# CAPACITY DROP IN FIRST-ORDER TRAFFIC FLOW MODELS: OVERVIEW AND REAL-DATA VALIDATION

**Maria Kontorinaki, Corresponding Author**
Dynamic Systems and Simulation Laboratory,
Technical University of Crete,
University Campus, Kounoupidiana, Chania 73100, Crete (Greece)
Tel: +30 28210-37307; Email: mkontorinaki@dssl.tuc.gr

**Anastasia Spiliopoulou**
Dynamic Systems and Simulation Laboratory,
Technical University of Crete,
University Campus, Kounoupidiana, Chania 73100, Crete (Greece)
Email: natasa@dssl.tuc.gr

**Claudio Roncoli**
Dynamic Systems and Simulation Laboratory,
Technical University of Crete,
University Campus, Kounoupidiana, Chania 73100, Crete (Greece)
Email: croncoli@dssl.tuc.gr

**Markos Papageorgiou**
Dynamic Systems and Simulation Laboratory,
Technical University of Crete,
University Campus, Kounoupidiana, Chania 73100, Crete (Greece)
Email: markos@dssl.tuc.gr

Word count: 5,230 words text + 8 tables/figures x 250 words (each) = 7,230 words + 35 References

Submitted for Presentation at the 2016 Transportation Research Board 95th Annual Meeting
2nd Submission Date: November 13th, 2015

## ABSTRACT

First-order traffic flow models are known for their simplicity and computational efficiency and have, for this reason, been widely used for various traffic engineering tasks. However, first-order models are not able to reproduce significant traffic phenomena of great interest such as the capacity-drop and stop-and-go waves. This paper presents an overview of the, so far, proposed modelling approaches which aim to introduce the capacity-drop phenomenon into first-order traffic flow models. The background and main characteristics of each approach are analyzed with a particular emphasis on the practical applicability of such models for traffic management and control. The presented modelling approaches are calibrated and tested using real data from a motorway network in U.K.

*Keywords*: Capacity-drop phenomenon, First-order traffic flow models, model calibration

## INTRODUCTION

Among numerous phenomena characterizing traffic flow behavior, one of the most known and puzzling is the so-called capacity drop. This phenomenon breeds the reduction in the mainstream flow of a motorway when a queue starts forming upstream of a bottleneck location *(1, 2)*. Bottleneck locations can be motorway merge areas, areas with particular infrastructure layout (such as lane drops, strong grade or curvature, tunnels etc.), areas with specific traffic conditions (e.g. strong weaving of traffic streams), areas with external capacity-reducing events (e.g. work-zones, incidents) etc. *(3–5)*. If the arriving demand is higher than the bottleneck capacity, i.e. the maximum flow that can pass through a reference point at a certain time period, the bottleneck is activated, i.e. congestion is formed upstream of the bottleneck location. Empirical observations show that, whenever a bottleneck is activated, the maximum outflow that materializes (also called discharge flow) may be 5 to 20 percent lower than the nominal bottleneck capacity. The capacity drop is then defined as the difference between these two values of flow, i.e. the capacity and the discharge flow. Certainly, the capacity drop reflects infrastructure degradation, leading to increased vehicles' travel times and longer delays. To avoid or delay the activation of a bottleneck, and the related capacity drop phenomenon, various traffic control measures have been proposed and applied *(6, 7)*.

The design and testing of new traffic control strategies requires the existence of accurate traffic flow models that are able to reproduce the motorway traffic conditions with satisfactory accuracy. The macroscopic first-order traffic flow models represent a valuable tool for the study of traffic behavior, as they are simple and effective in reproducing wave formation and propagation under congested conditions. However, their inherent formulation does not allow for capturing traffic phenomena such as the capacity drop and stop-and-go waves. In contrast, second-order models (e.g., *(8–10)*), are able to reproduce traffic instabilities, such as the aforementioned phenomena, but they are characterized by drawbacks such as higher complexity, the specification of parameters without clear physical significance and the higher computation effort that is needed for optimization problems built upon them.

The most common space-time continuous first-order macroscopic model is the Lighthill-Whitham-Richards (LWR) model *(11, 12)*, which is described by a single partial differential equation based on the conservation of vehicles. A significant amount of literature proposes and extends discrete approximations of LWR using the Godunov scheme *(13, 14)*. The most referenced among them is the Cell Transmission Model (CTM) *(15)*, where the flow is defined as a function of density via the definition of a triangular fundamental diagram (FD).

During the last years, attention has been drawn in the direction of including the capacity drop phenomenon into first-order models, since this seems of crucial importance for designing and testing motorway traffic control strategies *(16)*. Some researchers have tried to capture the capacity drop phenomenon based on the "inverse lambda" shaped FD, first proposed in *(17)*, suggesting that the flow-density relation (FD) can be discontinuous, characterized by a sharp flow drop within a small density range. This behavior can be theoretically modelled via definition of two flow values for a specific range of densities around the critical, where the different flows appear depending on the current traffic conditions *(18, 19)*. Other researchers *(20, 21)* propose two-phase traffic flow models, assuming bounded acceleration for certain traffic conditions via building a modified demand function. Following this concept, in *(22)* the authors introduce generalized versions of discrete approximations of the LWR, allowing for a wide range of demand functions to be taken into account for overcritical densities; while in *(23)*, the authors included the capacity drop phenomenon into a multi-lane first-order traffic flow model where the capacity drop is triggered by lateral and on-ramp flows. In *(24)* the authors utilize a macroscopic first-order multilane model, including capacity drop by decreasing the supply function of the cells located downstream of a congested one. In addition, some studies such as *(25)* and *(26)*, describe the capacity drop mechanism as a consequence of microscopic phenomena, such as lane-changing maneuvers, slow vehicles entering a merge cell, and heterogeneous lane behavior due to the variations of traffic states at merges, which prevent the system to reach the full motorway capacity before the breakdown *(27)*. Furthermore, there are also some studies that combine some of the aforementioned concepts such as in *(28)*, where the authors proposed a model which includes two approaches for capacity drop: a reduction of the mainstream demand and the introduction of a weaving parameter.

The rest of the paper is structured as follows: Firstly, the approaches selected for testing are described in more detail, highlighting the necessary modification of a basic discretized-LWR formulation. Later, the aforementioned approaches are tested using real traffic data from a motorway network in UK. Finally, some concluding remarks are provided.

## MODEL FORMULATIONS

For the subsequent description and testing of different capacity-drop strategies, a simple formulation of a discretized-LWR model is utilized. Despite the fact that some of the considered approaches are introduced for more sophisticated models, their implementation is here based on a common formulation, which also permits a clearer understanding and a fairer result comparison. Also the notation (see Table 1) is kept constant throughout the paper for all the described approaches. Moreover, a simple demonstrative example is constructed, which illustrates the qualitative behavior of the different approaches that are tested hereafter.

### Basic Discretized-LWR Formulation

A discretized first-order model considers the discretization of the network in a finite number of cells and the definition of rules for sending and receiving traffic flow. To this end, a motorway stretch is divided into $n$ cells as shown in Figure 1. The $i^{th}$ cell (where $i=1,...,n$) is characterized by a single state variable $\rho_i$, which corresponds to the density of vehicles, namely the number of vehicles divided by the length of the cell, implying that the state of the motorway is entirely described by the $n$-dimensional vector $\rho=\left(\rho_1,...,\rho_n\right)$ evolving according to a $n$-dimensional nonlinear difference equation. The movement of vehicles from one cell to the next is governed by the steady-state relation between flow and density, i.e. the corresponding FD. This relation is characterized by its concave branches, where the demand part and the supply part reflect, respectively, its increasing and decreasing branches *(13, 14)*. The density value (unimodal FDs) or values (multimodal FDs), for which the maximum flow is observed, is defined as the critical density of the cell. The equations considered in our formulation are Equations (1), (2), (3), and (4); while the definition of the model's variables and parameters are given in Table 1.

$$\begin{aligned}
&\rho_1(k+1)=\rho_1(k)+\frac{T}{l_1L_1}\left(-\frac{f_1(\rho_1(k))}{1-p_1}+r_1(k)\right),\\
&\rho_i(k+1)=\rho_i(k)+\frac{T}{l_iL_i}\left(f_{i-1}(\rho_{i-1}(k))-\frac{f_i(\rho_i(k))}{1-p_i}+r_i(k)\right)\ \text{for}\ i=2,...,n,
\end{aligned}\tag{1}$$

$$f_i(\rho_i(k))=\min\left\{f_{i,D}(\rho_i(k)),f_{i+1,S}(\rho_{i+1}(k))\right\}\ \text{for}\ i=1,...,n-1\,,\ f_n(\rho_n(k))=f_{n,D}(\rho_n(k))\,,\tag{2}$$

$$f_{i,D}(\rho_i(k))=\min\{g_i(\rho_i(k)),Q_i\}(1-p_i)\ \text{for}\ i=1,...,n-1\,,\ f_{n,D}(\rho_n(k))=\min\{g_n(\rho_n(k)),Q_n\}\tag{3}$$

$$f_{i+1,S}(\rho_{i+1}(k))=\min\left\{Q_{i+1},w_{i+1}\left(\rho_{\max,i+1}-\rho_{i+1}(k)\right)l_{i+1}\right\}-r_{i+1}(k)\ \text{for}\ i=1,...,n-1\,.\tag{4}$$

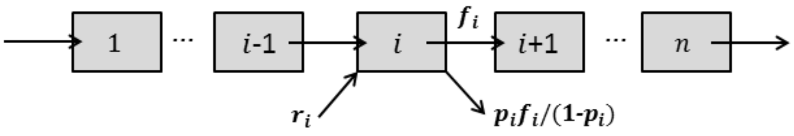

**FIGURE 1 The discretization of motorway network**

**TABLE 1 Models' variables and parameters**

| Symbol | Name | Units |
|---|---|---|
| $T$ | simulation time step | h |
| $l_i$ | number of lanes in the $i^{th}$ cell | dimensionless |
| $L_i$ | length of the $i^{th}$ cell | km |
| $\rho_i(k)$ | density of the $i^{th}$ cell at time $t = kT$ , $k = 0,1,2,...$ | veh/km/lane |
| $\rho_{\max,i}$ | storage capacity of the $i^{th}$ cell | veh/km/lane |
| $\rho_{cr,i}$ | critical density of the $i^{th}$ cell | veh/km/lane |
| $f_i(\rho_i(k))$ | actual outflow from the $i^{th}$ to $(i+1)^{th}$ cell in $(kT,(k+1)T]$, $k = 0,1,2,...$ | veh/h |
| $f_{i,D}(\rho_i(k))$ | demand part of the FD of the $i^{th}$ cell | veh/h |
| $f_{i,S}(\rho_i(k))$ | supply part of the FD of the $i^{th}$ cell | veh/h |
| $Q_i$ | capacity flow of the $i^{th}$ cell | veh/h |
| $r_i(k)$ | demand on-ramp flow of the $i^{th}$ cell in $(kT,(k+1)T]$, $k = 0,1,2,...$ | veh/h |
| $p_i$ | percentage of the actual flow exiting from the off-ramp of the $i^{th}$ cell (exit-rate of the $i^{th}$ cell) | dimensionless |
| $v_{f,i}$ | free-flow speed of the $i^{th}$ cell | km/h |
| $w_i$ | congestion wave speed of the $i^{th}$ cell | km/h |

Notice that, according to this formulation, it is assumed that vehicles entering from on-ramps have full priority. That is, the flow from an on-ramp always enters the cell (given that the maximum density is not exceeded) while only the remaining supply space can be occupied by the flow received from the upstream cell. Although this assumption may not be always correct, it helps significantly the calibration procedure, since boundary conditions are given, and therefore the actual on-ramp flow is allowed to feed directly the model. Note also, that for the 1st cell, the entering mainstream flow plus any possible on-ramp flow are considered within $r_1$; which can be fully accommodated by the same cell (no supply term); this implies that any appearing congestion in the stretch never reaches the upstream boundary.

Finally, notice that in Equations (3) and (4), the demand and supply functions, respectively, are completed by assuming capacity flow values $Q_i$ for overcritical and undercritical densities, respectively. Thus, this model predicts capacity flow (no capacity drop) for discharge flows, in accordance with the non-discretised LWR model. Note also, that the right-hand side of the FD of the $i^{th}$ cell in Equation (4) is assumed to be linear (with a negative slope $w_i$); while the left-hand side of the FD in Equation (3) is assumed to be a non-decreasing function $g_i(\rho_i)$.

## Different Shapes for the FD

Different functions $g_i(\rho_i)$ can be used within the demand function in Equation (3). The CTM formulation *(15)* considers a triangular-shaped FD (Figure 2(a)), where $g_i(\rho_i) = v_{f,i}\rho_i l_i$ , $g_i(\rho_{cr,i}) = Q_i$ and $w_i = Q_i \big/ \big((\rho_{\max,i} - \rho_{cr,i}) l_i\big)$. This formulation has two main drawbacks: first, when using realistic free-flow and congestion-wave speeds, it leads to high (and sometimes unrealistic) capacity flow; second, only one speed value is considered for all under-critical densities, which is often not compatible with traffic observations. To overcome the first issue, a trapezoidal FD can be used, where $g_i(\rho_i) = v_{f,i}\rho_i l_i$ , $g_i(\rho_{cr,i}) \geq Q_i$ and $w_i \geq Q_i \big/ \big((\rho_{\max,i} - \rho_{cr,i}) l_i\big)$, as illustrated in Figure 2(b). In this case, the critical density, instead of being a fixed point for both the FD parts, can be selected within an interval of densities, increasing also the degree of freedom for model calibration. Nevertheless, in real traffic, the observed speed may be characterized by a decreasing-slope behavior also for low densities, which can be reflected by using a nonlinear concave function $g_i$ (Figure 2(d)), where $g_i(\rho_{cr,i}) = Q_i$ and $w_i = Q_i / \left((\rho_{\max,i} - \rho_{cr,i}) l_i\right)$. An opportune calibration of such function may lead to more realistic results. As an example, a nonlinear exponential function, as proposed in *(10)*, can be employed. A similar behavior can also be obtained, without much loss of accuracy, considering a piecewise-linear approximation of the nonlinear function (Figure 2(c)), which is helpful in case linear constraints are needed for the formulation of an optimization problem (e.g., *(23, 29)*).

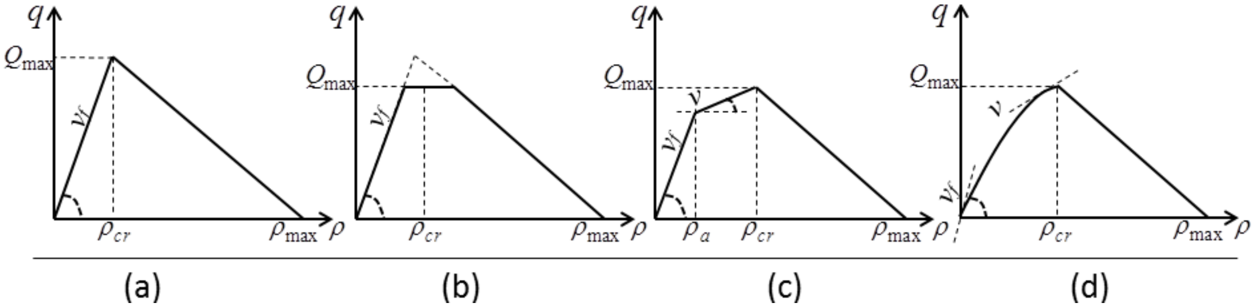

**FIGURE 2 Different choices for the left-hand side of the fundamental diagram corresponding to: (a) a triangular FD (CTM), (b) a trapezoidal FD, (c) a piecewise linear FD and (d) a nonlinear FD.**

## Demonstrative Example

In order to illustrate the behavior of each approach, a simple hypothetical motorway stretch is considered, consisting of a set of $n = 15$ homogeneous cells of equal length $L_i = 0.5$ km, 3 lanes ($l_i = 3$) and common FD parameters. The motorway stretch includes one on-ramp which is located at the upstream boundary of the cell $i = 13$. Furthermore, the FD parameters are set to be $\rho_{cr,i} = 20$ veh/km/lane, $v_{f,i} = 100$ km/h, $\rho_{\max,i} = 120$ veh/km/lane, $w_i = 20$ km/h, and $Q_i = 6000$ veh/h. The simulation time step is set to be $T = 5$ s, while the simulation time horizon is $T^{hor} = 4$ h for all the following tests. For the sake of simplicity, in all the following tests, the function $g_i$ utilized in the Equation (3) is selected to be $g_i(\rho_i) = v_{f,i} \rho_i l_i$ (for $i = 1,...,15$). Boundary conditions are set so that the exit can accommodate the receiving flow from the last cell (infinite supply). A hypothetical trapezoidal traffic demand scenario is applied to the network which, for some time period, exceeds the capacity of the merge area, generating congestion that spills back for several kilometers, without though reaching the network origin. Specifically, the mainstream traffic demand and the demand from the on-ramp are equal to $r_1 = 3500$ veh/h and $r_{13} = 500$ veh/h, respectively, for [0-0.5] hours and for [2.5-4] hours; after 0.5 hours, both demand flows start to increase linearly and then are constant and equal to $r_1 = 4500$ veh/h and $r_{13} = 1600$ veh/h, respectively, for [1-2] hours; while after 2 hours, they start again to decrease linearly for [2-2.5]hours. The initial state for every cell is set $\rho_i(0) = 11.7$ veh/km/lane for $i = 1,..,12$ and $\rho_i(0) = 13.3$ veh/km/lane for $i = 13,14,15$.

Figure 3 (a), (b), (c) illustrates some significant characteristics that appear in the well-known behavior of the CTM in case congestion is created at an on-ramp merge. Once the total demand flow (in this case, the sum of mainstream and ramp flows) exceeds the bottleneck's capacity, only a portion of the available mainstream flow is allowed to access the 13th cell, since full priority is given to the on-ramp flow. This causes an increase of density at the upstream cell (the 12th cell) (see Figure 3(a), red line), which eventually enters into a congested state, generating a congestion wave. During this period, the density in the merge cell remains at its critical value (see Figure 3(a), blue line), allowing an exit flow equal to the capacity (Figure 3(c)). As a consequence of the observations above, the speed at the 12th cell decreases (Figure 3(b), red line), while the speed at the merge cell (13th) remains constant and equal to the free speed. Note that, in contrast to this modelled behavior, the merge cell is typically congested in real traffic; while the exit flow is reduced upon the onset of congestion due to capacity drop.

In the following subsections, the selected approaches for including capacity drop are described.

## Approach 1: Switching Logic for Maximum Flow

One effective way to implement a FD characterized by the inverse-lambda shape is via the definition of an opportune switching logic to define the current maximum flow. An example can be found in *(18)*, where the authors proposed a set of rules to impose capacity drop in case VSL (variable speed limits) are applied in a certain area of the network. The concept is based on the coexistence of two FDs for the same location: a triangular-shaped one, active in case VSL are not applied (and no congestion is present); and a trapezoidal-shaped one, characterized by a lower capacity that materializes in case congestion is present. This method can be extended straightforwardly to the case of bottlenecks due to lane drops, tunnels, etc.; in addition, we show here that it is also effective in case congestion is generated because of a merging on-ramp.

The formulation is described by Equations (1), (2), (5), and (6) under Equations (7) and (8):

$$f_{i,D}(\rho_i(k)) = \min\left\{R_i(k), g(\rho_i(k))\right\}(1-p_i) \quad \text{for } i=1,\dots,n\,, \tag{5}$$

$$f_{i+1,S}(\rho_{i+1}(k)) = \min\left\{R_{i+1}(k), w_{i+1}\left(\rho_{\max,i+1}-\rho_{i+1}(k)\right)l_{i+1}\right\} - r_{i+1}(k) \quad \text{for } i=1,\dots,n-1\,, \tag{6}$$

where

$$R_1(k+1)=Q_1,\; R_2(k+1)=Q_2, \tag{7}$$

$$R_{i+1}(k+1)=\begin{cases}\bar{Q}_{i+1} & \text{if } \; w_i(\rho_{\max,i}-\rho_i(k))l_i < \min(f_{i-1,D}(\rho_{i-1}(k)), R_i(k)) \\ Q_{i+1} & o.w.\end{cases} \quad \text{for } i=2,\dots,n-1\,, \tag{8}$$

where $R_i$ are auxiliary variables that define the maximum flow for cell $i$ and $\bar{Q}_i$ is the queue discharge flow observed after the congestion onset. The queue discharge flow $\bar{Q}_i$, can also be viewed as $\bar{Q}_i = \alpha Q_i$, i.e., a portion $\alpha < 1$ of the capacity flow. For this simulation test, this portion is constant and equal to $\alpha = 0.95$. Equation (7) reflects the assumption that the spilling-back congestion does not reach the entrance of the network. Moreover, all cells are initially uncongested, thus $R_i(0) = Q_i$, for every $i = 1,\dots,n$.

Figure 3 (d), (e), (f), illustrates the behavior resulting from the application of this approach. The main idea lies in decreasing the capacity of the cell located immediately downstream of a congested one. More specifically, when the aggregated flow from the on-ramp and the mainstream exceeds the capacity of the merge cell, the density of the upstream cell starts increasing (see Figure 3(d), red line) while at the same time its speed starts decreasing (see Figure 3(e), red line); consequently, after some time, its supply function becomes smaller than the demand function of the upstream cell; this, according with Equation (8), triggers a reduction of the maximum flow for the downstream cell (see Figure 3(f)), which persists until the overall demand is sufficiently decreased. As a possible drawback, the flow reduction appears with some delay after the congestion starts; this is because this reduction materializes only when both the demand flow of the 11th cell and the maximum flow of the 12th cell become higher than the supply of the 12th cell. Furthermore, it is interesting to point out that despite the flow-drop, there is no congestion, i.e. over-critical density (Figure 3(d), blue line) and therefore also no speed-drop (Figure 3(e), blue line), at the merge cell.

### Approach 2: Introduction of a Weaving Parameter

Another option to achieve a reduced outflow at a merge cell is via the introduction of a weaving parameter that affects the supply function at the merge cell, as proposed in *(28)*. The purpose of this parameter is to take into account the "intensity" of lane changing maneuvers performed by vehicles just entered from the on-ramp, imposing a reduction of the available space for vehicles coming from upstream. The mathematical formulation consists of Equations (1), (2), (3), and (9):

$$f_{i+1,S}(\rho_{i+1}(k)) = \min\left\{Q_{i+1}, w_{i+1}\left(\rho_{\max,i+1} - \rho_{i+1}(k)\right)l_{i+1}\right\} - \eta_{i+1}^{r} r_{i+1}(k) \text{ for } i = 1,...,n-1 \tag{9}$$

where $\eta_{i+1}^{r} \geq 1$ is the weaving parameter which aims to further reduce the supply of the merge cell, in order to reduce the mainstream flow attempting to enter. We can see from Figure 3(i) (where $\eta_{13}^{r} = 1.2$ is used), that the capacity flow is never reached even in case of low on-ramp demand; for this reason, the flow reduction is barely visible. Notice also, that the merge cell is again not congested (Figure 3(g),(h), blue line).

### Approach 3: Reduction of the Demand Function

Another way to incorporate capacity drop, also utilized in *(28)*, consists of the definition of a discontinuous demand part of the FD at bottleneck locations and the implementation of a switching rule to determine which value of maximum flow should be used. In particular, the model can be described by Equations (1), (2), (4), and (10):

$$f_{i,D}(\rho_i(k)) = (1-p_i)\begin{cases} g_i(\rho_i(k)) & if\ \rho_i(k) \leq \rho_{cr,i} \\ \bar{Q}_i & o.w. \end{cases} \text{ for } i = 1,...,n\ . \tag{10}$$

This approach generates the same values as the basic LWR when the density in the 12th cell is undercritical (Figure 3(j)), reaching properly capacity flow; then, for overcritical densities, the flow drops to a value corresponding to $\bar{Q}_i + r_i(k)$; in this case, $\bar{Q}_i = \alpha Q_i$, and $\alpha = 0.7$ are used (Figure 3(l)). As a main drawback, traffic congestion persists longer than in the other cases, because, once formed, its disappearance can only be triggered by a decrease of the arriving demand below $\bar{Q}_i$, irrespectively of the ramp variations. Again, no congestion appears at the merge cell (Figure 3(k),(j), blue line).

### Approach 4: Linear Reduction of Maximum Flow

As mentioned earlier, the presence of capacity drop within traffic flow models plays a key role for the design and testing of motorway traffic control strategies. Among others, model-based control problems have been widely exploited in recent years because of the possibility to explicitly consider the system dynamics and physical constraints. In some works, the classic formulation of first-order models was implemented via use of integer variables and opportune switching rules; see e.g. *(28, 30, 31)*. In other works, e.g. in *(29, 32)*, linear inequalities (derived from the piecewise linear FD) were considered as constraints in the optimization problem; hereafter some modification of these models are presented, which allow to define linearly constrained formulations for corresponding optimization problems. A similar model is proposed in *(33)*.

The same concept as in Approach 1 can be used, albeit with the introduction of an additional linear term that reduces the supply function of a downstream cell. Specifically, when congestion starts in the cell $i$ ($\rho_i > \rho_{cr,i}$), the supply term of the downstream cell $i+1$ is linearly decreased as a function of $\rho_i$, according to the following equations:

$$f_{i+1,S}(\rho_{i+1}(k)) = \min\left\{F_{i+1}(\rho_i(k)), w_{i+1}\left(\rho_{\max,i+1} - \rho_{i+1}(k)\right)l_{i+1}\right\} - r_{i+1}(k) \text{ for } i = 1,\ldots,n-1\,, \quad (11)$$

where $F_{i+1}(\rho_i(k))$ is given by:

$$F_{i+1}(\rho_i(k)) = \begin{cases} Q_{i+1} & \text{if } \rho_i(k) \le \rho_{cr,i} \\ \bar{Q}_{i+1} + \dfrac{Q_{i+1} - \bar{Q}_{i+1}}{\rho_{cr,i} - \rho_{\max,i}}\left(\rho_i(k) - \rho_{\max,i}\right) & o.w. \end{cases} \text{ for } i = 1,\ldots,n-1 \quad (12)$$

with $\bar{Q}_i = \alpha Q_i$, and $\alpha = 0.9$. The proposed formulation is thus given by Equations (1), (2), (3) and (11) under Equation (12). For under-critical densities, $F_i$ is constant and equal to the capacity flow; however, in case the density of the $i^{th}$ cell increases beyond its critical value (Figure 3(m), red line), the maximum flow of the supply function of the $(i+1)^{th}$ cell is reduced linearly (Figure 3(o)). This approach appears to work appropriately also for bottlenecks due to lane drops, tunnels etc.; the capacity flow is reached before dropping, in contrast with Approach 2; furthermore, the capacity drop appears with a lower delay than the one observed in Approach 1. On the other hand, the merge segment is seen to remain uncongested as with all previous approaches as well.

## Approach 5: Increased Space for Vehicles Entering a Bottleneck Location.

Yet another approach may be conceived, which, in contrast to all previous approaches, allows for the bottleneck (e.g. merge) cell to get congested, as in real traffic. To achieve this, the bottleneck cell must be enabled to temporarily receive more flow than the cell's capacity parameter would allow. To this end, the supply function is modified, compared with the basic approach, in two ways. First, the upper bound (capacity) used in Equation (4) for undercritical densities is increased, e.g. by 5%; second, the wave speed is set slightly (e.g. 5%) higher than usual; thus, the merge cell can temporarily accommodate more inflow than what it can sent downstream. In addition, in order to also enable capacity drop, a linearly decreasing demand part for overcritical densities (similar to the one proposed in *(20, 21)*), is introduced. These changes may be applied either to bottleneck cells only or to all cells (as in the current test example), leading to very similar results. In Figure 3, this approach causes indeed a density increase in the merge cell whenever capacity is exceeded (Figure 3(p), blue line). This increase generates consequently a reduction of the cell outflow (Figure 3(r)), i.e. capacity drop, accompanied also by a reduction of speed (Figure 3(q), blue line). The resulting equations, that replace Equations (3) and (4), are Equations (13) and (14) respectively:

$$f_{i,D}(\rho_i(k)) = (1-p_i)\begin{cases} g_i(\rho_i(k)) & \text{if } \rho_i(k) \le \rho_{cr,i} \\ Q_i + \bar{Q}_i \dfrac{\rho_i(k) - \rho_{cr,i}}{\rho_{cr,i} - \rho_{\max,i}} & o.w. \end{cases} \text{ for } i = 1,\ldots,n\,, \quad (13)$$

$$f_{i+1,S}(\rho_{i+1}(k)) = \min\left(Q'_{i+1}, w'_{i+1}\left(\rho_{i+1,\max} - \rho_{i+1}(k)\right)\right) - r_{i+1}(k) \text{ for } i = 1,\ldots,n-1\,, \quad (14)$$

where $Q'_i > Q_i$ and $w'_i > w_i$. For this test, we selected $\bar{Q}_i = \alpha Q_i$ with $\alpha = 0.4$ while $Q'_i = 1.05Q_i = 6300$ veh/h and $w'_i = 1.05w_i = 21$ km/h for every $i = 1,\ldots,15$. It is interesting to point out that this approach, despite not being a direct derivation of the LWR model, still guarantees the conservation of vehicles; and furthermore, the congestion is first created at the merge cell and the flow drop occurs immediately after the maximum flow is reached, in accordance with real traffic observations.

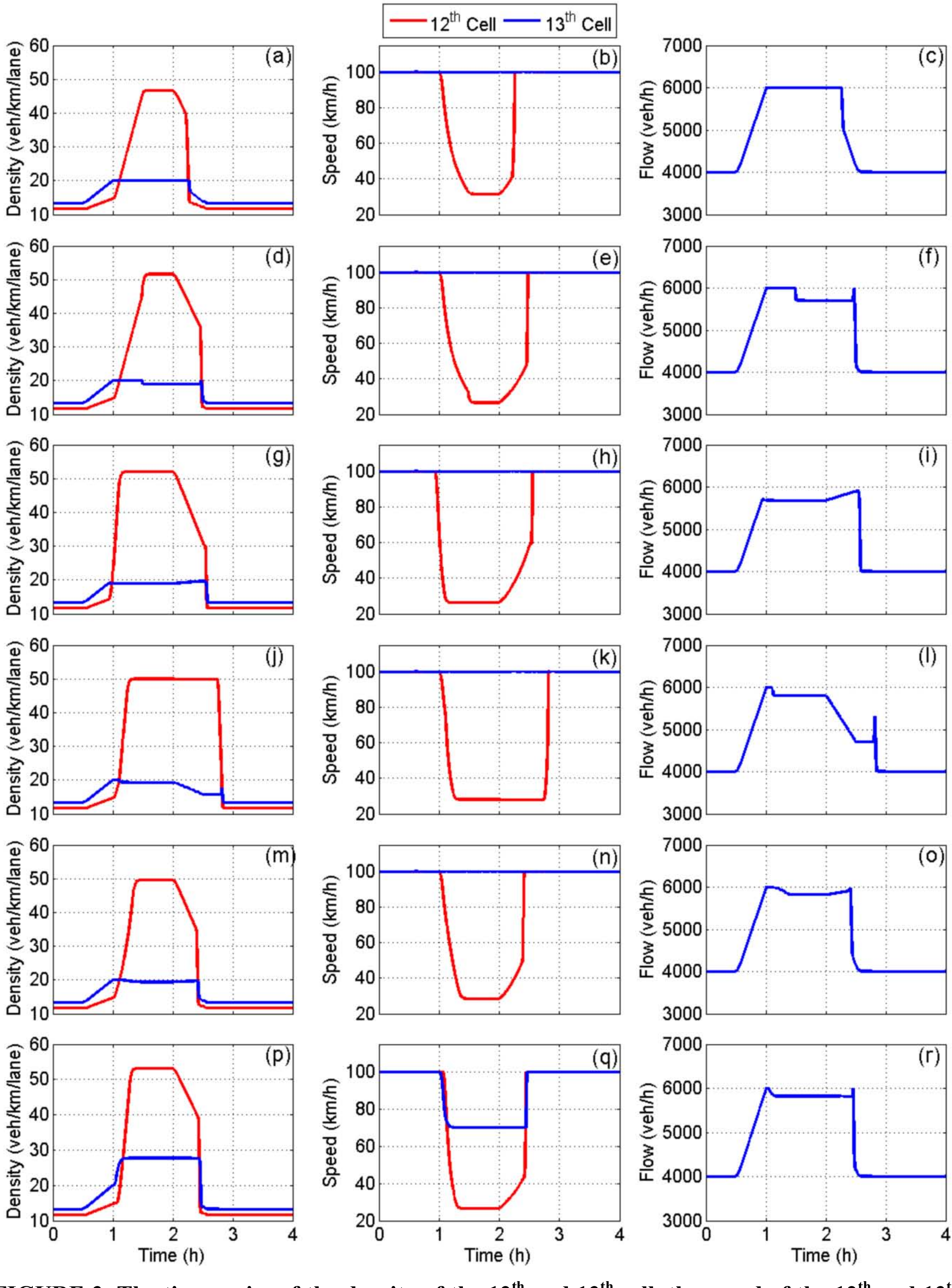

**FIGURE 3 The time-series of the density of the 12th and 13th cell, the speed of the 12th and 13th cell and the outflow from the 13th for the application of (a),(b),(c) CTM, (d),(e)(f) Approach 1, (g),(h),(i) Approach 2, (j),(k),(l) Approach 3, (m),(n),(o) Approach 4, (p),(q),(r) Approach 5.**

## CALIBRATION RESULTS

In this section the described approaches of the previous section are validated and compared regarding the representation of traffic conditions in a real motorway stretch with particular emphasis on the reproduction of the capacity drop phenomenon.

### Motorway Network and Calibration Set-up

The considered motorway stretch of 9.45 km in length is a part of the M56 motorway in the United Kingdom, direction from Chester to Manchester. This 3-lane motorway stretch includes one off-ramp and a two-lane on-ramp, which, before reaching the motorway, is divided into two separate lanes. The corresponding on-ramp flows of each lane enter the motorway at two different locations, is shown in Figure 4. In order to model the network by use of the selected traffic flow models, the examined motorway stretch is divided into 7 links (Figure 4) and each motorway link is subdivided in model cells of equal length (about 250 m each). Using this representation, the motorway cells are well-defined, and the model equations presented in the previous section, are directly applicable. Figure 4 displays the length of each link, the location of the on-ramps and off-ramp and the locations of the available detector stations.

The real traffic data used in this study were obtained from MIDAS database *(34)*. The traffic data includes flow and speed measurements at each detector location, with a time resolution of 60 s. The traffic data analysis showed that, within this motorway stretch, recurrent congestion is created during the morning peak hours due to the high on-ramp flow. In particular, Figure 5(a) displays the space-time diagram of the real speed measurements for 03/06/2014. It is observed that congestion is created upstream of the second on-ramp during 7–8 a.m. which spills back several kilometers. Moreover, downstream of the second on-ramp, the vehicles accelerate as they exit the congestion area. Figure 6 presents the time-series of the flow measurements (black line) from the detector station D 8180 which is located downstream of the congestion creation area (see also Figure 4). It is observed that the capacity drop is present here, as the merge area outflow drops visibly when congestion sets in (between 7:10 a.m. and 8:10 a.m.).

In order to apply the examined models to this motorway stretch and achieve a fair comparison, it is important to first calibrate the models using real traffic data. The model calibration procedure aims to appropriately specify the model parameter values, so that the representation of the network traffic conditions is as accurate as the model structure allows. This can be achieved by employing a suitable optimization methodology which aims at minimizing the discrepancy between the model estimations and the real traffic data. For more details on the considered model calibration procedure see *(35)*.

In the current study, the Nelder-Mead optimization method is employed for the calibration of the examined traffic flow models. The models are fed with boundary data (upstream boundary and ramps) and produce the stretch-internal traffic state according to the respective equations and parameter values. The utilized performance index (PI) under minimization is the root-mean-square error (RMSE) of the real versus the model-predicted speed values at all detector locations. The models are calibrated using real traffic data from 03/06/2014 and a simulation time step equal to $T = 5$ s. It should be stressed that all cells of the modelled motorway stretch are characterized by the same parameters of the FD. After the calibration procedure, the accuracy and robustness of the resulted models should be tested. This is achieved by the model validation, where the produced models are applied using different traffic data (from the same motorway site) than those used for their calibration. In this study, the models are validated using real traffic

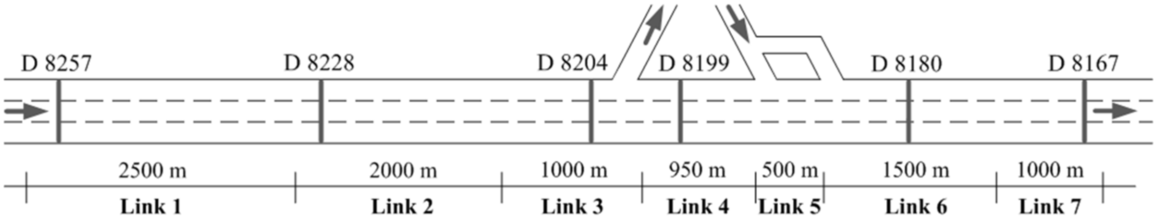

**FIGURE 4 Representation of the considered freeway stretch.**

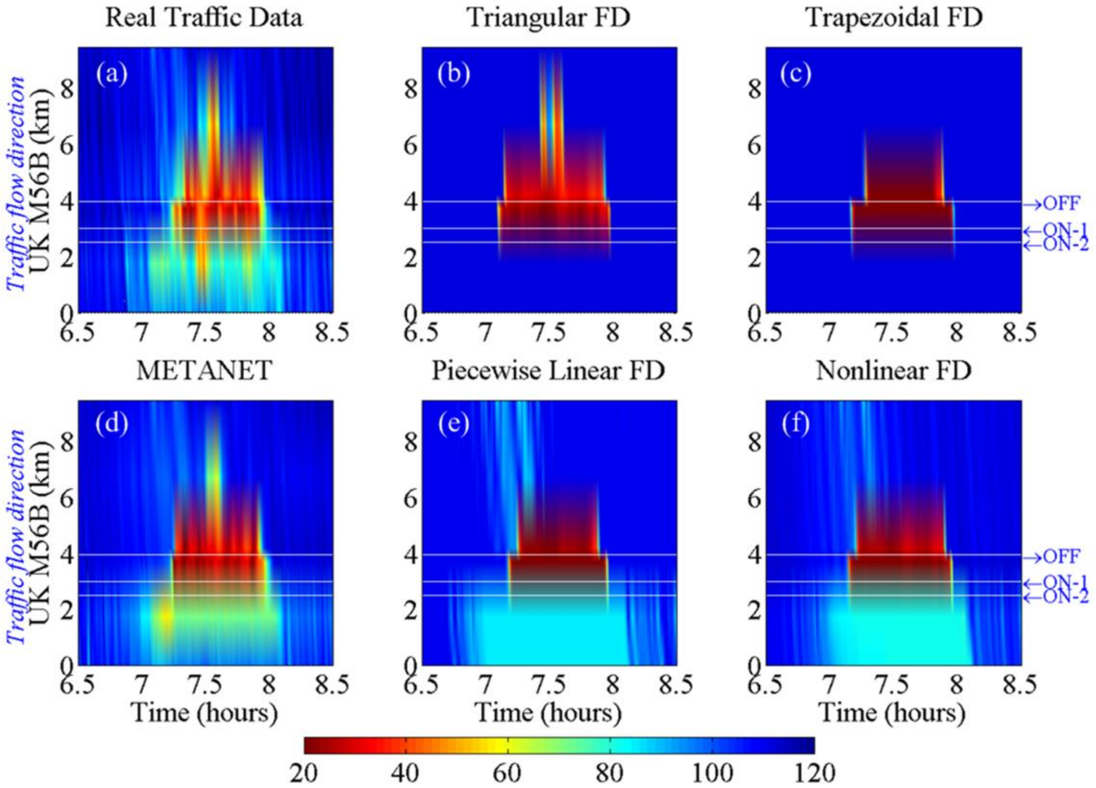

**FIGURE 5 Space-time diagrams of the real speed measurements and the models' estimations of speed for 03/06/2014.**

data from 24/06/2014.

## Basic Discretized-LWR Formulation

As mentioned before, the investigated capacity drop approaches are based on a simple first-order model. This basic first-order model cannot reflect the capacity drop phenomenon; however, different expressions for the FD may improve the model's accuracy. To this end, four different shapes of the FD were examined first, all applied to the basic model, i.e. triangular FD, trapezoidal FD, piecewise linear FD, and nonlinear exponential FD (see Figure 2).

Table 2 includes the estimated models' parameter values and the corresponding PI value for the calibration and the validation date. It is interesting to see that all four variations estimated similar value for the $Q$ (capacity) parameter. Moreover, as it was expected, the use of a triangular FD results in a low $\rho_{cr}$ value, lower than in the other formulations. Table 2 also includes the calibration results of a second-order model, METANET *(10)*, which was applied to this motorway stretch for comparison purposes. Note that Table 2 presents only some of the estimated METANET parameter values, due to the limited available space, while the rest parameters were estimated equal to: $\tau$=26.8 s, $\nu$=45.6 km$^2$/h, $\delta$=0.1 h/km, $\kappa$=10 veh/km/lane, $v_{min}$=7 km/h. Figure 5 illustrates the space-time diagrams of the real speed measurements and the corresponding model predictions of speed for the calibration date. It is observed that the models using a triangular or a trapezoidal FD predict free flow conditions at all areas outside congestion. In contrast, the use of a piecewise linear or non-linear FD allows for mean speed variations, also outside of the congestion area, thus achieving higher accuracy there, compared to the first two formulations. The second-order model METANET, produces, as expected, a more realistic representation of the prevailing traffic conditions thanks to the fact that this model takes into account the vehicle acceleration and the drivers reaction time. Considering the above results, first-order models with nonlinear FD are used for the subsequent investigations of capacity drop approaches.

## Capacity Drop Approaches

Five capacity drop approaches are implemented for this simple, but typical, motorway stretch, which were described in the previous section. Table 2 includes the estimated model parameter values for all five approaches. It should be mentioned that in all examined approaches the maximum capacity flow $Q$ was considered fixed and equal to 6900 veh/h, which is close to the highest flows observed in the network. This was done in order to achieve a fair comparison of the models regarding the reproduction of the capacity drop phenomenon.

Table 2 shows that in all five approaches similar values were estimated for the $v_f$ and $\rho_{cr}$ parameters, while quite different values were obtained for the parameters $w$ and $\rho_{\max}$ due to the different formulations adopted for the reproduction of the capacity drop phenomenon. Moreover, it should further be noted that although in all approaches the parameter $\alpha$ is related to the magnitude of the capacity drop, it is actually introduced to the model equations in a different way and for this reason the value of $\alpha$ varies in the different approaches.

Figure 6 (a)-(f) display the time-series of the real flow measurements and the corresponding model estimations, at the location of detector station D 8180, for all tested approaches. It is observed that, in contrast to the simple first-order model, all five approaches are able to reproduce the capacity drop, resulting in reduced merge area outflow during the congestion period, in accordance also with simulation results. Table 2 also includes the PI value for the calibration and the validation days. It is observed that the models achieve similar PI values, which implies that they are all able to reproduce the traffic conditions in this network with sufficient accuracy. Moreover, all approaches improve the PI value compared to the simple first-order model with non-linear FD.

Nevertheless, each approach is characterized by a different qualitative behavior. In particular, the discharge flow observed at detector D8180 for Approach 1 corresponds exactly to the pre-specified value $\bar{Q} = aQ = 6141$ veh/h. In Approach 2, the observed outflow never reaches capacity, even before the onset of congestion, in accordance with the behavior described in the previous section. For Approach 3, the discharge flow that materializes is also dependent on the on-ramp flow entering the merge cell (which causes the fluctuations that can be observed in the corresponding plot), whereas the mainstream flow exiting the cell upstream of the merge cell is constantly equal to $\bar{Q} = aQ = 4968$ veh/h. In Approach 4, the magnitude of the observed capacity drop increases according to the density of the upstream cell; i.e., in case

**TABLE 2 Calibration and validation results for all employed models.**

| *Model* | $v_f$ (*km/h*) | $\rho_{cr}$ (*veh/km/ lane*) | *w* (*km/h*) | $\rho_{max}$ (*veh/km/ lane*) | *Q* (*veh/h*) | $\rho_\alpha$ (*veh/km/ lane*) | *α* | *η* | *w'* (*km/h*) | *PI 3/6* | *PI 24/6* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Trian. FD** | 112.0 | 18.7 | 21.2 | 117.4 | 6282 | - | - | - | - | 18.0 | 19.3 |
| **Trap. FD** | 112.0 | - | 21.8 | 145.0 | 6192 | - | - | - | - | 18.0 | 16.9 |
| **PWL FD** | 110.5 | 24.7 | 14.8 | 165.5 | 6258 | 14.4 | - | - | - | 12.6 | 13.1 |
| **NL FD** | 114.2 | 26.0 | 19.4 | 133.9 | 6285 | - | - | - | - | 12.9 | 13.5 |
| **METANET** | 114.2 | 28.9 | - | - | 6525 | - | - | - | - | 7.9 | 11.5 |
| **Approach 1** | 123.2 | 36.4 | 25.9 | 124.9 | 6900 | - | 0.89 | - | - | 11.6 | 12.9 |
| **Approach 2** | 123.4 | 35.9 | 21.9 | 139.2 | 6900 | - | - | 1.56 | - | 11.9 | 13.7 |
| **Approach 3** | 122.8 | 33.5 | 21.4 | 149.1 | 6900 | - | 0.72 | - | - | 12.8 | 13.1 |
| **Approach 4** | 123.0 | 35.6 | 25.9 | 131.2 | 6900 | - | 0.63 | - | - | 11.3 | 13.1 |
| **Approach 5** | 122.8 | 35.7 | 33.6 | 106.7 | 6900 | - | 0.57 | - | 38.2 | 11.5 | 12.7 |

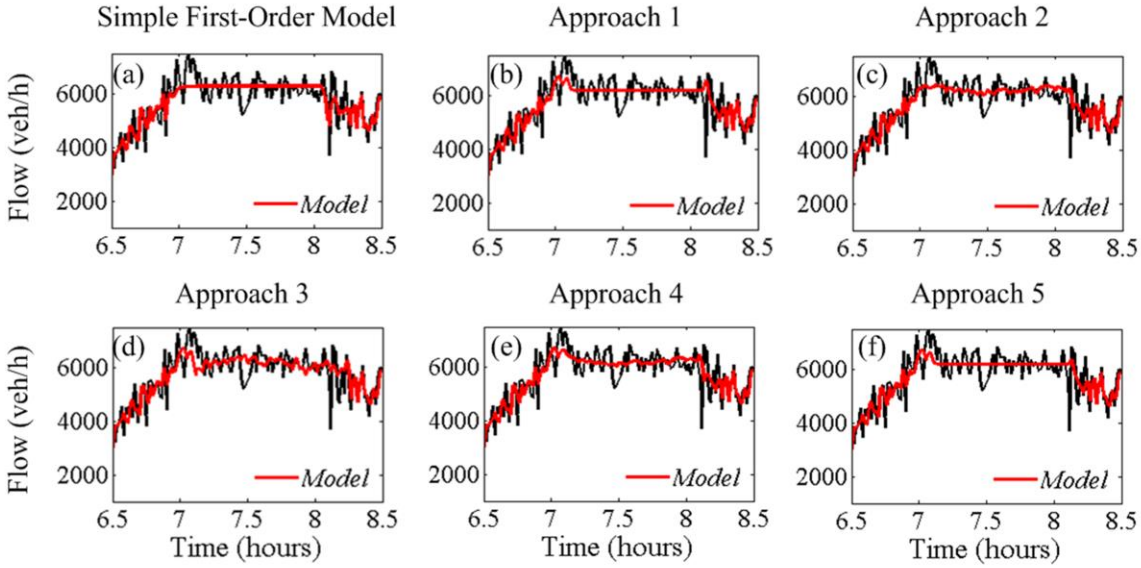

**FIGURE 6 Time-series of the real flow measurements at the location of detector station D 8180 for (a) the simple first-order model; (b) the Approach 1; (c) the Approach 2; (d) the Approach 3; (e) the Approach 4; and (f) the Approach 5 for 03/06/2014.**

of stronger congestion characterized by a lower internal speed, a stronger capacity drop is observed, which is in accordance with some traffic observations. Finally, in Approach 5, the discharge flow is the equilibrium resulting from the combined effect of the demand and supply functions for the merge cell. It should further be noted that all approaches, except for Approach 2, are capable to estimate a high merge area outflow just before the onset of congestion, which is in accordance with the real flow values observed.

Regarding the calibration procedure, Approaches 1, 2 and 4 were found to converge to similar optimal parameter valus, reflecting a capacity drop, when started with different initial parameter values. As a consequence, these models might be more easily applicable to different application scenarios. On the other hand, for Approaches 3 and 5 the selection of appropriate parameters was more challenging; in particular, several calibration runs starting from different initial parameter values led sometimes to solutions with a less pronounced capacity drop. These sensitivity and robustness issues are currently under more focussed investigation, and related findings may shed more light to the particularities of each approach.

## CONCLUSIONS

This study presents an overview of modelling approaches to include capacity drop into first-order traffic flow models. The presented approaches were tested first for a hypothetical network and traffic demand scenario to highlight their principal behaviour and qualitative properties; eventually the models were more rigorously calibrated and validated using real-data from a motorway in the U.K. The obtained results show that, although the tested models employ different mechanisms to include the capacity drop phenomenon, they are all able to produce an appropriate flow reduction at the merge area whenever traffic congestion is present. Furthermore, it is important to point out that Approaches 1, 3, 4, and 5 can be implemented in a similar way to other types of bottlenecks (e.g. due to lane-drop, tunnel etc.), while Approach 2, in its current form, can only be applied to merging bottlenecks because of its inherent formulation.

The obtained results were found to be quantitatively similar with respect to the achieved PI values, which is mainly attributed to a traffic situation with limited complexity and a limited number of internal comparison data. Future investigations involving more complex traffic situations and richer data might shed more light on the comparative quantitative accuracy of different approaches.

## ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. [321132], project TRAMAN21. The authors are grateful to Highways England for providing access to the MIDAS data.

## REFERENCES

1. Hall, F. L., and K. Agyemang-Duah. Freeway Capacity Drop and the Definition of Capacity. *Transportation Research Record*, No. 1320, 1991.
2. Banks, J. H. The Two-Capacity Phenomenon: Some Theoretical Issues. *Transportation Research Record*, No. 1320, 1991.
3. Chung, K., J. Rudjanakanoknad, and M. J. Cassidy. Relation between Traffic Density and Capacity Drop at Three Freeway Bottlenecks. *Transportation Research Part B: Methodological*, Vol. 41, No. 1, 2007, pp. 82–95.
4. Dixon, K., J. Hummer, and A. Lorscheider. Capacity for North Carolina Freeway Work Zones. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1529, 1996, pp. 27–34.
5. Smith, B. L., L. Qin, and R. Venkatanarayana. Characterization of Freeway Capacity Reduction Resulting from Traffic Accidents. *Journal of transportation engineering*, Vol. 129, No. 4, 2003, pp. 362–368.
6. Papageorgiou, M., H. Hadj-Salem, and J.-M. Blosseville. ALINEA: A Local Feedback Control Law for On-Ramp Metering. *Transportation Research Record*, No. 1320, 1991.
7. Cassidy, M. J., and J. Rudjanakanoknad. Increasing the Capacity of an Isolated Merge by Metering its On-Ramp. *Transportation Research Part B: Methodological*, Vol. 39, No. 10, 2005, pp. 896–913.
8. Payne, H. J. Models of Freeway Traffic and Control. *Mathematical Models of Public Systems*, 1971.
9. Whitham, G. B. *Linear and Nonlinear Waves*. John Wiley & Sons, 2011.
10. Messmer, A., and M. Papageorgiou. METANET: A Macroscopic Simulation Program for Motorway Networks. *Traffic engineering & control*, Vol. 31, No. 8-9, 1990, pp. 466–470.
11. Lighthill, M. J., and G. B. Whitham. On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Vol. 229, No. 1178, 1955, pp. 317–345.
12. Richards, P. I. Shock Waves on the Highway. *Operations Research*, Vol. 4, No. 1, 1956, pp. 42–51.
13. Godunov, S. A Difference Method for Numerical Calculation of Discontinuous Solutions of Hydrodynamic Equations. *Matematicheskii Sbornik,* Vol. 89, 1959, pp. 271–306.
14. Lebacque, J. P. The Godunov Scheme and What It Means for First Order Traffic Flow Models. *Presented at the International symposium on transportation and traffic theory*, 1996.
15. Daganzo, C. F. The Cell Transmission Model: A Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory. *Transportation Research Part B: Methodological*, Vol. 28, No. 4, 1994, pp. 269–287.
16. Papageorgiou, M. Some Remarks on Macroscopic Traffic Flow Modelling. *Transportation Research Part A: Policy and Practice*, Vol. 32, No. 5, 1998, pp. 323–329.
17. Edie, L. C. Car-Following and Steady-State Theory for Noncongested Traffic. *Operations Research*, Vol. 9, No. 1, 1961, pp. 66–76.
18. Torné, J. M., F. Soriguera, and N. Geroliminis. Coordinated Active Traffic Management Freeway Strategies Using Capacity-Lagged Cell Transmission Model, *Proceedings of the Transportation Research Board 93rd Annual Meeting*, 2014, paper no. 14-3941.
19. Srivastava, A., and N. Geroliminis. Empirical Observations of Capacity Drop in Freeway Merges with Ramp Control and Integration in a First-Order Model. *Transportation Research Part C: Emerging Technologies*, Vol. 30, 2013, pp. 161–177.
20. Lebacque, J. Two-Phase Bounded-Acceleration Traffic Flow Model: Analytical Solutions and Applications. *Transportation Research Record: Journal of the Transportation Research Board*, Vol.

1852, 2003, pp. 220–230.

21. Monamy, T., H. Haj-Salem, and J.-P. Lebacque. A Macroscopic Node Model Related to Capacity Drop. *Procedia - Social and Behavioral Sciences*, Vol. 54, 2012, pp. 1388–1396.
22. Karafyllis, I., M. Kontorinaki, and M. Papageorgiou. Global Exponential Stabilization of Freeway Models, to appear in *International Journal of Robust and Nonlinear Control*, doi 10.1002/rnc.3412.
23. Roncoli, C., M. Papageorgiou, and I. Papamichail. Traffic Flow Optimisation in Presence of Vehicle Automation and Communication Systems – Part I: A First-Order Multi-Lane Model for Motorway Traffic. *Transportation Research Part C: Emerging Technologies*, Vol. 57, 2015, pp. 241–259.
24. Landman, R. L., A. Hegyi, and S. P. Hoogendoorn. Coordinated Ramp Metering Based on On-ramp Saturation Time Synchronisation. *Presented at the Transportation Research Board 94th Annual Meeting*, 2015, paper no. 15-4037.
25. Leclercq, L., J. A. Laval, and N. Chiabaut. Capacity Drops at Merges: An Endogenous Model. *Transportation Research Part B: Methodological*, Vol. 45, No. 9, 2011, pp. 1302–1313.
26. Jin, W.-L. A Kinematic Wave Theory of Lane-Changing Traffic Flow. *Transportation Research Part B: Methodological*, Vol. 44, No. 8–9, 2010, pp. 1001–1021.
27. Treiber, M., A. Kesting, and D. Helbing. Understanding Widely Scattered Traffic Flows, The Capacity Drop, and Platoons as Effects of Variance-Driven Time Gaps. *Physical Review E*, Vol. 74, No. 1, 2006, p. 016123.
28. Muralidharan, A., and R. Horowitz. Computationally Efficient Model Predictive Control of Freeway Networks. *Transportation Research Part C: Emerging Technologies*, In Press.
29. Ziliaskopoulos, A. K. A Linear Programming Model for the Single Destination System Optimum Dynamic Traffic Assignment Problem. *Transportation Science*, Vol. 34, No. 1, 2000, pp. 37–49.
30. Sun, X., and R. Horowitz. A Localized Switching Ramp-Metering Controller with a Queue Length Regulator for Congested Freeways. *Presented at the American Control Conference*, 2005.
31. Ferrara, A., S. Sacone, and S. Siri. Event-Triggered Model Predictive Schemes for Freeway Traffic Control. *Transportation Research Part C: Emerging Technologies*, Vol. 58, pp. 554-567.
32. Roncoli, C., M. Papageorgiou, and I. Papamichail. Traffic Flow Optimisation in Presence of Vehicle Automation and Communication Systems – Part II: Optimal Control for Multi-Lane Motorways. *Transportation Research Part C: Emerging Technologies*, Vol. 57, 2015, pp. 260–275.
33. Han, Y., Y. Yuan, A. Hegyi, and S. Hoogendoorn. A New Variant of Discretized LWR Model to Reproduce Capacity Drop. *Presented at the 18th Euro Working Group on Transportation*, Delft, The Netherlands, 2015.
34. Highways Agency. *Motorway Incident Detection and Automatic Signalling (MIDAS) Design Standard.* Publication 1st ed. Bristol, UK, 2007.
35. Spiliopoulou, A., M. Kontorinaki, M. Papageorgiou, and P. Kopelias. Macroscopic Traffic Flow Model Validation at Congested Freeway Off-Ramp Areas. *Transportation Research Part C: Emerging Technologies*, Vol. 41, 2014, pp. 18–29.