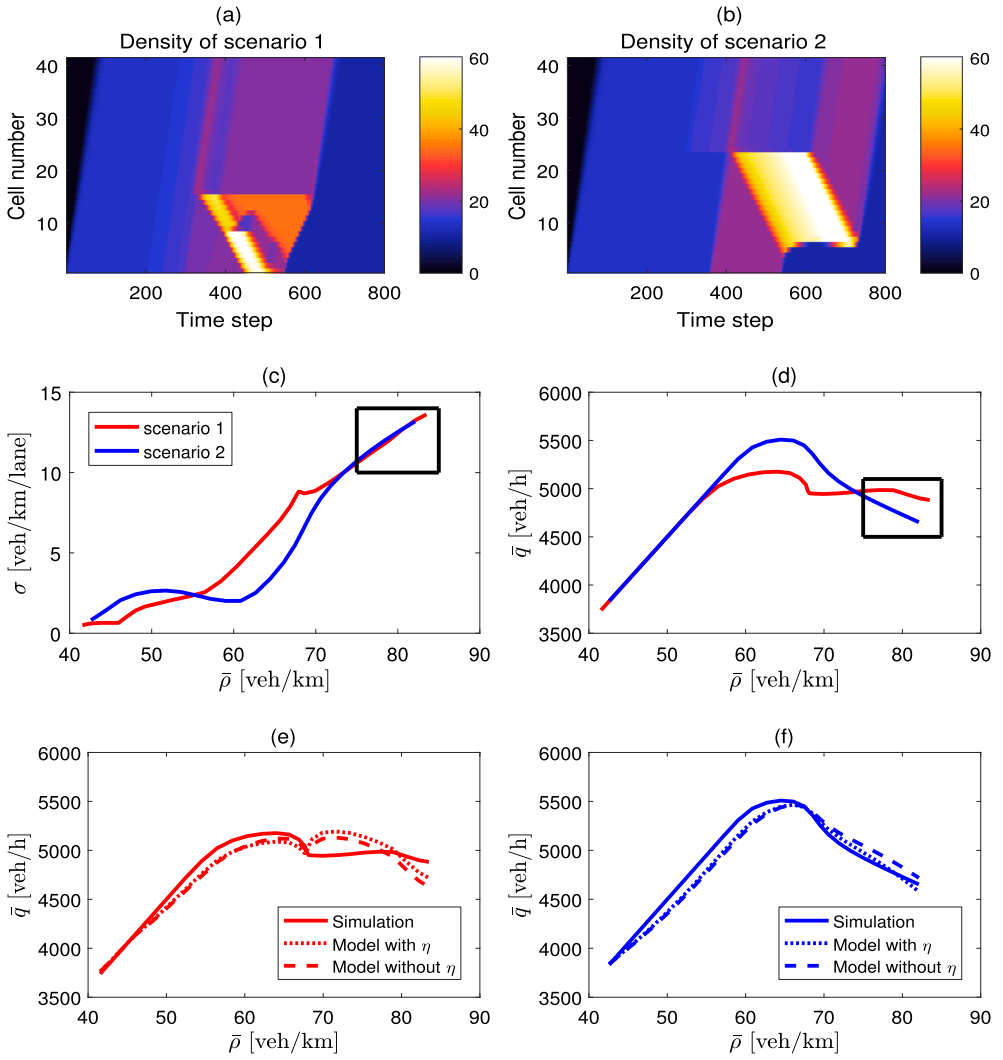

Fig. 7. (a) and (b): The density contour plots of scenario 1 and 2 in the synthetic case. The density of the jam head in (a) is around 35 (veh/km/lane) and in (b) is around 50 (veh/km/lane). (c): The $\bar{\rho}$-σ curves of the two scenarios. (d): The $\bar{\rho}$-$\bar{q}$ curves of the two scenarios. In (c) and (d), when $\bar{\rho}$ of the two scenario are around 73 (veh/km), the values of σ in the two scenarios are similar but the value of $\bar{q}$ in scenario 1 is substantially higher than $\bar{q}$ in scenario 2. (e) and (f) show the accuracy of the MFD model (10) with and without η. The solid lines represent the MFD of the simulation. The dotted lines and dashed lines respectively represent the MFDs with and without η in (10).

drop. Besides, it is expected that with more traffic jams the (negative) influence of the capacity drop on the MFD increases. Accordingly, we introduce a variable η to quantify the effect of the capacity drop, which is,

$$\eta(t) = \sqrt{\sum_{j=1}^{J} (v^{\mathrm{cr}} - v_j^{\mathrm{jam}}(t))^2}, \tag{9}$$

where j is the index of traffic jams, and J is the number of traffic jams occurring in the freeway stretch. v_j^{jam} is the speed in jam j, and v^{cr} is the critical speed associated with the FD critical density and the free-flow capacity. We use the square root of the summation because as the number of jam increases, the density heterogeneity is also expected to increase, which will dilute the influence of η. If the aggregated speed in an area is below the critical value and both upstream and downstream of the area have free flow speed, then the area is perceived as a jam area. v_j^{jam} can be calculated as the average speed in the jam area. Note that first-order traffic flow models (e.g., CTM) consider the density as a state variable. If these models are used, Eq. (9) can be reformulated by using ρ_i^{jam} and ρ^{cr} because there is a one-to-one relationship between those variables.

2.4. Functional form of the freeway MFD model

In (Ramezani et al., 2015), an exponential functional form is found to be a good estimation of MFD of heterogeneous urban

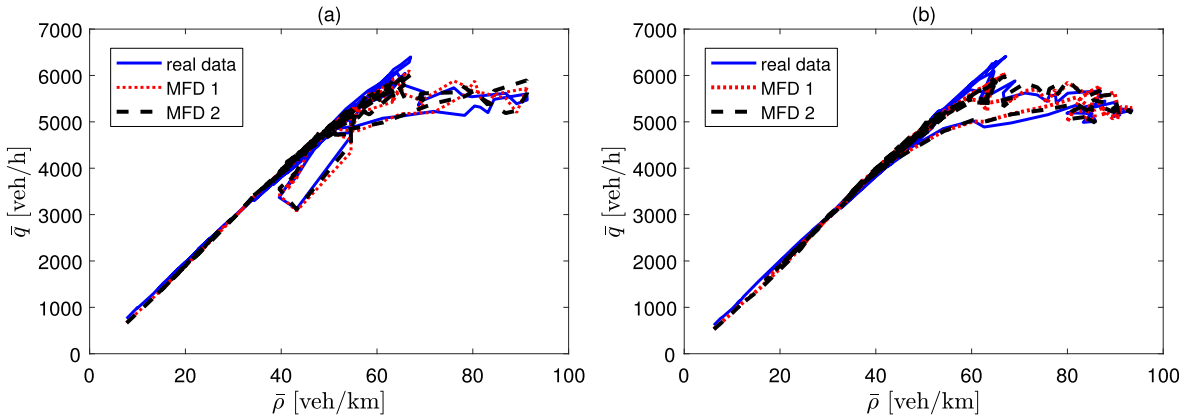

Fig. 8. The estimated MFDs for (a) April 2 and (b) March 27. MFD-1 is the MFD with η and MFD-2 is the one without η. The proposed freeway MFD model can capture the complex hysteresis patterns.

networks. In this paper we use the same functional form as,

$$\bar{q}(t) = (D_3 \cdot \bar{\rho}(t)^3 + D_2 \cdot \bar{\rho}(t)^2 + D_1 \cdot \bar{\rho}(t)) \cdot (a \cdot e^{b_1 \sigma(t) + b_2 \eta(t)} + (1 - a)) \tag{10}$$

where, a, b_1, b_2, D_1, D_2 and D_3 are estimated parameters. The function is a product of a third-order polynomial function which represents a low scatter MFD (i.e. the upper MFD envelope that is associated with the most homogeneous traffic condition) and an exponential function which represents the heterogeneity and capacity drop effects. Note that the choice of the third-order polynomial function is a balance between complexity and accuracy. A second-order function is not able to represent the low scatter MFD because the critical density will be half of the jam density, which is not realistic. Higher order functions increase the complexity of the MFD model.

To examine the accuracy of the proposed MFD model, the parameters of function (10) are estimated based on real data which are obtained from Dutch freeway A13-L. The field data cover March and April 2013. Data from April are used to estimate the parameters in (10), while the data from March are used to test the performance. $\bar{\rho}(t)$, $\sigma(t)$ and $\eta(t)$ are calculated respectively based on (3), (4), and (9). The MFDs with and without η are estimated and tested respectively. The MFD curves of one of the days from the estimation data set (i.e. in April) and from the testing data set (i.e. in March) are shown in Fig. 8. The root mean square error (RMSE) of two MFDs are shown in Table 1. It can be seen that for both MFDs, the estimation error and the testing error are below 5 %. Thus, both MFDs represent the average flow accurately, and the MFDs with η further improves the accuracy. The estimation and testing results of the MFD without σ and η are also shown in Table 1. It can be seen that the errors are significantly higher when the MFD model does not address the effect of heterogeneity and capacity drop. Note that unlike the synthetic case in which we consider different densities of jam heads, the day-to-day pattern of traffic congestion is similar in this field data. This is the reason why the accuracy of the MFD with the introduction of η does not improve significantly. To highlight the significance of η, we estimate the parameters of the MFDs based on simulation data in the previous section. The estimation results are shown in Fig. 7(e) and (f). It can be seen that both MFDs represent the critical point accurately, while the MFD with η reproduces $\bar{q}$ more accurately at the part where both $\bar{\rho}$ and σ have similar values (the rectangle area).

3. Control design

In this section, a hierarchical ramp metering approach is designed to improve the freeway traffic operation efficiency. At the upper level of the controller, a model predictive control (MPC) approach is applied to determine the optimal total inflow from the on-ramps to the freeway stretch. The upper level MPC approach is based on the proposed MFD model. The lower level controller locally distributes the optimal total inflow that is determined by the upper level controller to each on-ramp of the freeway.

Table 1
The estimation and testing results of different MFD models. The field data from April are used to estimate the parameters in (10), i.e. the training set, while the data from March are used to test the performance of (10). That is, once the parameters of (10) are estimated using April data, we keep the parameters fixed and compare the model (10) estimation against March data. In estimation of the MFD without σ and η, b_1 and b_2 are equal to 0. In estimation of the MFD with σ and without η, b_2 is equal to 0.

	Estimation RMSE (veh/h)	Estimation error percentage (%)	Testing RMSE (veh/h)	Testing error percentage (%)
MFD with both σ and η	160.7	4.28	176	4.65
MFD with σ and without η	167.5	4.55	184.8	4.87
MFD without σ and η	273.4	6.31	333.8	8.58

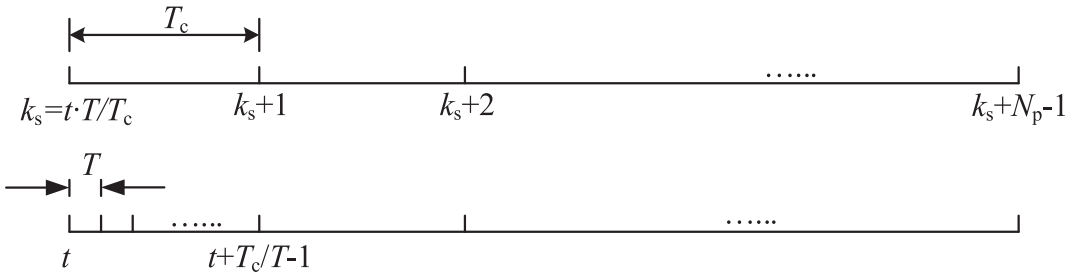

Fig. 9. Time scales of the process model (top) and the prediction model (bottom). Note that N_p adjusts the prediction period in the MPC framework.

3.1. The upper level controller: MPC approach

The MPC approach has largely been used in freeway traffic control problems (e.g. Hegyi et al., 2005a; Haddad et al., 2013). The MPC approach utilizes a traffic model to predict evolution of the traffic state based on the current state of the system, and determines the optimal control actions that lead to the optimum value of an objection function. This feature enables the controller to take advantage of potentially larger future gains at a current (smaller) cost, so as to avoid a myopic control actions. After optimization, the control value of the first sample of the optimal control action is applied to the process. The remaining part of the control signal is recalculated in the finite rolling horizon scheme. For a detailed description of the MPC approach, readers are referred to (Camacho and Bordons, 2012).

In this paper, the proposed MFD model is used as the prediction model of the MPC approach. The model predicts the dynamics of the freeway stretch accumulation A and the total number of vehicles R at all on-ramps. The dynamics of A is formulated as,

$$A(k+1|t) = A(k|t) + (U(k|t) - G(k|t) + f_1(k|t))\cdot T_c, \tag{11}$$

where f_1 is the (uncontrolled) flow from the freeway mainstream origin, t is the discrete time index for the process model and k is the discrete time index for the prediction model, see Fig. 9. For simplicity, we assume that the duration of the discrete time step of the prediction model is the same as the duration of the control time step, i.e. T_c. We assume that $\frac{T_c}{T}$ is an integer, and the controller is activated at time step t of the process model, when the time step of prediction model, $k_s = \frac{t\cdot T}{T_c}$, is an integer. The time scales depiction of the process model and the prediction model are shown in Fig. 9. U is the control variable, which represents the total flows that enter the freeway through all the on-ramps. G is the total outflow of the freeway estimated by the freeway MFD. Based on the empirical analysis in Section 2.1, we assume that the average trip length is time-invariant and $\bar{q}$ and G are readily transformable. Note that the varying trip length between peak hour and off-peak hour shown in Section 2.1 has very limited influence to the proposed control approach, because ramp metering control is activated only during rush hours when the average trip length is relatively constant. If the total outflow of the freeway stretch is not measurable, then G can be estimated as,

$$G(k|t) = \frac{\bar{q}(k|t)\cdot L}{\bar{x}}. \tag{12}$$

We estimate the parameters in (10) based on the measured data of the weighted average densities and the weighted average flows. Similarly, G can be estimated as,

$$G(k|t) = \left(D_3'\left(\frac{A(k|t)}{L}\right)^3 + D_2'\left(\frac{A(k|t)}{L}\right)^2 + D_1'\left(\frac{A(k|t)}{L}\right)\right)\cdot(a'\cdot e^{(b_1'\sigma(k|t)+b_2'\eta(t))} + (1-a')), \tag{13}$$

where parameters a', b_1', b_2', D_1', D_2', and D_3' are estimated based on the measured data of the accumulations and the total outflows. $\eta(t)$ is calculated according to (9) and is assumed to be constant over the prediction horizon, whereas the dynamic of σ is formulated as,

$$\sigma(k|t) = \begin{cases} s_1\cdot\frac{A(k|t)}{L}, & \text{if } \frac{A(k|t)}{L} \leqslant \bar{\rho}^{cr}(t) \\ s_2(\frac{A(k|t)}{L} - \bar{\rho}^{cr}(t))^3 + s_3(\frac{A(k|t)}{L} - \bar{\rho}^{cr}(t))^2 + s_4(\frac{A(k|t)}{L} - \bar{\rho}^{cr}(t)) + s_5, & \text{if } \frac{A(k|t)}{L} > \bar{\rho}^{cr}(t), \end{cases} \tag{14}$$

where s_1, s_2, s_3, s_4, and s_5 are model parameters. Based on the measurements from the field, $\bar{\rho}^{cr}(t)$ is calculated according to (6) and is assumed to be constant over the prediction horizon.

The number of vehicles on all the on-ramps, i.e. R, is modeled as,

$$R(k+1|t) = R(k|t) + T_c\left(\sum_{o=2}^{O} d_o(k|t) - U(k|t)\right), \tag{15}$$

where o is the index of origins of the freeway ($o = 1$ is the freeway mainstream origin, and $o = 2, 3, \ldots O$ are on-ramps), and $d_o(k)$ is the demand of origin o at time step k. Eq. 15 models the number of the vehicles on all the on-ramps as a single and continuum queuing system while neglecting the storage capacity of the on-ramps. Note that the controller guarantees R to remain positive (see Eq. 16)

while the lower level controller control and consider the dynamics of individual ramps.

We define the MPC objective function to minimize the total time spent (TTS) of vehicles on the freeway and on-ramps during the prediction horizon. The overall optimization problem is,

$$\min_{U(k|t)} \sum_{k=\frac{t\cdot T}{T_c}}^{\frac{t\cdot T}{T_c}+N_p-1} (A(k|t)+R(k|t))\cdot T_c,$$

subject to

(10)–(15),

$$A(k|t),\ R(k|t),\ \sigma(k|t),\ P(k|t),\ U(k|t) \geqslant 0,\ \forall\ k = k_s, k_s+1, \ldots, k_s+N_p-1. \tag{16}$$

3.2. The lower level controller: local ramp metering

The lower level controller is devised to distribute the optimal total inflow calculated by the MPC upper level controller (see Eq. (16)) to each on-ramp of the freeway stretch. Based on the optimal total inflow, U, obtained from the MPC, the total amount of flows that are needed to be restricted by the on-ramps, $Q(t)$, is

$$Q(t) = \sum_{o=2}^{O} f_o(t) - U\left(\frac{t\cdot T}{T_c}\ \middle|\ t\right) \tag{17}$$

where $f_o(t)$ is the on-ramp o flow without ramp metering control and is estimated by a simple node model as,

$$f_o(t) = \min\left(q_o, \frac{n_o(t)}{T_c}, S_o(t)\cdot\frac{q_o}{q_o+c_o}\right) \tag{18}$$

where q_o is the capacity of on-ramp o and c_o is the capacity of the mainstream at the downstream of on-ramp o. $n_o(t)$ is the queue length of on-ramp o. $S_o(t)$ is the receiving capacity at the downstream of on-ramp o, which is calculated according to the CTM based on the measurement of density at the downstream of on-ramp o.

To distribute $Q(t)$ to each on-ramp of the freeway stretch, we divide the freeway stretch into sections which include a number of segments. An on-ramp is located in the middle of each section such that flows around the section can be regulated by ramp metering control. $Q(t)$ is distributed to each on-ramp based on the weight of each section. The weight of section s, i.e. w_s, is

$$w_s = \begin{cases} 0 & \text{if } \forall\ i \in I_s,\ \rho_i \leqslant \rho_i^{cr} \\ \dfrac{\sum_{i\in I_s}\rho_i l_i}{\sum_{i\in I_s} l_i}\cdot(\max_{i\in I_s}(\rho_i) - \min_{i\in I_s}(\rho_i)) & \text{if } \exists\ i \in I_s,\ \rho_i > \rho_i^{cr} \end{cases} \tag{19}$$

where I_s is the set of segments in section s. $\max_{i\in I_s}(\rho_i)$ is the maximum measured density of segments in section s, and $\min_{i\in I_s}(\rho_i)$ is the minimum measured density of segments in section s. The term $\sum_{i\in I_s}\rho_i l_i / \sum_{i\in I_s} l_i$ represents the weighted average density of section s and the term $\max_{i\in I_s}(\rho_i) - \min_{i\in I_s}(\rho_i)$ approximates the density spread over section s. The weight w_s can be interpreted as the combination of the average density and density variance of section s. Then, the ramp metering rate of on-ramp o, which is located in section s is

$$r_o(t) = f_o(t) - Q(t)\cdot\frac{w_s}{\sum w_s}. \tag{20}$$

In case $\rho_i \leqslant \rho_i^{cr}$ for all freeway sections, all w_s become equal to zero that forces all the ramp meters to become inactive, i.e. $r_o(t) = f_o(t)$ for all on-ramps. Ultimately, the number of vehicles on on-ramp o, i.e. n_o, and the number of vehicles on all the on-ramps, i.e. R, are updated as

$$n_o(t+1) = n_o(t) + T_c(d_o(t) - r_o(t)) \tag{21}$$

$$R(t) = \sum_{o=2}^{O} n_o(t). \tag{22}$$

4. Numerical experiments

The presented hierarchical control approach is applied to several numerical experiments to test the performance of the proposed MFD based ramp metering. For comparison purposes, MPC controllers that are based on different prediction models are also applied to the numerical experiments. To scrutinize the robustness of the controllers, we use two different models as the process model, namely the extended CTM and the METANET model. Both models are able to reproduce the capacity drop and other freeway traffic phenomena, such as the propagation of moving jams, with reasonable accuracy. For detail discussion and comparison between the two models, readers are referred to (Han et al., 2017a,b).

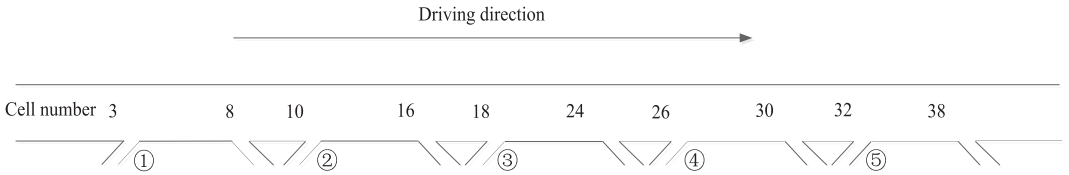

Fig. 10. The layout of the synthetic freeway stretch. Cell numbers are marked at the upstream of each on-ramp and off-ramp. The order of on-ramps are shown as circled numbers.

The performances in terms of total time spent (TTS) of all vehicles using the freeway and all on-ramps and the computation time of the controllers are evaluated. The prediction and control horizon parameters, N_p and N_c, are tuned based on the heuristic rules presented in (Hegyi et al., 2005b). To ensure a fair performance comparison among the MPC controllers with different prediction models, the prediction and control horizon parameters are set equal in all the case studies and scenarios. Note that in each case study, the parameters of MFD, CTM, and METANET are estimated based on the No Control scenario.

4.1. Case study 1-The extended CTM as the process model

In this case study, the extended CTM presented by Han et al. (2016) is used as the process model. The 3-lane synthetic freeway stretch is 16 km in length and it is divided into 40 cells. It includes 5 on-ramps and 5 off-ramps, as shown in Fig. 10. The parameters of the process model are set as the following. The free-flow speed v is 100 [km/h]. The congestion wave speed β is set to 20 [km/h]. The capacity c is 2000 [veh/h/lane]. The maximum extent of the capacity drop α is set to 30%. The capacities of all the on-ramps are set to 2000 [veh/h]. The time duration of a simulation step T is set to 10 [s]. Off-ramp splitting ratio is 0.2 for each off-ramp at every time step. The demand profiles of the mainstream origin and each on-ramp are shown in Fig. 11. The simulation period is 6 h (6:00–12:00) that contains 2 h of peak time (7:00–9:00). The density contour plot of the simulation is shown in Fig. 12(a). It can be seen that two traffic jams are triggered at on-ramps 3 and 5 at the peak time, due to increasing demand at the on-ramps.

Five control scenarios are investigated in this case study. The objective of each control scenario is the same, which is minimizing the TTS in the prediction horizon. The prediction horizon N_p is set to one hour. The control sampling time T_c is set to 2 min and the first control sample of the optimization is applied to the process. The set up of five scenarios are:

1. Scenario 1. The system optimal scenario, where in this scenario, the extended CTM is used as the prediction model, so the process and the prediction models are the same. This scenario is used as the upper achievable bound of control performance.
2. Scenario 2. The proposed control approach is applied in this scenario. That is, the proposed MFD model is used as the prediction model. The total outflow is measurable from the simulation, thus (13) is used to represent the dynamics of G. The parameters in (13) and (14) are estimated with the data from the simulation of the process model.
3. Scenario 3. The same control approach as scenario 2 is applied, except that the prediction model is the MFD without the heterogeneity and capacity drop reduction, i.e. the exponential term in (13) is set to one. This scenario intends to explore the significance of the density heterogeneity modeling in the performance of the controller.
4. Scenario 4. The original CTM is used as the prediction model in this scenario. The parameters of the original CTM are set to the same as the extended CTM (the process model), except that the parameter which represents the extent of the capacity drop is set to 0.
5. Scenario 5. The second-order model METANET is used as the prediction model in this scenario. The METANET model is estimated with the data from the simulation of the process model. The parameter estimation problem is formulated as a non-linear optimization problem which aims to minimize the discrepancy between the simulation results of the process model and the prediction

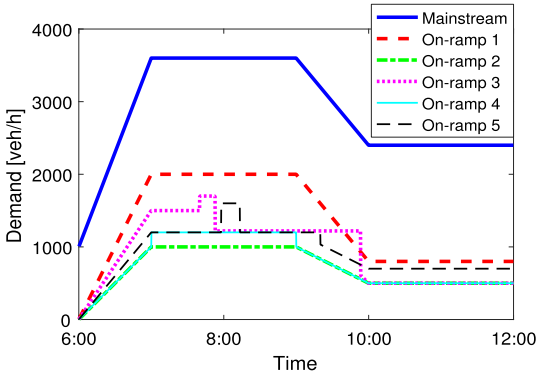

Fig. 11. The demand profile of the simulation.

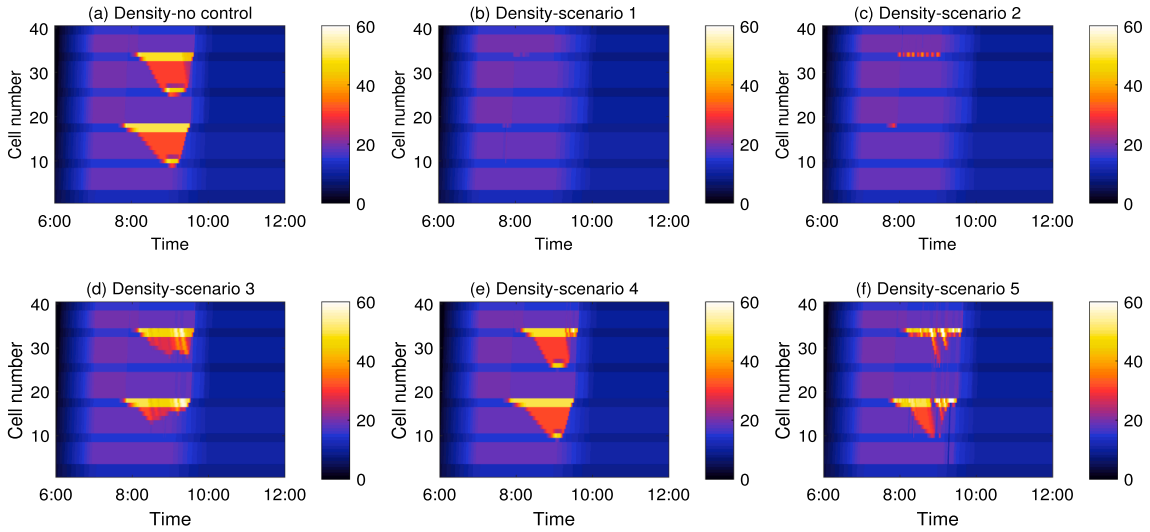

Fig. 12. The density (veh/km/lane) contour plots of the scenario without control and all five control scenarios in case study 1. On-ramps 3 and 5 are the main bottlenecks. Evidently, the proposed MFD-based ramp metering control strategy (i.e. scenario 2) efficiently reduces traffic density at on-ramp locations. Scenario 1 represents the theoretical upper bounds of MPC performance.

model. The non-linear and non-convex optimization problem is solved by MATLAB implementation of the SQP algorithm. To ensure that a global optimum is achieved, 100 starting points are used.

Simulation results are shown in Fig. 12 and Fig. 13, where Fig. 12(b-f) show the density contour plots of the five control scenarios and Fig. 13 shows the queue length and ramp flow of each ramp for every control scenario. The TTS and the computation time of each scenario are listed in Table 2, where it can be observed that the TTS of control scenario 2 (the proposed control approach) is the closest to the system optimal TTS (scenario 1). The control dynamics of both scenarios (1 and 2) are similar where ramp metering controls are activated at the proper locations (on-ramps 3 and 5) and time (when the bottlenecks are activated). Fig. 13 depicts that ramp queues are mainly accumulated at on-ramps 3 and 5 with both scenarios. The proposed controller is efficient and effective because density heterogeneity is precisely modeled and predicted so that the reduction of the outflow due to the increase in vehicle accumulation is captured by the freeway MFD.

In contrast to the proposed controller, the controller in scenario 3 has an inferior performance, because the prediction model of the controller does not take the density heterogeneity and capacity drop into consideration. When breakdown occurs in the freeway, the density heterogeneity increases rapidly with the increase of the accumulation, which results in a rapid decrease of the total outflow. Since the MFD in scenario 3 does not model this effect, the controller fails to generate effective ramp metering control signals to resolve the congestion. For the controller in scenario 4, since the prediction model does not incorporate the capacity drop, the only gain of the controller comes from avoiding the off-ramp blockade. The controller underestimates the benefits of applying ramp metering. Therefore, the ramp metering rates are not sufficiently low and on time to resolve the congestion, see Fig. 13 where scenario 4 ramp metering is active late at on-ramp 5. For the METANET-based MPC (scenario 5), due to the mismatch between the prediction model and the process model, control signals are ineffective that lead to inferior control performance. For example, as depicted in Fig. 13, it can be seen that when freeway mainstream congestion is triggered at around 8:00 at on-ramp 5, there is no control action at on-ramp 5. While at around 9:00, the controller performs unnecessary actions at on-ramps 2 and 4. These ineffective control actions significantly decrease the performance of the controller.

Fig. 14 shows MFDs of all control scenarios in this case study and Case study 2. The figure clearly demonstrates that the MFD-based (without σ) control scenario (Scenario 3), the CTM-based control scenario (Scenario 4), and the METANET-based control scenario (Scenario 5) exhibit hysteresis loops. MFDs of the control scenarios with the system optimal control strategy and the presented control approach donot exhibit hysteresis, thanks to effective ramp metering control schemes. The relation between ramp metering control strategy and the size of hysteresis loop will be investigated in the future.

The computation efficiency of all control scenarios are shown in Table 2. A computer with an E5-1620 processor and 16 GB RAM is used for the optimizations. The computation time of the controllers have strong correlation with the complexity of the prediction model. The MFD-based controllers (scenarios 2 and 3) have less computation times, the first-order model-based controllers (scenarios 1 and 4) have higher computation times, and the second-order model-based controller (scenario 5) has the highest computation time. This is of great importance for practical implementation of real-time control strategies. Note that some efficient computational algorithms, such as the feasible-direction algorithm, have been proposed to increase the computation speed of the METANET-based controllers (Kotsialos et al., 1999). In this paper, we use the standard SQP algorithm for every scenario.

Note that in this case study, the estimated value of parameter b_2' in (13), which represents the influence of capacity drop on MFD,

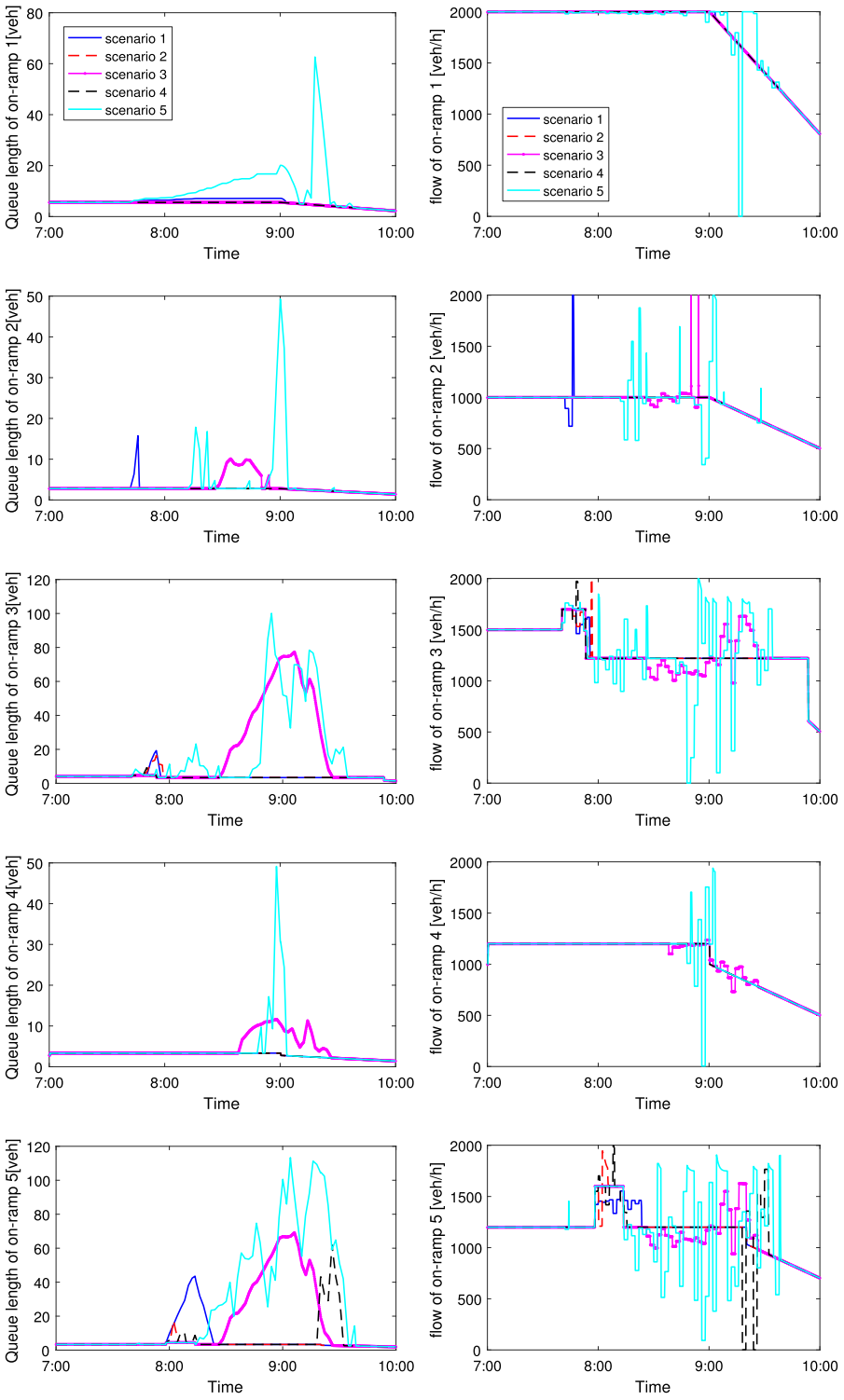

Fig. 13. The queue lengths (on left) and ramp flows (on right) of on-ramps 1–5 of the five control scenarios in case study 1.

is close to 0 because the effect of capacity drop is captured by the density heterogeneity. This is because the temporal duration of capacity drop phenomenon is limited compared to the period that the freeway is congested. Hence, to explore the influence of capacity drop on the control performance, traffic scenarios with extreme extent and number of capacity drop and moving jams should be studied. This is a future research direction.

Table 2
The performance of each control scenario in case study 1. Total Time Spent (TTS) considers both the freeway and all on-ramps.

Scenario	TTS (h) Mainstream + Onramp	Improvement (%)	Computation time (s/control step)
Without control	4218.8 (4141.0 + 77.8)	0	–
1. System optimal	3775.6 (3686.0 + 89.6)	10.51	20.9
2. MFD-based	3807.5 (3697.7 + 109.8)	9.74	3.1
3. MFD-based (no σ)	4170.2 (4009.7 + 160.5)	1.15	2.6
4. CTM-based	4173.2 (4087.2 + 86.0)	1.08	9.6
5. METANET-based	4164.0 (3950.9 + 213.1)	1.30	252.9

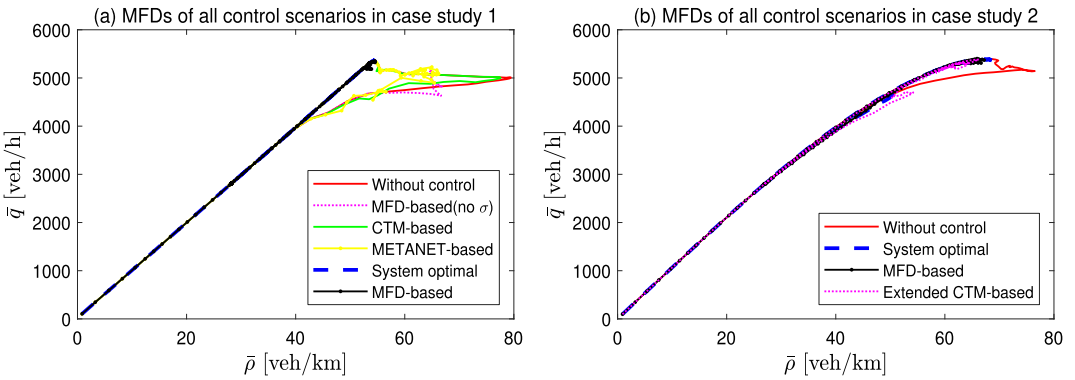

Fig. 14. MFDs of all control scenarios in case study 1 and case study 2. In both case studies, MFDs of the control scenarios with the system optimal control strategy and the presented control approach donot exhibit hysteresis, owning to effective ramp metering control schemes. The MFDs of other control scenarios exhibit different size of hysteresis loops.

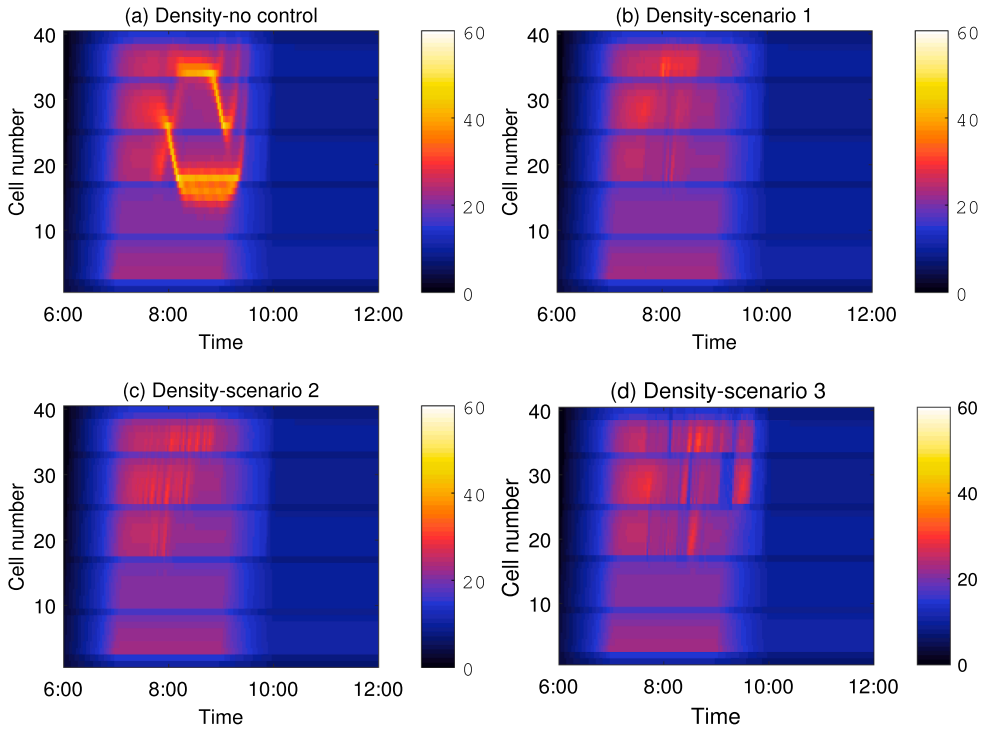

Fig. 15. The density contour plots of the scenario without control and all three control scenarios in case study 2.

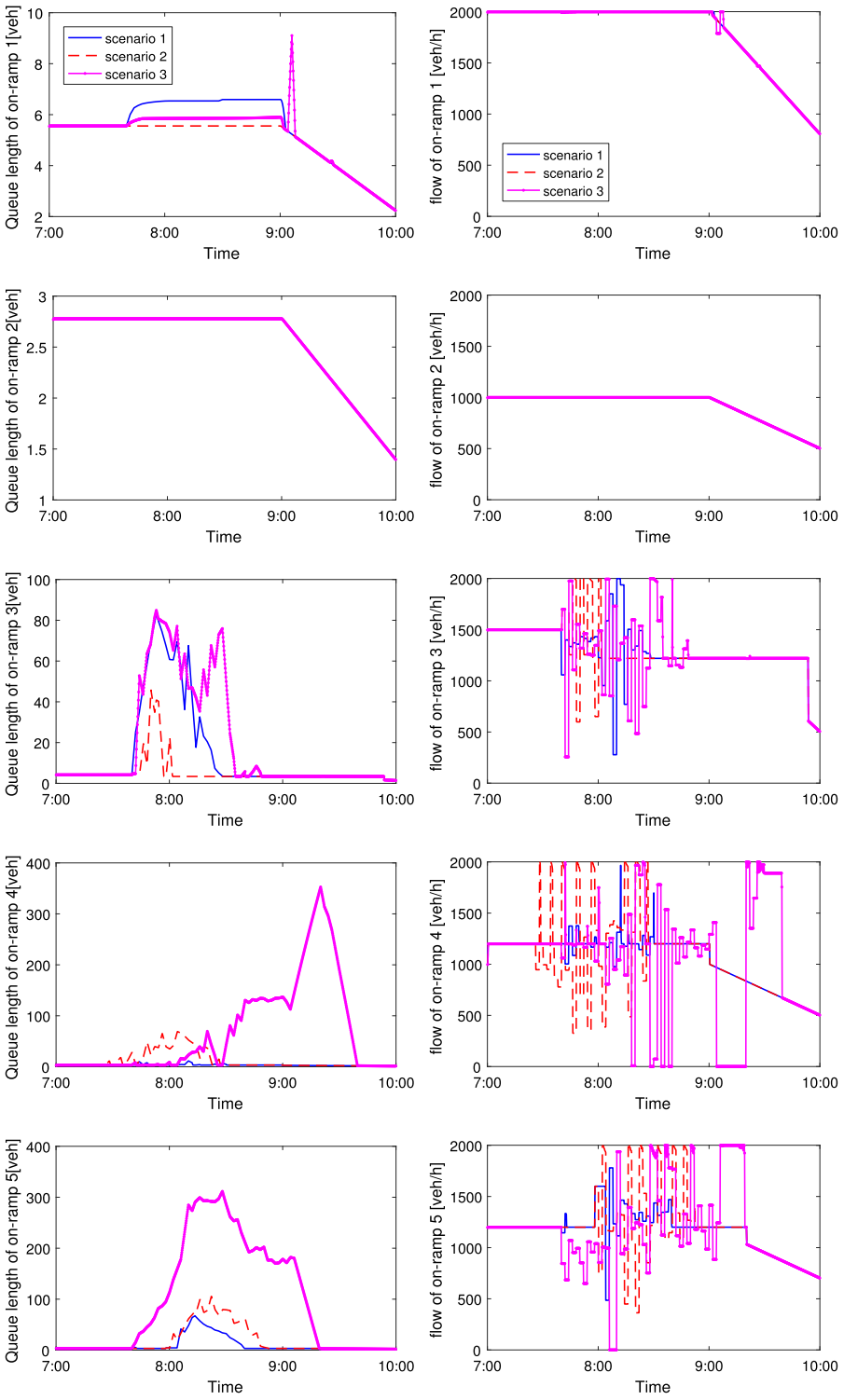

Fig. 16. The queue lengths (on left) and ramp flows (on right) of on-ramps 1–5 of the three control scenarios in case study 2.

Table 3
The performance of each control scenario in case study 2. Total Time Spent (TTS) considers both the freeway and all on-ramps.

Scenario	Without control	1. System optimal	2. MFD-based	3. Extended CTM-based
TTS (h) (Mainstream + Onramp)	4341.2 (4263.3 + 77.8)	4143.5 (4020.0 + 123.5)	4163.7 (4018.4 + 145.3)	4619.2 (4012.7 + 606.5)
Improvements (%)	0	4.6	4.1	−6.4
Computation time (s/control step)	–	246.3	3.0	22.1

4.2. Case study 2-METANET as the process model

To further explore properties and the robustness of the presented controller, the performance of the controller is tested where the METANET model is employed as the process model. The parameters of the METANET model are set to same values as the calibrated parameters values in scenario 5 of the previous section. We apply the same demand profile as in Fig. 11. The density contour plot of the no control scenario is shown in Fig. 15(a). Three control scenarios are considered as:

1. Scenario 1. The system optimal scenario where the METANET model is used as both the process and the prediction models. This scenario demonstrates the upper bound of MPC control performance.
2. Scenario 2. The proposed control approach is applied to this scenario. The parameters in (13) and (14) are calibrated with the data from the simulation of the process model.
3. Scenario 3. The extended CTM is used as the prediction model in this scenario. The extended CTM is calibrated with the data from the simulation of the process model.

Results are presented in Fig. 15 and Fig. 16, where Fig. 15(b-d) show the density contour plots of the three control scenarios and Fig. 16 shows the queue length and ramp flow of each ramp for every control scenario. The TTS achieved by the controllers are listed in Table 3. Fig. 15 shows that all three scenarios resolve the congestion in the freeway mainstream. However, for the extended CTM-based controller (scenario 3), the TTS is significantly higher than the ones achieved by the system optimal scenario and the proposed control approach (scenario 2). The differences of the control performance stem from the effectiveness of the control actions. Fig. 16 demonstrates that the queue lengths at the bottlenecks (on-ramps 3 and 5) are similar in scenarios 1 and 2. Whereas in scenario 3, the controller applies unnecessarily low ramp flows (e.g., after 8:00 at on-ramp 5 and after 9:00 at on-ramp 4), which results in longer ramp queues and thus larger time delays. The mismatch between the process model and the prediction model is the main reason that causes the inferior performance of scenario 3. The results of case studies 1 and 2 show that the presented control approach has a robust performance to overcome the mismatch between the prediction model and the process model in the MPC framework.

In this case study, the TTS improvement of the system optimal scenario is 4.6 % while it is 10.5 % in case study 1. The reason is that two plant models have different mechanisms and different parameter settings, which results in different congested situations. In this case study, the computation time of scenario 1 (system optimal), scenario 2 (MFD-based), and scenario 3 (extended CTM-based) are comparable with scenario 5 (METANET-based), scenario 2 (MFD-based), and scenario 1 (system optimal) in Section 4.1.

5. Conclusion

This paper has proposed a parsimonious large-scale model for coordinated ramp metering of a freeway stretch based on a hierarchical control structure. At the upper level, a model predictive control (MPC) approach is developed to optimize the total inflow from on-ramps to the freeway stretch. The prediction model integrated in the MPC approach is a parsimonious MFD-type model that relates the freeway network outflow to vehicle accumulation on the freeway, dispersion of vehicles density on the freeway, and the occurrence of capacity drops. The paper has investigated the properties of freeway MFD based on field data and presented a model to predict density heterogeneity and freeway traffic dynamics.

The lower level controller hypercritically distributes the advised total inflow by the MPC among each on-ramp of the freeway based on local traffic state feedback. The proposed control approach demonstrates a performance close to the optimal MPC control in reducing the total time spent and eliminating congestion. The robustness of the controller is tested by using different traffic flow models (the extended CTM and the METANET model) as the process model and compared with MPC controllers that are based on different prediction models. Simulation results highlight two advantages of the proposed controller, (i) the robustness to overcome the mismatch between the prediction model and the process model, and (ii) the applicability for field implementation as it requires less computation efforts than other non-linear optimal controllers.

For future research, the proposed freeway MFD model and control approach can be developed to tackle congestion control in a freeway network that consists of multiple freeway stretches. Furthermore, the lower level of the hierarchical control structure can be accommodated by other sophisticated local ramp metering strategies. Control measures will not only be limited to ramp metering, but also can include the variable speed limits and individual lane density management, e.g., (Ramezani and Ye, 2019). The influence of average trip length and capacity drop on MFD estimation also need further investigation with different filed datasets.

Appendix A. Parameter values

See Tables A.4–A.8.

Table A.4
The numerical values of the regression parameters in Fig. 5.

s_1	s_2	s_3	s_4	s_5
3.951×10^{-2}	1.522×10^{-3}	7.767×10^{-2}	1.388	1.702

Table A.5
The numerical values of the regression parameters in Fig. 7.

	D_3	D_2	D_1	a	b_1	b_2
MFD with both σ and η	96.964	−0.177	7.827×10^{-4}	8.116×10^{-2}	0.120	3.874×10^{-3}
MFD with σ and without η	65.325	1.320	1.669×10^{-2}	9501.942	-3.901×10^{-7}	-

Table A.6
The numerical values of the regression parameters in Table 1.

	D_3	D_2	D_1	a	b_1	b_2
MFD with both σ and η	80.337	1.170	1.247×10^{-2}	1.508	2.107×10^{-2}	1.189×10^{-3}
MFD with σ and without η	78.464	1.397	1.429×10^{-2}	0.971	4.800×10^{-2}	–
MFD without σ and η	89.780	0.645	1.085×10^{-2}	–	–	–

Table A.7
The numerical values of MFD parameters in numerical experiments.

	D_3'	D_2'	D_1'	a'	b_1'
MFD with σ - case 1	110.988	4.176	4.594×10^{-2}	316.044	-5.973×10^{-5}
MFD without σ - case 1	124.150	3.778	4.903×10^{-2}	-	-
MFD with σ - case 2	166.758	1.857	2.946×10^{-2}	38.759	1.479×10^{-4}

Table A.8
The numerical values of σ parameters in numerical experiments.

	s_1	s_2	s_3	s_4	s_5
case 1	3.951×10^{-2}	1.522×10^{-3}	-7.767×10^{-2}	1.388	1.702
case 2	3.835×10^{-2}	3.360×10^{-3}	-9.281×10^{-2}	1.133	2.206

References

Ampountolas, K., Zheng, N., Geroliminis, N., 2017. Macroscopic modelling and robust control of bi-modal multi-region urban road networks. Transp. Res. Part B: Methodol.

Batista, S., Leclercq, L., Geroliminis, N., 2019. Estimation of regional trip length distributions for the calibration of the aggregated network traffic models. Transp. Res. Part B: Methodol. 122, 192–217.

Buisson, C., Ladier, C., 2009. Exploring the impact of homogeneity of traffic measurements on the existence of macroscopic fundamental diagrams. Transp. Res. Rec. 2124 (1), 127–136.

Camacho, E.F., Bordons, C., 2012. Model Predictive Control in the Process Industry. Springer Science & Business Media.

Cassidy, M., Jang, K., Daganzo, C., 2011. Macroscopic fundamental diagrams for freeway networks: theory and observation. Transp. Res. Rec. J. Transp. Res. Board 2260, 8–15.

Chow, A.H., 2015. Optimisation of dynamic motorway traffic via a parsimonious and decentralised approach. Transp. Res. Part C: Emerg. Technol. 55, 69–84.

Daganzo, C.F., 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transp. Res. Part B: Methodol. 28 (4), 269–287.

Daganzo, C.F., 2007. Urban gridlock: macroscopic modeling and mitigation approaches. Transp. Res. Part B: Methodol. 41 (1), 49–62.

Daganzo, C.F., 2011. On the macroscopic stability of freeway traffic. Transp. Res. Part B: Methodol. 45 (5), 782–788.

Daganzo, C.F., Gayah, V.V., Gonzales, E.J., 2011. Macroscopic relations of urban traffic variables: bifurcations, multivaluedness and instability. Transp. Res. Part B:

Methodol. 45 (1), 278–288.
Gayah, V.V., Daganzo, C.F., 2011. Clockwise hysteresis loops in the macroscopic fundamental diagram: an effect of network instability. Transp. Res. Part B: Methodol. 45 (4), 643–655.
Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transp. Res. Part B: Methodol. 42 (9), 759–770.
Geroliminis, N., Sun, J., 2011a. Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transp. Res. Part B: Methodol. 45 (3), 605–617.
Geroliminis, N., Sun, J., 2011b. Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. Transp. Res. Part A: Policy Pract. 45 (9), 966–979.
Geroliminis, N., Haddad, J., Ramezani, M., 2013. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: a model predictive approach. IEEE Trans. Intell. Transp. Syst. 14 (1), 348–359.
Gomes, G., Horowitz, R., 2006. Optimal freeway ramp metering using the asymmetric cell transmission model. Transp. Res. Part C: Emerg. Technol. 14 (4), 244–262.
Haddad, J., 2015. Robust constrained control of uncertain macroscopic fundamental diagram networks. Transp. Res. Part C: Emerg. Technol. 59, 323–339.
Haddad, J., Ramezani, M., Geroliminis, N., 2013. Cooperative traffic control of a mixed network with two urban regions and a freeway. Transp. Res. Part B: Methodol. 54, 17–36.
Han, Y., Yuan, Y., Hegyi, A., Hoogendoorn, S.P., 2015. Linear quadratic mpc for integrated route guidance and ramp metering. In: Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on. IEEE, pp. 1150–1155.
Han, Y., Yuan, Y., Hegyi, A., Hoogendoorn, S.P., 2016. New extended discrete first-order model to reproduce propagation of jam waves. Transp. Res. Rec. J. Transp. Res. Board 2560, 108–118.
Han, Y., Hegyi, A., Yuan, Y., Hoogendoorn, S., 2017a. Validation of an extended discrete first-order model with variable speed limits. Transp. Res. Part C: Emerg. Technol. 83, 1–17.
Han, Y., Hegyi, A., Yuan, Y., Hoogendoorn, S., Papageorgiou, M., Roncoli, C., 2017b. Resolving freeway jam waves by discrete first-order model-based predictive control of variable speed limits. Transp. Res. Part C: Emerg. Technol. 77, 405–420.
Hegyi, A., De Schutter, B., Hellendoorn, H., 2005a. Model predictive control for optimal coordination of ramp metering and variable speed limits. Transp. Res. Part C: Emerg. Technol. 13 (3), 185–209.
Hegyi, A., De Schutter, B., Hellendoorn, J., 2005b. Optimal coordination of variable speed limits to suppress shock waves. IEEE Trans. Intell. Transp. Syst. 6 (1), 102–112.
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. J. Transp. Res. Board 2161, 40–48.
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 (10), 1393–1403.
Keyvan-Ekbatani, M., Yildirimoglu, M., Geroliminis, N., Papageorgiou, M., 2015. Multiple concentric gating traffic control in large-scale urban networks. IEEE Trans. Intell. Transp. Syst. 16 (4), 2141–2154.
Knoop, V., Hoogendoorn, S., 2013. Empirics of a generalized macroscopic fundamental diagram for urban freeways. Transp. Res. Rec. J. Transp. Res. Board 2391, 133–141.
Kotsialos, A., Papageorgiou, M., Messmer, A., 1999. Optimal coordinated and integrated motorway network traffic control. In: 14th International Symposium on Transportation and Traffic Theory.
Kotsialos, A., Papageorgiou, M., 2004. Nonlinear optimal control applied to coordinated ramp metering. IEEE Trans. Control Syst. Technol. 12 (6), 920–933.
Kouvelas, A., Saeedmanesh, M., Geroliminis, N., 2017. Enhancing model-based feedback perimeter control with data-driven online adaptive optimization. Transp. Res. Part B: Methodol. 96, 26–45.
Laval, J.A., Castrillón, F., 2015. Stochastic approximations for the macroscopic fundamental diagram of urban networks. Transp. Res. Part B: Methodol. 81, 904–916.
Laval, J.A., Leclercq, L., Chiabaut, N., 2017. Minimal parameter formulations of the dynamic user equilibrium using macroscopic urban models: Freeway vs city streets revisited. Transp. Res. Part B: Methodol.
Leclercq, L., Chiabaut, N., Trinquier, B., 2014. Macroscopic fundamental diagrams: a cross-comparison of estimation methods. Transp. Res. Part B: Methodol. 62, 1–12.
Leclercq, L., Parzani, C., Knoop, V.L., Amourette, J., Hoogendoorn, S.P., 2015. Macroscopic traffic dynamics with heterogeneous route patterns. Transp. Res. Part C: Emerg. Technol. 59, 292–307.
Mazloumian, A., Geroliminis, N., Helbing, D., 2010. The spatial variability of vehicle densities as determinant of urban network capacity. Philos. Trans. Roy. Soc. Lond. A: Math. Phys. Eng. Sci. 368 (1928), 4627–4647.
Papageorgiou, M., Kotsialos, A., 2000. Freeway ramp metering: An overview. In: Intelligent Transportation Systems, 2000. Proceedings. 2000 IEEE. IEEE, 2000, pp. 228–239.
Papageorgiou, M., Hadj-Salem, H., Blosseville, J.-M., 1991. ALINEA: a local feedback control law for on-ramp metering. Transp. Res. Rec. 1320 (1), 58–67.
Papamichail, I., Kotsialos, A., Margonis, I., Papageorgiou, M., 2010. Coordinated ramp metering for freeway networks–a model-predictive hierarchical control approach. Transp. Res. Part C: Emerg. Technol. 18 (3), 311–331.
Ramezani, M., Nourinejad, M., 2017. Dynamic modeling and control of taxi services in large-scale urban networks: a macroscopic approach. Transp. Res. Part C: Emerg. Technol.
Ramezani, M., Ye, E., 2019. Lane density optimisation of automated vehicles for highway congestion control. Transp. B: Transp. Dyn. 1–21.
Ramezani, M., Haddad, J., Geroliminis, N., 2015. Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control. Transp. Res. Part B: Methodol. 74, 1–19.
Saberi, M., Mahmassani, H., 2012. Exploring properties of networkwide flow-density relations in a freeway network. Transp. Res. Rec. J. Transp. Res. Board 2315, 153–163.
Saberi, M., Mahmassani, H., Hou, T., Zockaie, A., 2014. Estimating network fundamental diagram using three-dimensional vehicle trajectories: extending edie's definitions of traffic flow variables to networks. Transp. Res. Rec. J. Transp. Res. Board 2422, 12–20.
Yang, K., Zheng, N., Menendez, M., 2017. Multi-scale perimeter control approach in a connected-vehicle environment. Transp. Res. Part C: Emerg. Technol.
Yildirimoglu, M., Ramezani, M., Geroliminis, N., 2015. Equilibrium analysis and route guidance in large-scale networks with mfd dynamics. Transp. Res. Part C: Emerg. Technol. 59, 404–420.
Yuan, K., Knoop, V.L., Hoogendoorn, S.P., 2015. Capacity drop: Relationship between speed in congestion and the queue discharge rate. Transp. Res. Rec. 2491 (1), 72–80.