Queueing Models and
Service Management
Vol. 3, No. 1, page 37-87, 2020

QMSM

Multi-Layer *MMFF* Processes and the *MAP/PH/K+GI* Queue: Theory and Algorithms

Qi-Ming He* and Haoran Wu

Department of Management Sciences
University of Waterloo
200 University Avenue West
Waterloo, Ontario, Canada, N2L 3G1

(Received June 2019 ; accepted November 2019)

Abstract: This paper is concerned with the basic theory and algorithms of multi-layer Markov modulated fluid flow (*MMFF*) processes and an *MAP/PH/K* queue with customer abandonment. For multi-layer *MMFF* processes, we review and refine the existing theory to make it easy to understand, and to make related algorithms meticulously organized. For the queueing system, we combine the *MMFF* approach and the count-server-for-phase (*CSFP*) method to make it possible to analyze it, and to develop an algorithm for computing queueing quantities related to customer abandonment, waiting times, and queue lengths. Some of the quantities are difficult to compute through other means. For both the multi-layer *MMFF* processes and the queueing system, we try to make the analysis easy to follow and the algorithm easy to implement.

Keywords: Abandonment, impatient customers, Markov modulated fluid flow process, Markov process, matrix-analytic methods, queueing systems.

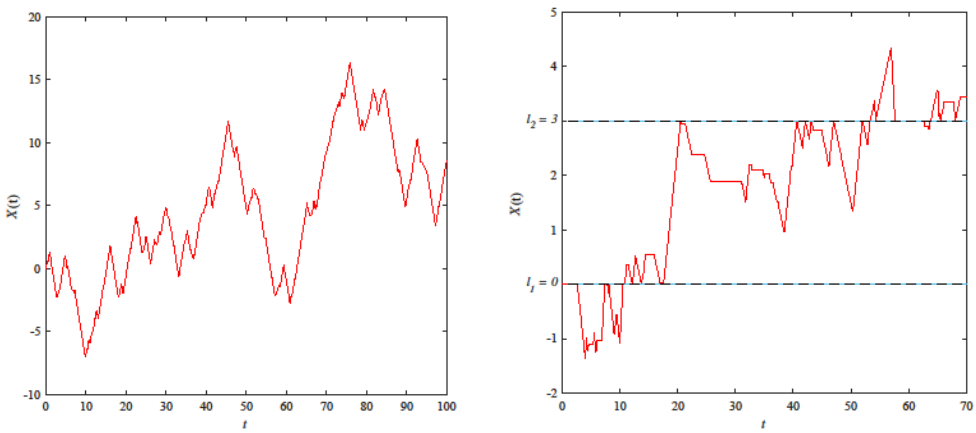
1. Introduction

Multi-layer Markov modulated fluid flow processes, as generalizations of single-layer *MMFF* processes, were introduced and investigated in the past decade. Such stochastic processes have found applications in areas such as queueing theory and risk analysis. Yet the full potential of such a tool is to be explored by researchers and practitioners. In this paper, we review and refine the theory on multi-layer *MMFF* processes. The main subject is the joint stationary distribution of the multi-layer *MMFF* process. We also apply the theory to a queueing model with customer abandonment to demonstrate the usefulness of multi-layer *MMFF* processes. The objective of the paper is to make the basic theory on multi-layer *MMFF* processes more accessible, easier to understand, and convenient to implement for researchers and practitioners.

MMFF processes are piece-wise linear stochastic processes in which the rate of fluid change is modulated by a continuous time Markov chain (to be called *the underlying Markov chain*). Consequently, the fluid level is continuous and can increase linearly,

*Corresponding author
Email : q7he@uwaterloo.ca

decrease linearly, or remain the same for exponential periods of time. Figure 1(a) plots a sample path of a typical *MMFF* process. *MMFF* processes are different from deterministic fluid models that have been used in various branches of queueing theory, especially in the study of the stability of queueing networks. Multi-layer *MMFF* processes are *MMFF* processes in which several underlying Markov chains are used to modulated the rate of fluid change. As such, the fluid level change rate can be different for different layers of the fluid level, which are separated by border lines. State changes of the underlying Markov chains are regulated as the fluid level passing, reflecting, entering, or leaving border lines so as to combine the underlying Markov chains. Figure 1(b) plots a sample path of a typical multi-layer *MMFF* process with two (dashed) border lines and three layers. Multi-layer *MMFF* processes are complicated stochastic processes with complicated solutions for basic quantities such as the stationary distribution of the process. They are amenable to stochastic systems in which key system variables/parameters are modulated by a stochastic process. They may not be the most convenient tools for analyzing simple stochastic systems such as the M/M/1 queue. Yet they are the power house for the investigation of complicated stochastic systems such as queueing models, risk/insurance models, and dam models.



(a) A sample path of a single-layer *MMFF* process. (b) A sample path of a three-layer *MMFF* process.

Figure 1. Sample paths of *MMFF* processes.

The main contributions of the paper are (i) reviewing and refining the theory and algorithm for computing the joint stationary distributions of multi-layer *MMFF* processes; and (ii) putting together several ideas in applied probability and queueing theory to develop algorithms for stochastic models arising from queueing systems and risk models. Specifically, that includes (a) Presenting and refining the existing theory on multi-layer *MMFF* processes; (b) Developing an easy way to implement computational procedure for the joint stationary distribution of multi-layer *MMFF* processes; and (c) Combining the *MMFF* approach and the *CSFP* method to develop a relatively simple and efficient

algorithm to analyze moderately large scale queueing systems such as the *MAP*/*PH*/*K* queue with customer abandonment and a moderately large number of servers. The algorithms presented in this paper can be useful for practitioners in their design of stochastic systems such as call centres. The algorithms can also be useful for researchers to do numerical experiments in their investigations of stochastic models. In addition, the queueing analysis finds quantities related to customers abandoning the queue before reaching the head of the waiting queue (e.g., the abandonment probability and abandonment (waiting) time), customers abandoning the queue at the head of the waiting queue, and the queue length distributions, which are difficult to derive.

The rest of the paper is organized as follows. In Section 2, we give a brief literature review on multi-layer *MMFF* processes, and the study of queues with customer abandonment. In Section 3, we define multi-layer *MMFF* processes. In Section 4, basic quantities and their properties related to *MMFF* processes are collected. In Section 5, we review and refine the theory on the joint stationary distribution of multi-layer *MMFF* processes. Step by step, we develop a computational procedure for the joint stationary distribution. In Section 6, we apply the *MMFF* approach to the *MAP*/*PH*/*K*+*GI* queue. Algorithms are developed for computing a variety of queueing quantities for the queue. We also present a few numerical examples and discuss some computation issues when the number of servers is moderately big. Section 7 concludes the paper.

2. Literature Review

In the literature, *MMFF* processes are also known as *fluid flow models*, *stochastic fluid flows*, or *Markovian fluid flows*. Early works on *MMFF* processes include Loynes [36], Anick *et al.* [5], Rogers [42], and Asmussen [6], which were motivated by an application in dam control. In those papers, *MMFF* processes were introduced and some basic quantities were obtained. By using Wiener-Hopf factorization, basic matrices such as Ψ, which represents the state change at regenerative epochs (e.g., the fluid level returns to zero), and $\mathcal{U}$, which represents the change of state as the fluid level reaches a new low level, were obtained. Since *MMFF* processes can approach positive infinity, negative infinity, or both (depending on the mean drift rate), they do not have stationary distributions. Nevertheless, stationary distributions exist for their truncated version, which is known as the Markov modulated fluid queues (*MMFQ*s). By using time-reversed Markov processes, the joint stationary distributions of the fluid level and the state of underlying Markov chain were obtained for *MMFQ*s (e.g., Rogers [42]). We shall use *MMFF* for *MMFQ* in this paper with the understanding that stationary distributions exist under a certain restriction. *MMFF* processes with a Brownian component were introduced and investigated. We do not review works in that direction since our focus is on *MMFF* processes without a Brownian

component.

Ramaswami [41] discovered a relationship between the basic quantities $\{\Psi, \mathscr{U}\}$ and the basic matrix G for quasi birth-and-death processes in matrix-analytic methods (Neuts [39] and Latouche and Ramaswami [35]), which led to a new method for computing Ψ, , in addition to the classical method of solving a quadratic Riccati equation. Ramaswami [41] also found a relationship between the joint stationary distribution and the crossing numbers of the fluid level, which led to a new approach to compute the joint stationary distribution and an application of matrix $\mathscr{K}$, another basic quantity of *MMFF* processes. Since then, the study of *MMFF* processes attracted the attention of many researchers and a large number of papers appeared with various applications including

i) In matrix-analytic methods: see Ramaswami [41], Ahn and Ramaswami [2, 3, 4], da Silva Soares and Latouche [19, 20, 21, 22], and Latouche and Nguyen [34];
ii) In risk analysis: Ahn *et al.* [1], Asmussen [7], Avram and Usabel [8], and Badescu *et al.* [9, 10, 11], and Badescu and Landriault [12, 13];
iii) In queueing theory: Horváth and Van Houdt [32], Van Houdt [43], and Horváth [31]; and
iv) In the theory of *MMFF* processes (e.g., two stage *MMFF* processes, first passage times, and two dimensional *MMFF* processes): Bean *et al*. [16, 17], and Bean and O'Reilly [14, 15].

A natural extension of (the single layer) *MMFF* processes are multi-layer *MMFF* processes, which were introduced in da Silva Soares and Latouche [21]. In fact, that paper considered the standard *MMFF* processes truncated from both above and below. The paper extended existing results on first passage probabilities and the joint stationary distribution. It was immediately clear from their work that multi-layer *MMFF* processes can be analyzed in a similar way, although the solution process is more involved and the presentation of results can be tedious. The main idea is to first analyze the process within individual layers and then combine results together through the transitions related border lines. Since then, more studies on multi-layer *MMFF* processes and their applications in queueing theory followed.

- The basic theory for the analysis on multi-layer *MMFF* processes was established in da Silva Soares and Latouche [21, 22], especially that are related to the joint stationary distribution of the processes. We review their results in this paper. We refine the theory on the joint stationary distribution, and present the theory and related algorithm in a systematic form. In Bean and O'Reilly [14], the multi-layer *MMFF* processes, in their full scale, were introduced. Their paper focused on the first passage time and first passage probabilities.

- Horváth and Van Houdt [32], Van Houdt [43], and Horváth [31] applied the theory on multi-layer *MMFF* processes to queueing models. Van Houdt [43] investigated a single server queue with multiple types of customers and customer abandonment, and obtained quantities related to customer abandonment and waiting times. Horváth [31] analyzed a single server queue with multiple types of customers with service priority. Our work on the queueing model is close to that in Van Houdt [43] in which a single server queue with customer abandonment is studied. We consider a queueing model with many servers and customer abandonment, and extend the analysis to more queueing quantities (e.g., different types of abandonment probabilities and waiting times, and the mean queue length).

Queueing systems with customer abandonment are important in the design of many stochastic systems such as call centres. The investigation of such queueing systems has been extensive (e.g., Dai and He [23, 24], Dai *et al.* [25], and references therein). Choi *et al.* [18] introduced a method to analyze the *MAP*/*M*/*K*+*GI* queue with constant abandonment time (i.e., $MAP/M/K+\tau$). Kim and Kim [33] adopted the same method to analyze the *M*/*PH*/1 queue with constant abandonment time. Following their approach, He *et al.* [18] investigated the *M*/*PH*/*K* queue with constant abandonment time. Unfortunately, the method cannot be applied to the *MAP*/*PH*/*K* queue with customer abandonment, due to the lack of commutability of some matrices.

MMFF processes have been proven to be an effective tool in analyzing queueing models. The basic idea of the approach is to introduce an *MMFF* process associated with the workload/age process of the queueing systems. If the stationary distribution of the fluid flow process can be found, then some queueing quantities can be obtained. Following previous works, we apply the multi-layer *MMFF* processes to the *MAP*/*PH*/*K*+*GI* queue, where the abandonment time distribution is assumed to be finite discrete. The queueing model is quite general since *MAP*s can approximate any arrival process and *PH* random variables can approximate any nonnegative random variables. To deal with a state space dimensionality issue, similar to He *et al.* [18], we use an approach developed in Ramaswami [40] (also see He and Alfa [29]), called *CSFP* (count-server-for-phase), to reduce the state space so that the algorithm developed in this paper can handle systems with up to one hundred servers. Thus, algorithms developed in this paper can be used by researchers and practitioners in their studies/design to gain insight on stochastic systems of interest.

The power of *MMFF* processes can be further demonstrated by their capacity in dealing with more complex queueing systems. For example, with minor modifications, the method presented in Section 6 can be used to analyze the *MAP*/*PH*/*K*+*GI* queue in which the abandonment time of the customer at the head of the waiting queue has a different distribution than that of the rest. Further, the method can also be extended to analyze queues

in which the customer arrival process and/or the service times depends on the age of the customer at the head of the waiting queue.

3. Multi-Layer *MMFF* Processes: Definition

We present the multi-layer *MMFF* processes first introduced in Bean and O'Reilly [14]. As mentioned earlier, a multi-layer *MMFF* process is a fluid flow process in which the fluid level is a piece-wise linear continuous function of the time and the change rate of its fluid level is modulated by a continuous time Markov chain. A two dimensional process $\{(X(t),\phi(t)), t\geq 0\}$ is called a *multi-layer Markov modulated fluid flow (MMFF) process* if the following conditions are satisfied.

1. There are $N+1$ constants $\{l_0=-\infty, l_1, ..., l_N=\infty\}$ such that $N\geq 1$ and $l_0<l_1<...<l_N$, to be called *Borders*. Those borders form N intervals (l_0,l_1), (l_1,l_2), ..., and (l_{N-1},l_N), to be called *Layer* 1, 2, ..., and N, respectively.
2. If $X(t)$ is in Layer n, for $n=1,...,N-1$, $\{\phi(t), t\geq 0\}$ is a continuous time irreducible Markov chain on finite state space $\mathcal{S}^{(n)}$ with infinitesimal generator $Q^{(n)}$.
3. The fluid process $\{X(t), t\geq 0\}$ is controlled by $\phi(.)$ such that the value of $X(t)$ changes linearly at rate $c_{\phi(t)}^{(L(X(t)))}$ at time t, where $L(x)=n$ if $l_{n-1}<x<l_n$, for $n=1,...,N$. The rate $c_i^{(n)}$ of fluid level change can be positive, negative, or zero. We put the rates into vectors $\mathbf{c}^{(n)}=\{c_i^{(n)}, i\in\mathcal{S}^{(n)}\}$, for $n=1,...,N$. For convenience, we partition the state space $\mathcal{S}^{(n)}$ into three subsets according to the sign of $c_i^{(n)}$ as follows:

$$\mathcal{S}_+^{(n)}=\{i\in\mathcal{S}^{(n)}:\ c_i^{(n)}>0\},\ \ \mathcal{S}_-^{(n)}=\{i\in\mathcal{S}^{(n)}:\ c_i^{(n)}<0\},\ \ \mathcal{S}_0^{(n)}=\{i\in\mathcal{S}^{(n)}:\ c_i^{(n)}=0\}. \quad (3.1)$$

We further divide $\mathbf{c}^{(n)}$, according the signs of its elements, and the infinitesimal generator $Q^{(n)}$ of the underlying Markov chain as

$$\mathbf{c}^{(n)}=(\mathbf{c}_+^{(n)},\mathbf{c}_-^{(n)},0);$$
$$Q^{(n)}=\begin{matrix}\mathcal{S}_+^{(n)}\\ \mathcal{S}_-^{(n)}\\ \mathcal{S}_0^{(n)}\end{matrix}\begin{pmatrix} Q_{++}^{(n)} & Q_{+-}^{(n)} & Q_{+0}^{(n)} \\ Q_{-+}^{(n)} & Q_{--}^{(n)} & Q_{-0}^{(n)} \\ Q_{0+}^{(n)} & Q_{0-}^{(n)} & Q_{00}^{(n)} \end{pmatrix}. \quad (3.2)$$

We note that $\mathcal{S}_+^{(n)}, \mathcal{S}_-^{(n)}$, and $\mathcal{S}_0^{(n)}$ are placed in the above definition to show the directions of transitions, and are not a part of $Q^{(n)}$.

4. If $X(t)=l_n$, for $n=1,...,N-1$, $\{\phi(t), t\geq 0\}$ is a continuous time irreducible Markov chain on finite state space $\mathcal{S}_b^{(n)}$ with sub-generator $Q_{bb}^{(n)}$. During the period that $\phi(t)$ is in $\mathcal{S}_b^{(n)}$, $X(t)$ remains at l_n until $\phi(t)$ switches from $\mathcal{S}_b^{(n)}$ to either $\mathcal{S}_-^{(n)}$ or

$\mathcal{S}_+^{(n+1)}$. The process $\phi(t)$ can go from $\mathcal{S}_b^{(n)}$ to (i) $\mathcal{S}_+^{(n+1)}$ with transition rate matrix $Q_{b+}^{(n)}$; and (ii) $\mathcal{S}_-^{(n)}$ with transition rate matrix $Q_{b-}^{(n)}$.

5. If $X(t)$ reaches l_n from below, for $n=1,...,N-1$, the process $\{\phi(t), t\geq 0\}$ can switch from $\mathcal{S}^{(n)}$ (actually from $\mathcal{S}_+^{(n)}$) to (i) $\mathcal{S}_-^{(n)}$ (i.e., reflecting back to Layer n) with probability (matrix) $P_{+b-}^{(n)}$; (ii) $\mathcal{S}_+^{(n+1)}$ (i.e., passing Border l_n to Layer $n+1$) with probability $P_{+b+}^{(n)}$; or (iii) into $\mathcal{S}_b^{(n)}$ with probability $P_{+bb}^{(n)}$.
6. If $X(t)$ reaches l_n from above, for $n=1,...,N-1$, the process $\{\phi(t), t\geq 0\}$ can switch from $\mathcal{S}^{(n+1)}$ (actually $\mathcal{S}_-^{(n+1)}$) to (i) $\mathcal{S}_+^{(n+1)}$ (i.e., reflecting back to Layer $n+1$) with probability $P_{-b+}^{(n)}$; (ii) $\mathcal{S}_-^{(n)}$ (i.e., passing Border l_n to Layer n) with probability $P_{-b-}^{(n)}$; or (iii) into $\mathcal{S}_b^{(n)}$ with probability $P_{-bb}^{(n)}$.

By the above definition, for $n=1,2,...,N-1$, we must have (i) $Q_{bb}^{(n)}\mathbf{e}+Q_{b+}^{(n)}\mathbf{e}+Q_{b-}^{(n)}\mathbf{e}=0$, where $\mathbf{e}$ is the column vector of ones and an appropriate size; (ii) $P_{+b+}^{(n)}\mathbf{e}+P_{+b-}^{(n)}\mathbf{e}+P_{+bb}^{(n)}\mathbf{e}=\mathbf{e}$; and (iii) $P_{-b+}^{(n)}\mathbf{e}+P_{-b-}^{(n)}\mathbf{e}+P_{-bb}^{(n)}\mathbf{e}=\mathbf{e}$. If we define $c_i^{(n)}=0$ for all n and $i\in\mathcal{S}_b^{(n)}$, then $X(t)$ is controlled by $\phi(t)$ explicitly as

$$X(t)=X(0)+\int_0^t c_{\phi(s)}^{(L(X(s)))}\mathrm{d}s, \quad \text{or} \quad \frac{\mathrm{d}X(t)}{\mathrm{d}t}=c_{\phi(t)}^{(L(X(t)))}. \tag{3.3}$$

Based on the above equations, the process $\{X(t), t\geq 0\}$ can be analyzed by using the ordinary different equation (ODE) method (e.g., Anick *et al.* [5]). The more popular approach is matrix-analytic methods, which are used in this paper.

In da Silva Soares and Latouche [22], a border with nonempty $\mathcal{S}_b^{(n)}$ is called a *sticky border*. We shall call a border a *passing border* if one of $P_{-b-}^{(n)}$ and $P_{+b+}^{(n)}$ is nonzero, and a *reflecting border* if one of $P_{-b+}^{(n)}$ and $P_{+b-}^{(n)}$ is nonzero.

The process has N layers separated by $N-1$ borders. If $N=1$, the process is the classical *MMFF* process. The classical *MMFQ* is a special case with $N=2$, Border $l_1=0$ is a reflecting border, and Layer 1 has an empty set of underlying states. The *MMFQ* can be considered as a two-layer *MMFF* process truncated at Border $l_1=0$. Such an *MMFF* process is called an *MMFQ* since the fluid level is always nonnegative and its dynamics reflects the change of queue length, workload, or the age of a customer in queueing systems.

Example 3.1. Parameters of a multi-layer *MMFF* process with $N=3$ are presented in Table 1. Figure 1(b), Figures 2, 3, 4, and Figure 5(a) are generated from this example.

Table 1. Parameters for Example 3.1 with $N=3$.

Borders / Layers	**Parameters**
Border ($L_3=\infty$)	Not defined
Layer 3	$\mathbf{c}^{(3)}=(0.5,-2,-1,0);\ Q^{(3)}=\begin{pmatrix}-1 & 0.5 & 0 & 0.5\\ 1 & -2 & 0 & 1\\ 1 & 0 & -2 & 1\\ 0 & 1 & 0 & -1\end{pmatrix}$
Border ($L_2=3$)	$Q^{(2)}_{bb}=\begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix};Q^{(2)}_{b+}=\begin{pmatrix}0.5\\ 0.5\end{pmatrix};Q^{(2)}_{b-}=\begin{pmatrix}0.5\\ 0.5\end{pmatrix};$ $P^{(2)}_{+b+}=\begin{pmatrix}0.1\end{pmatrix};\quad P^{(2)}_{+b-}=\begin{pmatrix}0.4\end{pmatrix};P^{(2)}_{+bb}=\begin{pmatrix}0.5 & 0\end{pmatrix};$ $P^{(2)}_{-b+}=\begin{pmatrix}0.5\\ 0.5\end{pmatrix};\quad P^{(2)}_{-b-}=\begin{pmatrix}0.1\\ 0.1\end{pmatrix};\ P^{(2)}_{-bb}=\begin{pmatrix}0.4\\ 0.4\end{pmatrix}.$
Layer 2	$\mathbf{c}^{(2)}=(1,-0.5,0,0);\ Q^{(2)}=\begin{pmatrix}-1 & 0.5 & 0.5 & 0\\ 1 & -2 & 1 & 0\\ 0 & 1 & -1 & 0\\ 1 & 0 & 0 & -1\end{pmatrix}.$
Border ($L_1=0$)	$Q^{(1)}_{bb}=\begin{pmatrix}-1\end{pmatrix};\ Q^{(1)}_{b+}=\begin{pmatrix}0.5\end{pmatrix};\ Q^{(1)}_{b-}=\begin{pmatrix}0.5\end{pmatrix};$ $P^{(1)}_{+b+}=\begin{pmatrix}0.5\\ 0.5\end{pmatrix};P^{(1)}_{+b-}=\begin{pmatrix}0.1\\ 0.1\end{pmatrix};P^{(1)}_{+bb}=\begin{pmatrix}0.4\\ 0.4\end{pmatrix};$ $P^{(1)}_{-b+}=\begin{pmatrix}0.4\end{pmatrix};\ P^{(1)}_{-b-}=\begin{pmatrix}0.1\end{pmatrix};\ P^{(1)}_{-bb}=\begin{pmatrix}0.5\end{pmatrix}.$
Layer 1	$\mathbf{c}^{(1)}=(2,1,-1,0);\ Q^{(1)}=\begin{pmatrix}-1 & 0 & 1 & 0\\ 0 & -1 & 0 & 1\\ 1 & 1 & -2 & 0\\ 1 & 0 & 1 & -2\end{pmatrix}.$
Border ($L_0=-\infty$)	Not defined

In general, the multi-layer *MMFF* process does not have the independent incremental property, and its evolutions in individual layers interact with each other through the borders. On the other hand, it evolves conditionally independently within individual layers. This observation implies that one can first investigate the process in individual layers and then combine them together. The study of the process within independent layers is equivalent to that of the single layer *MMFF* process. Thus, we shall first introduce a number of basic quantities that only associated with a single layer *MMFF* process.

4. Preliminaries: Basic Quantities

In this section, we shall introduce quantities $\{\Psi,\ \widehat{\Psi},\ \mathscr{K},\ \widehat{\mathscr{K}},\ \mathscr{U},\ \widehat{\mathscr{U}}\}$ for each layer. For that purpose, we assume in this section that there is only one layer, and remove the superscript/subscript "n". We refer to Latouche and Nguyen [34] for a detailed review on those quantities.

Let $\boldsymbol{\alpha}$ be the stationary distribution of infinitesimal generator Q, which is the unique solution to linear system $\boldsymbol{\alpha}Q=0$ and $\boldsymbol{\alpha}\mathbf{e}=1$. We define

$$\mu=\boldsymbol{\alpha}\mathbf{c}, \tag{4.1}$$

which is the mean drift of the fluid flow per unit time in steady state. Intuitively, as $t\to\infty$, if $\mu>0$, the process will drift to $+\infty$; if $\mu<0$, the process will drift to $-\infty$; and if $\mu=0$, $|X(t)|\to\infty$. It has been shown mathematically rigorously that the three limits hold with probability one.

Matrices Ψ and $\widehat{\Psi}$ are the most important quantities in the analysis of *MMFF* processes. Many other key quantities can be expressed explicitly in Ψ and $\widehat{\Psi}$. To define Ψ and $\widehat{\Psi}$, we introduce embedded regenerative processes in $\{(X(t),\phi(t)),t\geq 0\}$. Define, $\delta_0=\inf\{t>0: X(t)>0\}$, and for $n>0$,

$$\begin{aligned}\theta_n&=\inf\{t>\delta_{n-1}: X(t)=0\},\\ \delta_n&=\inf\{t>\theta_n: X(t)>0\},\end{aligned} \tag{4.2}$$

which are called *regenerative epochs* (see Figure 2), if the underlying process $\phi(t)$ is in $\mathcal{S}_+$ or $\mathcal{S}_-$. For example, $\{(X(\theta_n),\phi(\theta_n)),n=1,2,...\}$ is a regenerative process with state space $\{0\}\times\mathcal{S}_-$. Elements of matrices Ψ and $\widehat{\Psi}$ are defined as follows:

$$\begin{aligned}\Psi_{i,j}&=P\{\theta_{n+1}-\delta_n<\infty,\phi(\theta_{n+1})=j\,|\,\phi(\delta_n)=i\},\ \text{for } i\in\mathcal{S}_+,\ j\in\mathcal{S}_-;\\ \widehat{\Psi}_{i,j}&=P\{\delta_n-\theta_n<\infty,\phi(\delta_n)=j\,|\,\phi(\theta_n)=i\},\ \text{for } i\in\mathcal{S}_-,\ j\in\mathcal{S}_+.\end{aligned} \tag{4.3}$$

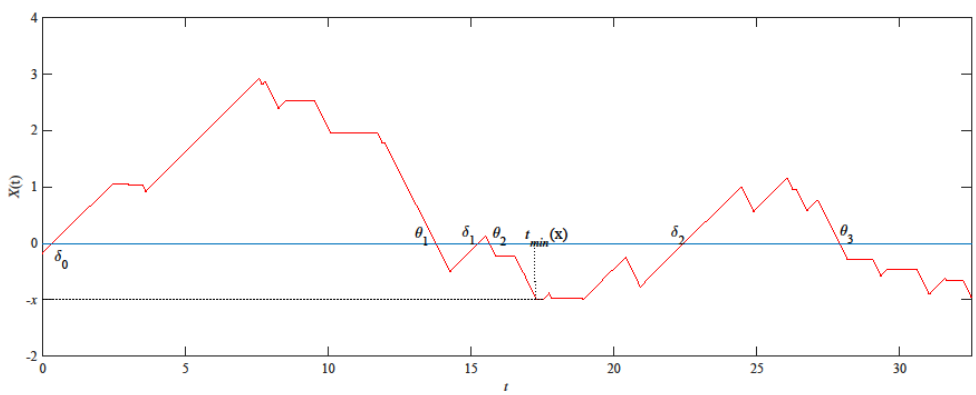


Figure 2. δ_n, θ_n, and $t_{\min}(x)$.

It is easy to see that Ψ records the transition of the state of the underlying Markov chain Q from an epoch that the fluid level $X(t)$ starts to increase from zero to the next first epoch that $X(t)$ reaches zero. Matrix $\widehat{\Psi}$ can be interpreted similarly.

If $\phi(t)\in\mathcal{S}_0$, the fluid flow level $X(t)$ can remain unchanged for a period of time. To analyze Ψ and $\widehat{\Psi}$, as demonstrated in da Silva Soares and Latouche [21], it is useful and

without loss of generality to consider the process obtained by censoring the time periods that $\phi(t)$ is in $\mathcal{S}_0$. The censored underlying Markov chain is defined by

$$T=\begin{pmatrix} T_{++} & T_{+-} \\ T_{-+} & T_{--} \end{pmatrix}=\begin{pmatrix} Q_{++} & Q_{+-} \\ Q_{-+} & Q_{--} \end{pmatrix}+\begin{pmatrix} Q_{+0} \\ Q_{-0} \end{pmatrix}(-Q_{00})^{-1}(Q_{0+}, Q_{0-}), \tag{4.4}$$

where T_{++}, for example, contains the transition rates from $\mathcal{S}_+$ to $\mathcal{S}_+$ directly, which are given by Q_{++}, and indirectly via $\mathcal{S}_0$, which are given by $Q_{+0}(-Q_{00})^{-1}Q_{0+}$.

In the rest of this section, we work with both infinitesimal generators T and Q. For convenience, we also define positive diagonal matrices C_+ and C_- as follows:

$$C_+ = diag(\mathbf{c}_+) \quad \text{and} \quad C_- = -diag(\mathbf{c}_-). \tag{4.5}$$

Lemma 1. *(Rogers [42]) Matrices Ψ and $\widehat{\Psi}$ are the minimal nonnegative solution to the following quadratic Riccati equations, respectively:*

$$\begin{aligned} &C_+^{-1}T_{+-}+C_+^{-1}T_{++}\Psi+\Psi C_-^{-1}T_{--}+\Psi C_-^{-1}T_{-+}\Psi=0; \\ &C_-^{-1}T_{-+}+C_-^{-1}T_{--}\widehat{\Psi}+\widehat{\Psi}C_+^{-1}T_{++}+\widehat{\Psi}C_+^{-1}T_{+-}\widehat{\Psi}=0. \end{aligned} \tag{4.6}$$

We refer to Latouche and Nguyen [34], Guo [26, 27], Ramaswami [41], and Meini [37] for more details and algorithms for computing Ψ and $\widehat{\Psi}$.

Second, we consider the underlying Markov chain when the fluid level reaches a new low/high point. We define matrices $\mathcal{U}$ and $\widehat{\mathcal{U}}$ as

$$\begin{aligned} &\mathcal{U}=C_-^{-1}T_{--}+C_-^{-1}T_{-+}\Psi; \\ &\widehat{\mathcal{U}}=C_+^{-1}T_{++}+C_+^{-1}T_{+-}\widehat{\Psi}. \end{aligned} \tag{4.7}$$

The following probabilistic interpretations of $\mathcal{U}$ holds:

$$\begin{aligned} &\tau_x^-=\inf\{t: X(t)<X(0)-x\}, \text{ for } x>0; \\ &(e^{\mathcal{U}x})_{i,j}=P\{\tau_x^-<\infty, \phi(\tau_x^-)=j \mid \phi(0)=i\}, \text{ for } i, j\in\mathcal{S}_-; \\ &(\Psi e^{\mathcal{U}x})_{i,j}=P\{\tau_x^-<\infty, \phi(\tau_x^-)=j \mid \phi(0)=i\}, \text{ for } i\in\mathcal{S}_+, j\in\mathcal{S}_-; \end{aligned} \tag{4.8}$$

Thus, $\mathcal{U}$ plays the role of an infinitesimal generator of a continuous time Markov chain for which the time is the minimal fluid level and the state space is $\mathcal{S}_-$. That is: $\mathcal{U}$ is related to the state of the underlying Markov chain every time the fluid level reaches a new low point (Asmussen [6]). Define, for $x\geq 0,$,

$$\begin{aligned} &t_{\min}(x)=\min\{t: X(t)=-x\}; \\ &i_{\min}(x)=\phi(t_{\min}(x)), \end{aligned} \tag{4.9}$$

where $i_{\min}(x)$ is the state of the underlying Markov chain at the first time epoch that $X(t)$

reaches $-x$.

Lemma 2. *(Asmussen [6]) If* $\mu \le 0$, $\{i_{\min}(x), x \ge 0\}$ *is a continuous time Markov chain with infinitesimal generator* $\mathcal{U}$. *If* $\mu > 0$, *then* $\{i_{\min}(x), x \ge 0\}$ *is an absorption Markov chain with state space* $\mathcal{S}_- \bigcup \{\Delta\}$, *where* Δ *is defined as an absorption state, and infinitesimal generator*

$$\begin{array}{c} \\ \mathcal{S}_- \\ \Delta \end{array} \begin{pmatrix} \mathcal{U} & -\mathcal{U}\mathbf{e} \\ 0 & 0 \end{pmatrix}. \tag{4.10}$$

Similarly, one can consider the underlying Markov chain when the fluid level reaches a new high point, which is related to a continuous time Markov chain with infinitesimal generator (or subgenerator) $\widehat{\mathcal{U}}$.

Third, we consider matrices $\mathcal{K}$ and $\widehat{\mathcal{K}}$, which are defined as

$$\begin{aligned} \mathcal{K} &= C_+^{-1} T_{++} + \Psi C_-^{-1} T_{-+}; \\ \widehat{\mathcal{K}} &= C_-^{-1} T_{--} + \widehat{\Psi} C_+^{-1} T_{+-}. \end{aligned} \tag{4.11}$$

Matrix $\mathcal{K}$ is associated with numbers of visits to a certain fluid level and state during first passage periods. We assume that $\delta_0 = 0$ (then $X(0) = 0$) and $\phi(0) = i$.

- For $i, j \in \mathcal{S}_+$ and $x > 0$, we define $(N_{++}(x))_{i,j}$ as the mean number of visits of the process $(X(t), \phi(t))$ to state (x, j) from below before $X(t)$ returns to zero. This type of visits are called *upcrossings* of fluid level x.
- For $i \in \mathcal{S}_+$, $j \in \mathcal{S}_-$, and $x > 0$, we define $(N_{+-}(x))_{i,j}$ as the expected number of visits of the process $(X(t), \phi(t))$ to state (x, j) from above before $X(t)$ returns to zero. Such visits are called *downcrossings* of fluid level x.
- For $i, j \in \mathcal{S}_-$ and $x < 0$, we define $(N_{--}(x))_{i,j}$ as the mean number of visits of the process $(X(t), \phi(t))$ to state (x, j) from above before $X(t)$ returns to zero.
- For $i \in \mathcal{S}_-$, $j \in \mathcal{S}_+$, and $x < 0$, we define $(N_{-+}(x))_{i,j}$ as the expected number of visits of the process $(X(t), \phi(t))$ to state (x, j) from below before $X(t)$ returns to zero.

Lemma 3. *(Ramaswami [41]) For* $x > 0$, *we have (i)* $N_{++}(x) = \exp\{\mathcal{K}x\}$; *and (ii)* $N_{+-}(x) = N_{++}(x)\Psi = \exp\{\mathcal{K}x\}\Psi$. *For* $x < 0$, *we have (iii)* $N_{--}(x) = \exp\{\widehat{\mathcal{K}}(-x)\}$; *and (iv)* $N_{-+}(x) = N_{--}(x)\widehat{\Psi} = \exp\{\widehat{\mathcal{K}}(-x)\}\widehat{\Psi}$.

Now, we summarize the relationship between μ and our basic matrices, which will be referenced repeatedly throughout this paper.

Lemma 4. *(Rogers [42], Asmussen [6], Ramaswami [41]) The relationships between* μ *and basic quantities are as follows.*

4.1 If $\mu>0$, then we have (i) $\Psi\mathbf{e}<\mathbf{e}$ and $\widehat{\Psi}\mathbf{e}=\mathbf{e}$; (ii) $\mathcal{U}\mathbf{e}<0$ and $\widehat{\mathcal{U}}\mathbf{e}=0$; and (ii) $\mathcal{K}$ is non-invertible and $\widehat{\mathcal{K}}$ is invertible.

4.2 If $\mu=0$, then we have (i) $\Psi\mathbf{e}=\mathbf{e}$ and $\widehat{\Psi}\mathbf{e}=\mathbf{e}$; (ii) $\mathcal{U}\mathbf{e}=0$ and $\widehat{\mathcal{U}}\mathbf{e}=0$; and (iii) $\mathcal{K}$ and $\widehat{\mathcal{K}}$ are non-invertible.

4.3 If $\mu<0$, then we have (i) $\Psi\mathbf{e}=\mathbf{e}$ and $\widehat{\Psi}\mathbf{e}<\mathbf{e}$; (ii) $\mathcal{U}\mathbf{e}=0$ and $\widehat{\mathcal{U}}\mathbf{e}<0$; and (iii) $\mathcal{K}$ is invertible and $\widehat{\mathcal{K}}$ is non-invertible.

For extensions to a multi-layer *MMFF* process, we need quantities when the process is constrained to an interval, say (a,b). Therefore, we define, for $a<x<b$,

- $(N_+^{(a,b)}(x))_{i,j}$ be the expected number of crossings of level x at state $j\in\mathcal{S}$ before the process reaches level a or level b, given that the process started in (a,i) for $i\in\mathcal{S}_+$. (See Figure 3)
- $(\widehat{N}_-^{(a,b)}(x))_{i,j}$ be the expected number of crossings of level x at state $j\in\mathcal{S}$ before the process reaches level b or level a, given that the process started in (b,i) for $i\in\mathcal{S}_-$.

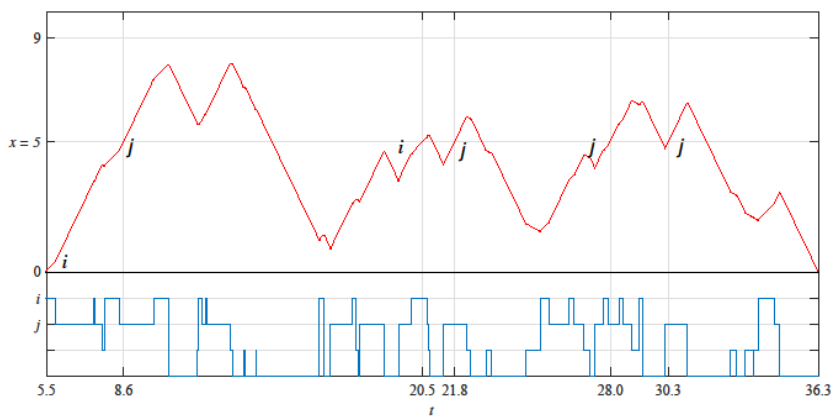


Figure 3. Upcrossings of level x, starting from level $a=0$, without visiting leve $b=9$.

Matrix $N_+^{(a,b)}(x)$ ($\widehat{N}_-^{(a,b)}(x)$) can be divided into two subblocks $N_{++}^{(a,b)}(x)$ ($\widehat{N}_{-+}^{(a,b)}(x)$) for upcrossings and $N_{+-}^{(a,b)}(x)$ ($\widehat{N}_{--}^{(a,b)}(x)$) for downcrossings according to $j\in\mathcal{S}_+$ or $j\in\mathcal{S}_-$, respectively.

Lemma 5. *(da Silva Soares and Latouche [21]) For $a<x<b$, we have*

$$\begin{pmatrix} I & e^{\mathcal{K}(b-a)}\Psi \\ e^{\widehat{\mathcal{K}}(b-a)}\widehat{\Psi} & I \end{pmatrix}\begin{pmatrix} N_+^{(a,b)}(x) \\ \widehat{N}_-^{(a,b)}(x) \end{pmatrix}=\begin{pmatrix} e^{\mathcal{K}(x-a)} & 0 \\ 0 & e^{\widehat{\mathcal{K}}(b-x)} \end{pmatrix}\begin{pmatrix} I & \Psi \\ \widehat{\Psi} & I \end{pmatrix}. \tag{4.12}$$

The first matrix on the left hand side in the above equation is invertible if $\mu \neq 0$.

For the first passage probabilities from one fluid level to another (e.g., from a to b or vice versa), we define

- $\Psi_{+-}^{(b-a)}$ is defined similar to Ψ except that the process does not reach fluid level b and the process starts in fluid level a; $\widehat{\Psi}_{-+}^{(b-a)}$ is defined similar to $\widehat{\Psi}$ except that the process does not reach fluid level a and the process starts in fluid level b.
- $\Lambda_{++}^{(b-a)}$ is defined as the probabilities for the process to go from level a to level b before returning to level a. $\widehat{\Lambda}_{--}^{(b-a)}$ is defined as the probabilities for the process to go from level b to level a before returning to level b.

Lemma 6. *(da Silva Soares and Latouche [21]) The matrices of first passage probabilities satisfy the following equations:*

$$\begin{pmatrix} \Lambda_{++}^{(b-a)} & \Psi_{+-}^{(b-a)} \\ \widehat{\Psi}_{-+}^{(b-a)} & \widehat{\Lambda}_{--}^{(b-a)} \end{pmatrix} \begin{pmatrix} I & \Psi e^{\mathscr{U}(b-a)} \\ \widehat{\Psi} e^{\widehat{\mathscr{U}}(b-a)} & I \end{pmatrix} = \begin{pmatrix} e^{\widehat{\mathscr{U}}(b-a)} & \Psi \\ \widehat{\Psi} & e^{\mathscr{U}(b-a)} \end{pmatrix}. \tag{4.13}$$

The second matrix on the left-hand-side of the above equation is invertible if $\mu \neq 0$.

Although Lemmas 5 and 6 are developed for *MMFF* processes with only one layer, they play a key role in the analysis of multi-layer *MMFF* processes and will be used repeatedly in the next section.

5. Joint Stationary Distribution and Algorithm

In this section, we review and refine an algorithm for computing the joint stationary distribution of the fluid level and the state of the underlying Markov chain developed in da Silva Soares and Latouche [21]. A censored CTMC and a linear system are introduced for border probabilities and limits of the density function, which are constants and coefficients used in the solutions of the joint stationary distribution. Define, for $-\infty < x < \infty$,

$$\begin{aligned} p_j^{(n)} &= \lim_{t\to\infty} P\{X(t) = l_n, \phi(t) = j \mid X(0), \phi(0)\}, \text{ for } j \in \mathcal{S}_b^{(n)}, n = 1,2,...,N-1; \\ g_j^{(n)}(x) &= \lim_{t\to\infty} P\{X(t) < x, \phi(t) = j \mid X(0), \phi(0)\}, \text{ for } j \in \mathcal{S}^{(n)}, n = 1,2,...,N; \\ \pi_j^{(n)}(x) &= \frac{\mathrm{d}g_j^{(n)}(x)}{\mathrm{d}x}, \text{ for } j \in \mathcal{S}^{(n)}, n = 1,2,...,N. \end{aligned} \tag{5.1}$$

Let $\mathbf{p}^{(n)} = (p_j^{(n)} : j \in \mathcal{S}_b^{(n)})$, for $n = 1,2,...,N-1,$, and for $-\infty < x < \infty$,

$$\boldsymbol{\pi}^{(n)}(x) = (\pi_j^{(n)}(x) : j \in \mathcal{S}^{(n)}), \text{ for } n = 1,2,...,N. \tag{5.2}$$

In the rest of the section, we focus primarily on the joint density function

$\boldsymbol{\pi}^{(n)}(x)=(\pi_j^{(n)}(x): j\in\mathcal{S}^{(n)})$, for $n=1,2,...,N$, and the border probabilities $\{\mathbf{p}^{(n)}, n=1,2,...,N-1\}$. Our analysis consists of five steps, with a separate subsection for each step.

- In Subsection 5.1, we use the semi-Markov chain theory to establish the relationship between the density function, the border probabilities, and an integral of a conditional density function;
- In Subsection 5.2, we construct a censored CTMC to find the border probabilities;
- In Subsection 5.3, we develop a linear system for the limits of the density function;
- In Subsection 5.4, we put things together to derive expressions for the joint density function; and
- In Subsection 5.5, we present the computation steps for computing the density function, distribution function, and the mean fluid level.

5.1. Density function and number of level crossing

Let

- $f_j(x,t)$ be the density at the state (x, j) at time t, given the initial state $(X(0), \phi(0))$; and
- $\gamma_{k,j}^{(n)}(y,x,t)$ be the taboo conditional density of (x, j) at time t, avoiding both Border l_{n-1} and Border l_n in the time interval $(0,t)$, given that the initial state is (y,k), for $l_{n-1}<x<l_n$ and $y=l_{n-1}$ or l_n.

We note that $f_j(x,t)h\approx P\{x<X(t)<x+h, \phi(t)=j\}$ for initial condition $(X(0), \phi(0))$, and $\gamma_{k,j}^{(n)}(y,x,t)h$ is approximately the taboo conditional probability that the fluid level is in $(x,x+h)$ at time t.

For $l_{n-1}<x<l_n$, we condition on the state at which the process is either in Border l_{n-1} or l_n for the last time before reaching state (x, j) at time t. After that time point, denoted as $t-\tau,$, the process will be between the two borders until it reaches (x, j) at t (see Figure 4). At the point $t-\tau$, the fluid level either touches one of the borders and enters into the interval (l_{n-1}, l_n) or goes from one of the two borders into the interval (l_{n-1}, l_n), a total of six cases. The corresponding probabilities for the occurrence for the six cases are given approximately as follows.

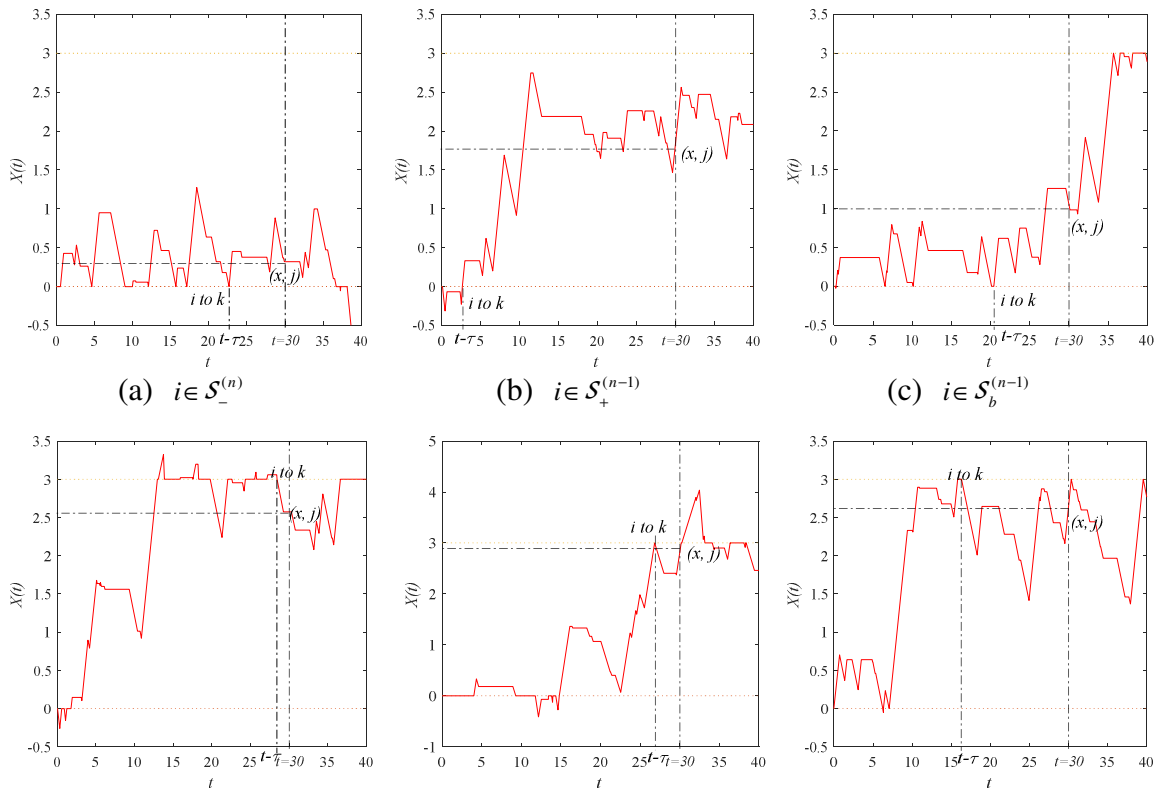


Figure 4. The fluid process in $(0,t)$ with $(X(t),\phi(t))=(x,j)$ and the time epoch $t-\tau$.

1. From $i\in\mathcal{S}_{-}^{(n)}$, the probability is $\dfrac{P_{X(0),\phi(0)}\{l_{n-1}<X(t-\tau)<l_{n-1}+c_i\mathrm{d}\tau,\phi(t-\tau)=i\}}{c_i^{(n)}\mathrm{d}\tau}c_i^{(n)}\mathrm{d}\tau$, which can be written in density function as: $f_i(l_{n-1}+,t-\tau)c_i^{(n)}\mathrm{d}\tau$. Then the process can be reflected at Border l_{n-1} at epoch $t-\tau$ with (matrix) probability $P_{-b+}^{(n-1)}$. (Remark: In state i, when the time elapses $\mathrm{d}\tau$ units, the fluid level changes by $c_i\mathrm{d}\tau$. That is why we need to use $c_i\mathrm{d}\tau$, instead of only $\mathrm{d}\tau$ in the expression.) (See Figure 4(a).)

2. From $i\in\mathcal{S}_{+}^{(n-1)}$, the probability is $\dfrac{P_{X(0),\phi(0)}\{l_{n-1}-c_i\mathrm{d}\tau<X(t-\tau)<l_{n-1},\phi(t-\tau)=i\}}{c_i^{(n-1)}\mathrm{d}\tau}c_i^{(n-1)}\mathrm{d}\tau$, which can be written in density function as $f_i(l_{n-1}-,t-\tau)c_i^{(n-1)}\mathrm{d}\tau$. Then the process can upcross Border l_{n-1} at epoch $t-\tau$ with (matrix) probability $P_{+b+}^{(n-1)}$. (See Figure 4(b).)

3. From $i\in\mathcal{S}_{b}^{(n-1)}$, the probability is $p_i^{(n-1)}$. Then the process can enter Layer n at epoch $t-\tau$ with (matrix) probability $Q_{b+}^{(n-1)}\mathrm{d}\tau$. (See Figure 4(c).)

4. From $i \in \mathcal{S}_-^{(n+1)}$, the probability is $\frac{P_{X(0),\phi(0)}\{l_n < X(t-\tau) < l_n + c_i \mathrm{d}\tau, \phi(t-\tau) = i\}}{c_i^{(n+1)} \mathrm{d}\tau} c_i^{(n+1)} \mathrm{d}\tau$, which can be written in density function as $f_i(l_n+, t-\tau) c_i^{(n+1)} \mathrm{d}\tau$. Then the process can downcross Border l_n at epoch $t-\tau$ with (matrix) probability $P_{-b-}^{(n)}$. (See Figure 4(d).)
5. From $i \in \mathcal{S}_+^{(n)}$, the probability is $\frac{P_{X(0),\phi(0)}\{l_n - c_i \mathrm{d}\tau < X(t-\tau) < l_n, \phi(t-\tau) = i\}}{c_i^{(n)} \mathrm{d}\tau} c_i^{(n)} \mathrm{d}\tau$, which can be written in density function as $f_i(l_n-, t-\tau) c_i^{(n)} \mathrm{d}\tau$. Then the process can be reflected at Border l_n at epoch $t-\tau$ with (matrix) probability $P_{+b-}^{(n)}$. (See Figure 4(e).)
6. From $i \in \mathcal{S}_b^{(n)}$, the probability is $p_i^{(n)}$. Then the process can enter Layer n at epoch $t-\tau$ with (matrix) probability $Q_{b-}^{(n)} \mathrm{d}\tau$. (See Figure 4(f).)

Using the arguments given in da Silva Soares and Latouche [22], given $(X(0), \phi(0))$, and conditioning on the state change (i.e., $i \to k$) at epoch $t-\tau$, we have, for $l_{n-1} < x < l_n$,

$$
\begin{aligned}
f_j(x,t)h = & \sum_{i \in \mathcal{S}_-^{(n)}} \sum_{k \in \mathcal{S}_+^{(n)}} \int_0^t f_i(l_{n-1}+, t-\tau) c_i^{(n)} (P_{-b+}^{(n-1)})_{i,k} \gamma_{k,j}^{(n)}(l_{n-1}, x, \tau) h \mathrm{d}\tau \\
& + \sum_{i \in \mathcal{S}_+^{(n-1)}} \sum_{k \in \mathcal{S}_+^{(n)}} \int_0^t f_i(l_{n-1}-, t-\tau) c_i^{(n-1)} (P_{+b+}^{(n-1)})_{i,k} \gamma_{k,j}^{(n)}(l_{n-1}, x, \tau) h \mathrm{d}\tau \\
& + \sum_{i \in \mathcal{S}_b^{(n-1)}} \sum_{k \in \mathcal{S}_+^{(n)}} \int_0^t p_i^{(n-1)} (Q_{b+}^{(n-1)})_{i,k} \gamma_{k,j}^{(n)}(l_{n-1}, x, \tau) h \mathrm{d}\tau \\
& + \sum_{i \in \mathcal{S}_-^{(n+1)}} \sum_{k \in \mathcal{S}_-^{(n)}} \int_0^t f_i(l_n+, t-\tau) c_i^{(n+1)} (P_{-b-}^{(n)})_{i,k} \gamma_{k,j}^{(n)}(l_n, x, \tau) h \mathrm{d}\tau \\
& + \sum_{i \in \mathcal{S}_+^{(n)}} \sum_{k \in \mathcal{S}_-^{(n)}} \int_0^t f_i(l_n-, t-\tau) c_i^{(n)} (P_{+b-}^{(n)})_{i,k} \gamma_{k,j}^{(n)}(l_n, x, \tau) h \mathrm{d}\tau \\
& + \sum_{i \in \mathcal{S}_b^{(n)}} \sum_{k \in \mathcal{S}_-^{(n)}} \int_0^t p_i^{(n)} (Q_{b-}^{(n)})_{i,k} \gamma_{k,j}^{(n)}(l_n, x, \tau) h \mathrm{d}\tau \\
& + g_j(x,t)h + o(h),
\end{aligned}
\tag{5.3}
$$

where $g_j(x,t)$ is the conditional density such that the fluid level is always in Layer n in $(0,t)$. Recall that $f_j(x,t)h \approx P\{x < X(t) < x+h, \phi(t) = j\}$ for initial condition $(X(0), \phi(0))$, and $\gamma_{j,k}^{(n)}(y,x,t)h$ is approximately the taboo conditional probability that the fluid level is in $(x, x+h)$ at time t. We have the term $o(h)$ because in a short period of time $h / c_j^{(n)}$, there can still be more than one transitions occurring. The sum of the probabilities of all those events is $o(h)$.

Let $\boldsymbol{\alpha}^{(n)}$ be the stationary distribution of $Q^{(n)}$, for $n = 1, 2, ..., N$. The mean drift of the process associated with $\{\mathbf{c}^{(n)}, Q^{(n)}\}$ is defined as $\mu_n = \boldsymbol{\alpha}^{(n)} \mathbf{c}^{(n)}$.

We assume that $\mu_1 > 0$, $\mu_N < 0$, and the process is irreducible. Then the stochastic

process is ergodic. Consequently, the joint stationary distribution exists, is given by the limit of equation (5.3), and is independent of the initial status at $t=0$. Letting $h\to 0$ and $t\to\infty$ in equation (5.3), in matrix form, we obtain:

Theorem 1. *We assume that* $\mu_1>0,\ \mu_N<0,$ *and the process is irreducible. Then the joint stationary distribution exist. For* $l_{n-1}<x<l_n$ *and* $n=1,2,...,N$, *we have*

$$\begin{aligned}\boldsymbol{\pi}^{(n)}(x)=&\left(\boldsymbol{\pi}_{-}^{(n)}(l_{n-1})C_{-}^{(n)}P_{-b+}^{(n-1)}+\boldsymbol{\pi}_{+}^{(n-1)}(l_{n-1})C_{+}^{(n-1)}P_{+b+}^{(n-1)}+\mathbf{p}^{(n-1)}Q_{b+}^{(n-1)}\right)\int_0^\infty\gamma^{(n)}(l_{n-1},x,s)\mathrm{d}s\\&+\left(\boldsymbol{\pi}_{-}^{(n+1)}(l_{n})C_{-}^{(n+1)}P_{-b-}^{(n)}+\boldsymbol{\pi}_{+}^{(n)}(l_{n})C_{+}^{(n)}P_{+b-}^{(n)}+\mathbf{p}^{(n)}Q_{b-}^{(n)}\right)\int_0^\infty\gamma^{(n)}(l_{n},x,s)\mathrm{d}s.\end{aligned}\tag{5.4}$$

(Remark: For notational convenience, we have added $\gamma^{(1)}(l_0,x,s)=0$ *and* $\gamma^{(N)}(l_N,x,s)=0$ *to the above equation. Recall that the underlying Markov chain* $\{\phi(t),t\ge 0\}$ *is irreducible when the fluid level is in between a certain layer.)*

Next, we find closed form solutions for the two integrals in the above expression. For $n=2,3,...,N$, the integrands can be approximated by: for $k\in\mathcal{S}_{+}^{(n)}$ and $j\in\mathcal{S}_{+}^{(n)}$,

$$\begin{aligned}\gamma_{k,j}^{(n)}(l_{n-1},x,s)\mathrm{d}s&\approx\left(\frac{P\{x<X(s)<x+\mathrm{d}x,l_{n-1}<X(t)<l_n,0<t<s,\phi(s)=j\mid F_0\}}{\mathrm{d}x}\right)\mathrm{d}s\\&\approx\mathbb{E}[1_{\{X(s)=x,l_{n-1}<X(t)<l_n,0<t<s,\phi(s)=j\}}\mid F_0]\frac{\mathrm{d}s}{\mathrm{d}x},\end{aligned}\tag{5.5}$$

where $F_0=\{X(0)=l_{n-1},\phi(0)=k\}$, and $1_{\{\cdot\}}$ is the indicator function. We remark that $\mathrm{d}s/(\mathrm{d}x)=1/c_j^{(n)}$ if $\phi(s)=j$. For $k\in\mathcal{S}_{+}^{(n)}$ and $j\in\mathcal{S}_{+}^{(n)}$, we obtain (abusing the notation a little bit)

$$\begin{aligned}\int_0^\infty\gamma_{k,j}^{(n)}(l_{n-1},x,s)\mathrm{d}s&\approx\frac{\sum_{m=1}^{\infty}\mathbb{E}[1_{\{X(m\times\mathrm{d}s)=x,l_{n-1}<X(t)<l_n,0<t<m\times\mathrm{d}s,\phi(m\times\mathrm{d}s)=j\}}\mid F_0]}{c_j^{(n)}}\\&\xrightarrow{\mathrm{d}s\to 0}\frac{(N_{++}^{(l_{n-1},l_n)}(x))_{k,j}}{c_j^{(n)}}\end{aligned}\tag{5.6}$$

Intuitively, the left-hand-side can be interpreted as the (conditional) total time the process visiting state (x,j). Since the fluid generated per unit time is $c_j^{(n)}$ for state j, the time to generate one unit of fluid is $1/c_j^{(n)}$. Thus, the right hand side is also the (conditional) total time the process visiting state (x,j).

For $k\in\mathcal{S}_{+}^{(n)}$ and $j\in\mathcal{S}_{-}^{(n)}$, similarly, we obtain $(N_{+-}^{(l_{n-1},l_n)}(x))_{k,j}/c_j^{(n)}$. For $k\in\mathcal{S}_{+}^{(n)}$ and $j\in\mathcal{S}_{0}^{(n)}$, the process will be in a state in $\mathcal{S}_{+}^{(n)}$ or $\mathcal{S}_{-}^{(n)}$ before entering $\mathcal{S}_{0}^{(n)}$. By conditioning on the state at such time points, we obtain, for $k\in\mathcal{S}_{+}^{(n)}$ and $j\in\mathcal{S}_{0}^{(n)}$,

$$\int_0^\infty \gamma_{k,j}^{(n)}(l_{n-1},x,s)\mathrm{d}s=\left((N_{++}^{(l_{n-1},l_n)}(x)C_+^{-1}Q_{+0}^{(n)}+N_{+-}^{(l_{n-1},l_n)}(x)C_-^{-1}Q_{-0}^{(n)})(-Q_{00}^{(n)})^{-1}\right)_{k,j}. \tag{5.7}$$

In matrix form, we obtain, for $n=2,3,...,N$,

$$\begin{aligned}&\int_0^\infty \boldsymbol{\gamma}^{(n)}(l_{n-1},x,s)\mathrm{d}s\\&=\left(N_{++}^{(l_{n-1},l_n)}(x),N_{+-}^{(l_{n-1},l_n)}(x)\right)\begin{pmatrix}(C_+^{(n)})^{-1} & 0 & (C_+^{(n)})^{-1}Q_{+0}^{(n)}(-Q_{00}^{(n)})^{-1}\\ 0 & (C_-^{(n)})^{-1} & (C_-^{(n)})^{-1}Q_{-0}^{(n)}(-Q_{00}^{(n)})^{-1}\end{pmatrix}.\end{aligned} \tag{5.8}$$

Similarly, we have, for $n=1,2,...,N-1$,

$$\begin{aligned}&\int_0^\infty \boldsymbol{\gamma}^{(n)}(l_n,x,s)\mathrm{d}s\\&=\left(\widehat{N}_{-+}^{(l_{n-1},l_n)}(x),\widehat{N}_{--}^{(l_{n-1},l_n)}(x)\right)\begin{pmatrix}(C_+^{(n)})^{-1} & 0 & (C_+^{(n)})^{-1}Q_{+0}^{(n)}(-Q_{00}^{(n)})^{-1}\\ 0 & (C_-^{(n)})^{-1} & (C_-^{(n)})^{-1}Q_{-0}^{(n)}(-Q_{00}^{(n)})^{-1}\end{pmatrix}.\end{aligned} \tag{5.9}$$

Combining equations (5.8) and (5.9) with Lemma 5, we obtain

Lemma 7. *Matrices of the integrals satisfy the following equation:*

$$\begin{aligned}&\begin{pmatrix} I & e^{\mathcal{K}^{(n)}(l_n-l_{n-1})}\Psi^{(n)}\\ e^{\widehat{\mathcal{K}}^{(n)}(l_n-l_{n-1})}\widehat{\Psi}^{(n)} & I\end{pmatrix}\begin{pmatrix}\int_0^\infty \boldsymbol{\gamma}^{(n)}(l_{n-1},x,s)\mathrm{d}s\\ \int_0^\infty \boldsymbol{\gamma}^{(n)}(l_n,x,s)\mathrm{d}s\end{pmatrix}\\&=\begin{pmatrix}e^{\mathcal{K}^{(n)}(x-l_{n-1})} & 0\\ 0 & e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\end{pmatrix}\begin{pmatrix}(C_+^{(n)})^{-1} & \Psi^{(n)}(C_-^{(n)})^{-1} & \Gamma^{(n)}\\ \widehat{\Psi}^{(n)}(C_+^{(n)})^{-1} & (C_-^{(n)})^{-1} & \widehat{\Gamma}^{(n)}\end{pmatrix}.\end{aligned} \tag{5.10}$$

where

$$\begin{aligned}\Gamma^{(n)}&=\left((C_+^{(n)})^{-1}Q_{+0}^{(n)}+\Psi^{(n)}(C_-^{(n)})^{-1}Q_{-0}^{(n)}\right)(-Q_{00}^{(n)})^{-1};\\ \widehat{\Gamma}^{(n)}&=\left(\widehat{\Psi}^{(n)}(C_+^{(n)})^{-1}Q_{+0}^{(n)}+(C_-^{(n)})^{-1}Q_{-0}^{(n)}\right)(-Q_{00}^{(n)})^{-1}.\end{aligned} \tag{5.11}$$

If $\mu_n\neq 0$, *the first matrix on the left hand side of equation (5.10) is invertible.*

According to Theorem 1, to find the joint stationary distribution, we still need the following sets of border probabilities and limits of the density function in vector form:

1. $\{\mathbf{p}^{(n)}, n=1,2,...,N\}$; (Remark: We use $\mathbf{p}^{(N)}=0$ for convenience.)
2. $\{\boldsymbol{\pi}_+^{(n)}(l_{n-1}), \boldsymbol{\pi}_-^{(n)}(l_{n-1}), \boldsymbol{\pi}_+^{(n)}(l_n), \boldsymbol{\pi}_-^{(n)}(l_n), n=1,2,...,N\}$ (To be called *density limits*).

We find those vectors in the next two subsections.

5.2. An embedded discrete time Markov chain for border transitions and a censored continuous time Markov chain for border probabilities

We want to find out, after the process leaves a border, which border it will enter next. For that purpose, we introduce two fictitious sets of states for Border l_n: (i) a set of states "above" the border: which is $\mathcal{S}_+^{(n+1)}$; and (ii) a set of states "below" the border: which is $\mathcal{S}_-^{(n)}$. The "above" set contains all the states that the *MMFF* process leaves Border l_n by increasing the fluid level. The "below" set contains all the states that the *MMFF* process leaves Border l_n by decreasing the fluid level. Plus the border states $\mathcal{S}_b^{(n)}$, we have three sets of states associated with each border. We arrange the states in the order: $(\mathcal{S}_-^{(1)}, \mathcal{S}_+^{(2)}, \mathcal{S}_-^{(2)}, \mathcal{S}_+^{(3)}, ..., \mathcal{S}_-^{(N-1)}, \mathcal{S}_+^{(N)}, \mathcal{S}_b^{(1)}, ..., \mathcal{S}_b^{(N-1)})$.

We construct a discrete time Markov chain such that the border states are absorption. The embedded discrete time Markov chain is defined at the time epochs the *MMFF* process is leaving (e.g., upcrossing, downcrossing, reflecting, and entering) a border. The transition probability matrix D of the Markov chain has the following structure:

$$D = \begin{pmatrix} A & B \\ 0 & I \end{pmatrix}, \tag{5.12}$$

where matrix A contains all the transition blocks from $\{\mathcal{S}_-^{(1)}, \mathcal{S}_+^{(2)}, \mathcal{S}_-^{(2)}, \mathcal{S}_+^{(3)}, ..., \mathcal{S}_-^{(N-1)}, \mathcal{S}_+^{(N)}\}$ to themselves, and matrix B contains all the transition blocks from $\{\mathcal{S}_-^{(1)}, \mathcal{S}_+^{(2)}, \mathcal{S}_-^{(2)}, \mathcal{S}_+^{(3)}, ..., \mathcal{S}_-^{(N-1)}, \mathcal{S}_+^{(N)}\}$ to $\{\mathcal{S}_b^{(1)}, ..., \mathcal{S}_b^{(N-1)}\}$. The transition blocks in *A* and *B* are identified explicitly as follows.

- From $\mathcal{S}_-^{(n)}$ (i.e., the set below Border l_n), the process can
  1. return to itself (i.e., $\mathcal{S}_-^{(n)}$) with probabilities in matrix $\widehat{\Psi}_{-+}^{(l_n - l_{n-1})} P_{+b-}^{(n)}$, for $n = 1, 2, ..., N-1$;
  2. go to the set above Border l_n (i.e., $\mathcal{S}_+^{(n+1)}$) with probabilities in matrix $\widehat{\Psi}_{-+}^{(l_n - l_{n-1})} P_{+b+}^{(n)}$, for $n = 1, 2, ..., N-1$;
  3. enter Border l_n (i.e., $\mathcal{S}_b^{(n)}$) with probabilities in matrix $\widehat{\Psi}_{-+}^{(l_n - l_{n-1})} P_{+bb}^{(n)}$, for $n = 1, 2, ..., N-1$;
  4. go to the set above Border l_{n-1} (i.e., $\mathcal{S}_+^{(n)}$) with probabilities in matrix $\widehat{\Lambda}_{--}^{(l_n - l_{n-1})} P_{-b+}^{(n-1)}$, for $n = 2, 3, ..., N$;
  5. go to the set below Border l_{n-1} (i.e., $\mathcal{S}_-^{(n-1)}$) with probabilities in matrix $\widehat{\Lambda}_{--}^{(l_n - l_{n-1})} P_{-b-}^{(n-1)}$, for $n = 2, 3, ..., N$; and
  6. enter Border l_{n-1} (i.e., $\mathcal{S}_b^{(n-1)}$) with probabilities in matrix $\widehat{\Lambda}_{--}^{(l_n - l_{n-1})} P_{-bb}^{(n-1)}$, for $n = 2, 3, ..., N$.

• From $\mathcal{S}_+^{(n+1)}$ (i.e., the set above Border l_n), the process can

1. return to itself (i.e., $\mathcal{S}_+^{(n+1)}$) with probabilities in matrix $\Psi_{+-}^{(l_{n+1}-l_n)}P_{-b+}^{(n)}$, for $n=1,2,...,N-1$;
2. go to the set below Border l_n (i.e., $\mathcal{S}_-^{(n)}$) with probabilities in matrix $\Psi_{+-}^{(l_{n+1}-l_n)}P_{-b-}^{(n)}$, for $n=1,2,...,N-1$;
3. enter the Border l_n (i.e., $\mathcal{S}_b^{(n)}$) with probabilities in matrix $\Psi_{+-}^{(l_{n+1}-l_n)}P_{+bb}^{(n)}$, for $n=1,2,...,N-1$;
4. go to the set above Border l_{n+1} (i.e., $\mathcal{S}_+^{(n+2)}$) with probabilities in matrix $\Lambda_{++}^{(l_{n+1}-l_n)}P_{+b+}^{(n+1)}$, for $n=1,2,...,N-2$,
5. go to the set below Border l_{n+1} (i.e., $\mathcal{S}_-^{(n+1)}$) with probabilities in matrix $\Lambda_{++}^{(l_{n+1}-l_n)}P_{+b-}^{(n+1)}$, for $n=1,2,...,N-2$; and
6. enter Border l_{n+1} (i.e., $\mathcal{S}_b^{(n+1)}$) with probabilities in matrix $\Lambda_{++}^{(l_{n+1}-l_n)}P_{+bb}^{(n+1)}$, for $n=1,2,...,N-2$.

It is easy to see that $(I-A)^{-1}B$ contains the absorption probabilities from those "above" or "below" sets to the border sets. Let

$$(I-A)^{-1}B = \begin{matrix} \\ \mathcal{S}_-^{(m)} \\ \mathcal{S}_+^{(m+1)} \\ \vdots \end{matrix} \begin{matrix} \mathcal{S}_b^{(n)} \\ \vdots \end{matrix} \begin{pmatrix} \cdots & H_{-b}^{(m,n)} & \cdots \\ \cdots & H_{+b}^{(m,n)} & \cdots \\ & \vdots & \end{pmatrix}, \tag{5.13}$$

where $H_{-b}^{(m,n)}$ contains the probabilities that the first border entered by the original *MMFF* process, started from the set below Border l_m, is $\mathcal{S}_b^{(n)}$, and $H_{+b}^{(m,n)}$ contains the probabilities that the first border reached by the original *MMFF* process, started from the set above Border l_m, is $\mathcal{S}_b^{(n)}$.

Now, we construct a censored continuous time Markov chain $Q_{\mathbf{p}}$ for the border probabilities $\{\mathbf{p}^{(n)}, n=1,2,...,N-1\}$. The CTMC $Q_{\mathbf{p}}$ is obtained by censoring out the periods that the original *MMFF* process is between borders. Thus, the state space of $Q_{\mathbf{p}}$ constitutes (only) all the border states $\mathcal{S}_b^{(1)} \bigcup \mathcal{S}_b^{(2)} \bigcup ... \bigcup \mathcal{S}_b^{(N-1)}$. The infinitesimal generator $Q_{\mathbf{p}}$ can be divided into blocks as follow:

$$Q_{\mathbf{p}} = \begin{matrix} \\ \mathcal{S}_b^{(m)} \end{matrix} \begin{matrix} \mathcal{S}_b^{(n)} \\ \vdots \end{matrix} \begin{pmatrix} \cdots & Q_{m,n} & \cdots \\ & \vdots & \end{pmatrix}, \tag{5.14}$$

where, for $m,n=1,2,...,N-1$,

$$Q_{m,n}=\begin{cases} Q_{bb}^{(m)}+Q_{b-}^{(m)}H_{-b}^{(m,m)}+Q_{b+}^{(m)}H_{+b}^{(m,m)}, & \text{if } m=n; \\ Q_{b-}^{(m)}H_{-b}^{(m,n)}+Q_{b+}^{(m)}H_{+b}^{(m,n)}, & \text{if } m\neq n. \end{cases} \tag{5.15}$$

Lemma 8. *Border probabilities* $\{\mathbf{p}^{(n)},n=1,2,...,N-1\}$ *satisfies* $(\mathbf{p}^{(1)},\mathbf{p}^{(2)},...,\mathbf{p}^{(N-1)})\,Q_{\mathbf{p}}=0.$

In Algorithms I and II, we solve the linear system $\mathbf{x}Q_{\mathbf{p}}=0$ and $\mathbf{xe}=1$ for vector $\mathbf{x}$ first. Then we normalize vector $\mathbf{x}$ to determine the border probabilities $\{\mathbf{p}^{(n)},n=1,2,...,N-1\}$.

5.3. A linear system for density limits

Next, we introduce a linear system to find the density limits $\{\boldsymbol{\pi}_{+}^{(n)}(l_{n-1}),\boldsymbol{\pi}_{-}^{(n)}(l_{n-1}),$ $\boldsymbol{\pi}_{+}^{(n)}(l_n),\boldsymbol{\pi}_{-}^{(n)}(l_n),n=1,2,...,N\}$. Recall that $\boldsymbol{\pi}^{(n)}(x)=(\boldsymbol{\pi}_{+}^{(n)}(x),\boldsymbol{\pi}_{-}^{(n)}(x),\boldsymbol{\pi}_{0}^{(n)}(x))$. Thus, to find the density limits, it is sufficient to determine $\{\boldsymbol{\pi}^{(n)}(l_{n-1}),\boldsymbol{\pi}^{(n)}(l_n)\}$. For all borders $n=1,2,...,N-1$, we have two equations: one for $\boldsymbol{\pi}^{(n)}(l_n)$ and one for $\boldsymbol{\pi}^{(n+1)}(l_n)$. Alternatively, for each density function $\boldsymbol{\pi}^{(n)}(x)$, for $n=1,2,...,N-1$, there are two equations: one for $\boldsymbol{\pi}^{(n)}(l_{n-1})$ (except for $n=1$) and one for $\boldsymbol{\pi}^{(n)}(l_n)$ (except for $n=N$). Therefore, we have in total $2N-2$ equations for $2N-2$ unknown vectors. Denote by

$$\begin{aligned} &\mathbf{v}_{L}^{(n)}=\boldsymbol{\pi}^{(n)}(l_{n-1}), \quad \text{for } n=2,3,...,N; \\ &\mathbf{v}_{U}^{(n)}=\boldsymbol{\pi}^{(n)}(l_n), \quad \text{for } n=1,2,...,N-1. \end{aligned} \tag{5.16}$$

Vector $\mathbf{v}_{L}^{(n)}$ can be divided into three subvectors as $\mathbf{v}_{L}^{(n)}=(\mathbf{v}_{L,+}^{(n)},\mathbf{v}_{L,-}^{(n)},\mathbf{v}_{L,0}^{(n)})$, where $\mathbf{v}_{L,+}^{(n)}=\boldsymbol{\pi}_{+}^{(n)}(l_{n-1})$, $\mathbf{v}_{L,-}^{(n)}=\boldsymbol{\pi}_{-}^{(n)}(l_{n-1})$, and $\mathbf{v}_{L,0}^{(n)}=\boldsymbol{\pi}_{0}^{(n)}(l_{n-1})$. The same relationship holds between $\mathbf{v}_{U}^{(n)}$ and $\boldsymbol{\pi}^{(n)}(l_n)$: $\mathbf{v}_{U}^{(n)}=(\mathbf{v}_{U,+}^{(n)},\mathbf{v}_{U,-}^{(n)},\mathbf{v}_{U,0}^{(n)})=(\boldsymbol{\pi}_{+}^{(n)}(l_n),\boldsymbol{\pi}_{-}^{(n)}(l_n),\boldsymbol{\pi}_{0}^{(n)}(l_n))$. Denote by $\mathbf{v}^{(n)}=(\mathbf{v}_{L}^{(n)},\mathbf{v}_{U}^{(n)})$, for $n=1,2,...,N$. Note that we define $\mathbf{v}_{L}^{(1)}=0$ and $\mathbf{v}_{U}^{(N)}=0$. We define vectors:

$$\begin{aligned} \Xi_1^{(n)}&=\mathbf{p}^{(n-1)}Q_{b+}^{(n-1)}\int_0^{\infty}\gamma^{(n)}(l_{n-1},l_{n-1},s)\mathrm{d}s; \\ \Xi_2^{(n)}&=\mathbf{p}^{(n-1)}Q_{b+}^{(n-1)}\int_0^{\infty}\gamma^{(n)}(l_{n-1},l_n,s)\mathrm{d}s; \\ \Xi_3^{(n)}&=\mathbf{p}^{(n)}Q_{b-}^{(n)}\int_0^{\infty}\gamma^{(n)}(l_n,l_{n-1},s)\mathrm{d}s; \\ \Xi_4^{(n)}&=\mathbf{p}^{(n)}Q_{b-}^{(n)}\int_0^{\infty}\gamma^{(n)}(l_n,l_n,s)\mathrm{d}s. \end{aligned} \tag{5.17}$$

All the above four vectors are of size $|\mathcal{S}^{(n)}|$. We define matrices

$$
\begin{aligned}
&M_1^{(n)}=\begin{pmatrix} C_+^{(n-1)}P_{+b+}^{(n-1)} \\ 0 \\ 0 \end{pmatrix}\int_0^\infty \gamma^{(n)}(l_{n-1},l_{n-1},s)\mathrm{d}s;\ M_2^{(n)}=\begin{pmatrix} C_+^{(n-1)}P_{+b+}^{(n-1)} \\ 0 \\ 0 \end{pmatrix}\int_0^\infty \gamma^{(n)}(l_{n-1},l_n,s)\mathrm{d}s;\\
&M_3^{(n)}=\begin{pmatrix} 0 \\ C_-^{(n)}P_{-b+}^{(n-1)} \\ 0 \end{pmatrix}\int_0^\infty \gamma^{(n)}(l_{n-1},l_{n-1},s)\mathrm{d}s;\ \ M_4^{(n)}=\begin{pmatrix} 0 \\ C_-^{(n)}P_{-b+}^{(n-1)} \\ 0 \end{pmatrix}\int_0^\infty \gamma^{(n)}(l_{n-1},l_n,s)\mathrm{d}s;\\
&M_5^{(n)}=\begin{pmatrix} C_+^{(n)}P_{+b-}^{(n)} \\ 0 \\ 0 \end{pmatrix}\int_0^\infty \gamma^{(n)}(l_n,l_{n-1},s)\mathrm{d}s;\qquad M_6^{(n)}=\begin{pmatrix} C_+^{(n)}P_{+b-}^{(n)} \\ 0 \\ 0 \end{pmatrix}\int_0^\infty \gamma^{(n)}(l_n,l_n,s)\mathrm{d}s;\\
&M_7^{(n)}=\begin{pmatrix} 0 \\ C_-^{(n+1)}P_{-b-}^{(n)} \\ 0 \end{pmatrix}\int_0^\infty \gamma^{(n)}(l_n,l_{n-1},s)\mathrm{d}s;\quad M_8^{(n)}=\begin{pmatrix} 0 \\ C_-^{(n+1)}P_{-b-}^{(n)} \\ 0 \end{pmatrix}\int_0^\infty \gamma^{(n)}(l_n,l_n,s)\mathrm{d}s.
\end{aligned}
\tag{5.18}
$$

Matrices $M_1^{(n)}$ and $M_2^{(n)}$ are $|\mathcal{S}^{(n-1)}|$ by $|\mathcal{S}^{(n)}|$ matrices; $M_3^{(n)}$, $M_4^{(n)}$, $M_5^{(n)}$, and $M_6^{(n)}$ are $|\mathcal{S}^{(n)}|$ by $|\mathcal{S}^{(n)}|$ matrices; and $M_7^{(n)}$ and $M_8^{(n)}$ are $|\mathcal{S}^{(n+1)}|$ by $|\mathcal{S}^{(n)}|$ atrices.

By Theorem 1, the following linear system for the density limits can be established,

$$
\begin{aligned}
&\mathbf{v}_U^{(1)}=\mathbf{v}_L^{(2)}M_8^{(1)}+\mathbf{v}_U^{(1)}M_6^{(1)}+\Xi_4^{(1)};\\
&\mathbf{v}_L^{(n)}=\mathbf{v}_L^{(n)}M_3^{(n)}+\mathbf{v}_U^{(n-1)}M_1^{(n)}+\Xi_1^{(n)}+\mathbf{v}_L^{(n+1)}M_7^{(n)}+\mathbf{v}_U^{(n)}M_5^{(n)}+\Xi_3^{(n)},\ n=2,...,N-1;\\
&\mathbf{v}_U^{(n)}=\mathbf{v}_L^{(n)}M_4^{(n)}+\mathbf{v}_U^{(n-1)}M_2^{(n)}+\Xi_2^{(n)}+\mathbf{v}_L^{(n+1)}M_8^{(n)}+\mathbf{v}_U^{(n)}M_6^{(n)}+\Xi_4^{(n)},\ n=2,...,N-1;\\
&\mathbf{v}_L^{(N)}=\mathbf{v}_L^{(N)}M_3^{(N)}+\mathbf{v}_U^{(N-1)}M_1^{(N)}+\Xi_1^{(N)}.
\end{aligned}
\tag{5.19}
$$

Further, the above linear system can be written as follows:

$$
\begin{aligned}
&\mathbf{v}^{(1)}=\left(0,\ \Xi_4^{(1)}(I-M_6^{(1)})^{-1}\right)+\mathbf{v}^{(2)}\begin{pmatrix} 0 & M_8^{(1)}(I-M_6^{(1)})^{-1} \\ 0 & 0 \end{pmatrix};\\
&\mathbf{v}^{(n)}=\left(\begin{array}{l}\left(\Xi_1^{(n)}+\Xi_3^{(n)},\Xi_2^{(n)}+\Xi_4^{(n)}\right)+\mathbf{v}^{(n-1)}\begin{pmatrix} 0 & 0 \\ M_1^{(n)} & M_2^{(n)} \end{pmatrix}\\ \qquad +\mathbf{v}^{(n+1)}\begin{pmatrix} M_7^{(n)} & M_8^{(n)} \\ 0 & 0 \end{pmatrix}\end{array}\right)\begin{pmatrix} I-M_3^{(n)} & -M_4^{(n)} \\ -M_5^{(n)} & I-M_6^{(n)} \end{pmatrix}^{-1},\ n=2,...,N-1;\\
&\mathbf{v}^{(N)}=\left(\Xi_1^{(N)}(I-M_3^{(N)})^{-1},0\right)+\mathbf{v}^{(N-1)}\begin{pmatrix} 0 & 0 \\ M_1^{(N)}(I-M_3^{(N)})^{-1} & 0 \end{pmatrix}.
\end{aligned}
\tag{5.20}
$$

Define

$$\begin{aligned}
&\Delta_1 = \left(0, \Xi_4^{(1)}(I - M_6^{(1)})^{-1}\right);\\
&\Delta_n = \left(\Xi_1^{(n)} + \Xi_3^{(n)}, \Xi_2^{(n)} + \Xi_4^{(n)}\right)\begin{pmatrix} I - M_3^{(n)} & -M_4^{(n)} \\ -M_5^{(n)} & I - M_6^{(n)} \end{pmatrix}^{-1}, n = 2,..., N-1; \qquad (5.21)\\
&\Delta_N = \left(\Xi_1^{(N)}(I - M_3^{(N)})^{-1}, 0\right),
\end{aligned}$$

$$\begin{aligned}
&\Phi_{1,2} = \begin{pmatrix} 0 & M_8^{(1)}(I - M_6^{(1)})^{-1} \\ 0 & 0 \end{pmatrix};\\
&\Phi_{n,n-1} = \begin{pmatrix} 0 & 0 \\ M_1^{(n)} & M_2^{(n)} \end{pmatrix}\begin{pmatrix} I - M_3^{(n)} & -M_4^{(n)} \\ -M_5^{(n)} & I - M_6^{(n)} \end{pmatrix}^{-1};\\
&\Phi_{n,n+1} = \begin{pmatrix} M_7^{(n)} & M_8^{(n)} \\ 0 & 0 \end{pmatrix}\begin{pmatrix} I - M_3^{(n)} & -M_4^{(n)} \\ -M_5^{(n)} & I - M_6^{(n)} \end{pmatrix}^{-1};\\
&\Phi_{N,N-1} = \begin{pmatrix} 0 & 0 \\ M_1^{(1)}(I - M_3^{(N)})^{-1} & 0 \end{pmatrix},
\end{aligned} \qquad (5.22)$$

and

$$\begin{aligned}
&\Upsilon_N = \Phi_{N,N-1};\\
&\Upsilon_n = \Phi_{n,n-1}\left(I - \Upsilon_{n+1}\Phi_{n,n+1}\right)^{-1}, \text{ for } n = 2,3,..., N-1;\\
&\widetilde{\Delta}_N = \Delta_N;\\
&\widetilde{\Delta}_n = (\Delta_n + \widetilde{\Delta}_{n+1}\Phi_{n,n+1})(I - \Upsilon_{n+1}\Phi_{n,n+1})^{-1}, \text{ for } n = 2,3,..., N-1.
\end{aligned} \qquad (5.23)$$

Finally, we obtain

$$\begin{aligned}
&(\mathbf{v}_L^{(1)}, \mathbf{v}_U^{(1)}) = \mathbf{v}^{(1)} = (\Delta_1 + \widetilde{\Delta}_2\Phi_{1,2})\left(I - \Upsilon_2\Phi_{1,2}\right)^{-1};\\
&(\mathbf{v}_L^{(n)}, \mathbf{v}_U^{(n)}) = \mathbf{v}^{(n)} = \widetilde{\Delta}_n + \mathbf{v}^{(n-1)}\Upsilon_n, \text{ for } n = 2,3,..., N,
\end{aligned} \qquad (5.24)$$

which can be used to compute the density limits recursively. We note that the invertibility of the matrices in the above equations is guaranteed by the existence of the joint stationary distribution.

5.4. Putting things together and results

Now, we assume that $\mu_1 > 0$ and $\mu_N < 0$. Then the stationary distribution of the *MMFF* process exists. With all the vectors in place, we now find the expressions for the

density functions and distribution functions. By Theorem 1, we can write

$$\boldsymbol{\pi}^{(n)}(x)=\mathbf{w}_L^{(n)}\int_0^\infty \gamma^{(n)}(l_{n-1},x,s)\mathrm{d}s+\mathbf{w}_U^{(n)}\int_0^\infty \gamma^{(n)}(l_n,x,s)\mathrm{d}s. \tag{5.25}$$

where $\mathbf{w}_L^{(1)}=0,\ \ \mathbf{w}_U^{(N)}=0,$ and for $n=1,2,...,N-1,$

$$\begin{aligned}\mathbf{w}_U^{(n)}&=\boldsymbol{\pi}_-^{(n+1)}(l_n)C_-^{(n+1)}P_{-b-}^{(n)}+\boldsymbol{\pi}_+^{(n)}(l_n)C_+^{(n)}P_{+b-}^{(n)}+\mathbf{p}^{(n)}Q_{b-}^{(n)}\\&=\mathbf{v}_{L,-}^{(n+1)}C_-^{(n+1)}P_{-b-}^{(n)}+\mathbf{v}_{U,+}^{(n)}C_+^{(n)}P_{+b-}^{(n)}+\mathbf{p}^{(n)}Q_{b-}^{(n)};\\\mathbf{w}_L^{(n+1)}&=\boldsymbol{\pi}_-^{(n+1)}(l_n)C_-^{(n+1)}P_{-b+}^{(n)}+\boldsymbol{\pi}_+^{(n)}(l_n)C_+^{(n)}P_{+b+}^{(n)}+\mathbf{p}^{(n)}Q_{b+}^{(n)}\\&=\mathbf{v}_{L,-}^{(n+1)}C_-^{(n+1)}P_{-b+}^{(n)}+\mathbf{v}_{U,+}^{(n)}C_+^{(n)}P_{+b+}^{(n)}+\mathbf{p}^{(n)}Q_{b+}^{(n)};\end{aligned} \tag{5.26}$$

Then we define, for $n=1,2,...,N,$

$$(\mathbf{u}_+^{(n)},\mathbf{u}_-^{(n)})=(\mathbf{w}_L^{(n)},\mathbf{w}_U^{(n)})\begin{pmatrix} I & e^{\mathcal{K}^{(n)}(l_n-l_{n-1})}\Psi^{(n)}\\ e^{\widehat{\mathcal{K}}^{(n)}(l_n-l_{n-1})}\widehat{\Psi}^{(n)} & I\end{pmatrix}^{-1}. \tag{5.27}$$

Combining equations (5.25)-(5.27) and Lemma 7, we obtain a closed form expression of the joint density function.

Theorem 2. *We assume that* $\mu_1>0,\ \mu_N<0,$ *and* $\mu_n\neq 0$[1] *for* $n=2,3,...,N-1.$ *For* $n=1,2,...,N,$ *we have, for* $l_{n-1}<x<l_n,$

$$\begin{aligned}\boldsymbol{\pi}^{(n)}(x)=&\,\mathbf{u}_+^{(n)}e^{\mathcal{K}^{(n)}(x-l_{n-1})}((C_+^{(n)})^{-1},\Psi^{(n)}(C_-^{(n)})^{-1},\Gamma^{(n)})\\&+\mathbf{u}_-^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}(\widehat{\Psi}^{(n)}(C_+^{(n)})^{-1},(C_-^{(n)})^{-1},\widehat{\Gamma}^{(n)}).\end{aligned} \tag{5.28}$$

Now, we construct the joint stationary distribution function. Let $\mathbf{G}^{(n)}(x)=\int_{l_{n-1}}^x \boldsymbol{\pi}^{(n)}(x)\mathrm{d}x.$ We obtain, for $l_{n-1}<x<l_n$ and $n=1,2,...,N,$

$$\begin{aligned}\mathbf{G}^{(n)}(x)=&\,\mathbf{u}_+^{(n)}\int_{l_{n-1}}^x e^{\mathcal{K}^{(n)}(y-l_{n-1})}\mathrm{d}y\left((C_+^{(n)})^{-1},\Psi^{(n)}(C_-^{(n)})^{-1},\Gamma^{(n)}\right)\\&+\mathbf{u}_-^{(n)}\int_{l_{n-1}}^x e^{\widehat{\mathcal{K}}^{(n)}(l_n-y)}\mathrm{d}y\left(\widehat{\Psi}^{(n)}(C_+^{(n)})^{-1},(C_-^{(n)})^{-1},\widehat{\Gamma}^{(n)}\right).\end{aligned} \tag{5.29}$$

Finally, we need to normalize the coefficients in the joint density function and the joint distribution function. By the law of total probability, the normalization factor is given by

[1] We note that results for the case with $\mu_n=0$ for some $n=2,3,...,N-1$ are much more involved. We choose not to touch that case. Yet it is an interesting topic for future research.

$$
\begin{aligned}
c_{norm} = & \sum_{n=1}^{N-1} \mathbf{p}^{(n)}\mathbf{e} + \sum_{n=1}^{N} \mathbf{u}_{+}^{(n)} \int_{l_{n-1}}^{l_n} e^{\mathcal{K}^{(n)}(y-l_{n-1})} \mathrm{d}y \left((C_{+}^{(n)})^{-1}, \Psi^{(n)} (C_{-}^{(n)})^{-1}, \Gamma^{(n)} \right) \mathbf{e} \\
& + \sum_{n=1}^{N} \mathbf{u}_{-}^{(n)} \int_{l_{n-1}}^{l_n} e^{\widehat{\mathcal{K}}^{(n)}(l_n-y)} \mathrm{d}y \left(\widehat{\Psi}^{(n)} (C_{+}^{(n)})^{-1}, (C_{-}^{(n)})^{-1}, \widehat{\Gamma}^{(n)} \right) \mathbf{e}.
\end{aligned} \tag{5.30}
$$

Many quantities of interest can then be obtained. For example, the (steady state) mean fluid level can be obtaned as:

$$
\begin{aligned}
\mathbb{E}[X(t)] = & \sum_{n=1}^{N-1} l_n \mathbf{p}^{(n)}\mathbf{e} + \sum_{n=1}^{N} \int_{l_{n-1}}^{l_n} x \mathrm{d}\mathbf{G}^{(n)}(x)\mathbf{e} \\
= & \sum_{n=1}^{N-1} l_n \mathbf{p}^{(n)}\mathbf{e} + \sum_{n=1}^{N} \mathbf{u}_{+}^{(n)} \int_{l_{n-1}}^{l_n} y e^{\mathcal{K}^{(n)}(y-l_{n-1})} \mathrm{d}y \left((C_{+}^{(n)})^{-1}, \Psi^{(n)} (C_{-}^{(n)})^{-1}, \Gamma^{(n)} \right) \mathbf{e} \\
& + \sum_{n=1}^{N} \mathbf{u}_{-}^{(n)} \int_{l_{n-1}}^{l_n} y e^{\widehat{\mathcal{K}}^{(n)}(l_n-y)} \mathrm{d}y \left(\widehat{\Psi}^{(n)} (C_{+}^{(n)})^{-1}, (C_{-}^{(n)})^{-1}, \widehat{\Gamma}^{(n)} \right) \mathbf{e}.
\end{aligned} \tag{5.31}
$$

The integrals in equations (5.29)-(5.31) can be evaluated by using expressions in the results given in the next subsection (Lemma 9).

5.5. Computation details, Algorithm I, and numerical examples

First, we present a lemma that can be used for computing the distribution function, mean, and higher moments of the fluid flow processes. Let $\mathbf{v}_L$ and $\mathbf{v}_R$ be the left and right eigenvectors, corresponding to eigenvalue zero, of a matrix M, i.e., $\mathbf{v}_L M = 0$ and $M\mathbf{v}_R = 0$, and are normalized by $\mathbf{v}_L\mathbf{e} = 1$ and $\mathbf{v}_L\mathbf{v}_R = 1$. It can be shown that $M - \mathbf{v}_R\mathbf{v}_L$ is invertible. Define

$$
\begin{aligned}
& \mathcal{L}_{a,b}^{M} = \int_a^b \exp(M(x-a))\mathrm{d}x; \quad \widetilde{\mathcal{L}}_{a,b}^{M} = \int_a^b \exp(M(b-x))\mathrm{d}x; \\
& \mathcal{M}_{a,b}^{M} = \int_a^b x\exp(M(x-a))\mathrm{d}x; \ \widetilde{\mathcal{M}}_{a,b}^{M} = \int_a^b x\exp(M(b-x))\mathrm{d}x.
\end{aligned} \tag{5.32}
$$

Lemma 9. *Assume that* $-\infty < a < b < \infty$. *If matrix* M *is invertible, we have*

$$
\begin{aligned}
& \mathcal{L}_{a,b}^{M} = \widetilde{\mathcal{L}}_{a,b}^{M} = M^{-1}(e^{M(b-a)} - I); \\
& \mathcal{M}_{a,b}^{M} = M^{-1}\left(M^{-1} - aI + (bI - M^{-1})e^{M(b-a)} \right); \\
& \widetilde{\mathcal{M}}_{a,b}^{M} = M^{-1}\left(-M^{-1} - bI + (aI + M^{-1})e^{M(b-a)} \right).
\end{aligned} \tag{5.33}
$$

If matrix M *is non-invertible, we have*

$$\begin{aligned}
\mathcal{L}_{a,b}^{M} &= \widetilde{\mathcal{L}}_{a,b}^{M} = (M-\mathbf{v}_R\mathbf{v}_L)^{-1}\left(e^{M(b-a)}-I\right)+(b-a)\mathbf{v}_R\mathbf{v}_L;\\
\mathcal{M}_{a,b}^{M} &= (M-\mathbf{v}_R\mathbf{v}_L)^{-1}\left(be^{M(b-a)}-aI\right)+\frac{(b^2-a^2)}{2}\mathbf{v}_R\mathbf{v}_L\\
&\quad -(M-\mathbf{v}_R\mathbf{v}_L)^{-2}\left(e^{M(b-a)}-I\right)+(b-a)\mathbf{v}_R\mathbf{v}_L;\\
\widetilde{\mathcal{M}}_{a,b}^{M} &= (M-\mathbf{v}_R\mathbf{v}_L)^{-1}\left(ae^{M(b-a)}-bI\right)+\frac{(b^2-a^2)}{2}\mathbf{v}_R\mathbf{v}_L\\
&\quad +(M-\mathbf{v}_R\mathbf{v}_L)^{-2}\left(e^{M(b-a)}-I\right)-(b-a)\mathbf{v}_R\mathbf{v}_L.
\end{aligned} \tag{5.34}$$

Proof. First, we consider $\mathcal{L}_{a,b}^{M}$ and $\mathcal{M}_{a,b}^{M}$. It is easy to obtain:

$$\begin{aligned}
&M\int_a^b e^{M(x-a)}\mathrm{d}x = \int_a^b \mathrm{d}e^{M(x-a)} = e^{M(b-a)}-I;\\
&M\int_a^b xe^{M(x-a)}\mathrm{d}x = \int_a^b x\mathrm{d}e^{M(x-a)} = be^{M(b-a)}-aI-\int_a^b e^{M(x-a)}\mathrm{d}x;\\
&\mathbf{v}_R\mathbf{v}_L\int_a^b e^{M(x-a)}\mathrm{d}x = \mathbf{v}_R\mathbf{v}_L(b-a);\\
&\mathbf{v}_R\mathbf{v}_L\int_a^b xe^{M(x-a)}\mathrm{d}x = \mathbf{v}_R\mathbf{v}_L\frac{(b^2-a^2)}{2}.
\end{aligned} \tag{5.35}$$

If M is invertible, the results are obtained directly from the first two equalities in the above equation. If M is non-invertible, then we use the fact that $M-\mathbf{v}_R\mathbf{v}_L$ is invertible. The results for $\mathcal{L}_{a,b}^{M}$ are obtained by routine calculations using all the equalities in the above equation.

Results for $\widetilde{\mathcal{L}}_{a,b}^{M}$ and $\widetilde{\mathcal{M}}_{a,b}^{M}$ can be obtained similarly.

We summarize the computational steps for the joint density function in Algorithm I.

Algorithm I.

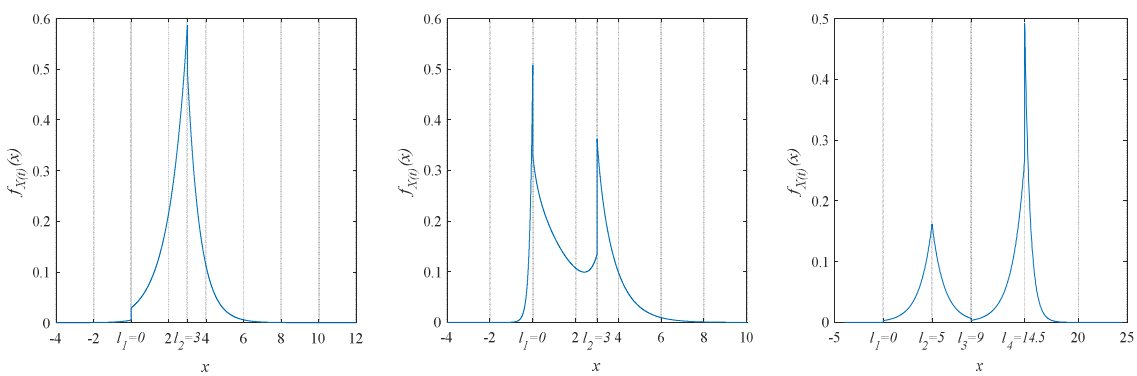
1. Input Parameters: $\{l_0=-\infty, l_1, ..., l_{N-1}, l_N=\infty\}$, $\{Q^{(n)}, C_+^{(n)}, C_-^{(n)}\}$, $n=1,2,...,N$, and $\{P_{+b+}^{(n)},\ P_{+bb}^{(n)},\ P_{+b-}^{(n)},\ P_{-b+}^{(n)},\ P_{-bb}^{(n)},\ P_{-b-}^{(n)},\ Q_{bb}^{(n)},\ Q_{b+}^{(n)},\ Q_{b-}^{(n)}\}$, for $n=1,2,...,N-1$.
2. Compute $\{\Psi^{(n)}, \mathcal{K}^{(n)}, \mathcal{U}^{(n)}, \widehat{\Psi}^{(n)}, \widehat{\mathcal{K}}^{(n)}, \widehat{\mathcal{U}}^{(n)}\}$ for $\{Q^{(n)}, C_+^{(n)}, C_-^{(n)}\}$ by using equations in Section 4, for $n=1,2,...,N$; Compute $\{\Gamma^{(n)}, \widehat{\Gamma}^{(n)}\}$ by equation (5.11) for $n=1,2,...,N$;
3. Compute $\{\Psi_{+-}^{(l_n-l_{n-1})}, \widehat{\Psi}_{-+}^{(l_n-l_{n-1})}, \Lambda_{++}^{(l_n-l_{n-1})}, \widehat{\Lambda}_{--}^{(l_n-l_{n-1})}\}$ for $\{Q^{(n)}, C_+^{(n)}, C_-^{(n)}\}$, $n=1,2,...,N$, by using equation (4.13);
4. Construct matrices A and B (equation (5.12)). Compute $\{H_{-b}^{(m,n)}, H_{+b}^{(m,n)}\}$ for $m,n=1,2,...,N-1$ by using equation (5.13);
5. Construct $Q_{\mathbf{p}}$ by using equations (5.14) and (5.15); Solve linear system $\mathbf{x}Q_{\mathbf{p}}=0$ and $\mathbf{xe}=1$ for $\{\mathbf{p}^{(n)}, n=1,2,...,N-1\}$;

6. Compute matrices $\{M_i^{(n)}, i=1,2,...,8, n=1,2,...,N\}$ by using Lemma 7 (equation (5.10)) and equation (5.18);
7. Use equations (5.17), (5.21), (5.22), and (5.24) to compute $\mathbf{v}_L^{(n)}$ and $\mathbf{v}_U^{(n)}$, for $n=1,2,...,N;$;
8. Compute $\{\mathbf{w}_+^{(n)}, \mathbf{w}_-^{(n)}, n=1,2,...,N\}$ by equation (5.26);
9. Compute $\{\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)}, n=1,2,...,N\}$ by equation (5.27);
10. Compute c_{norm} by using equation (5.30) and Lemma 9;
11. Use c_{norm} to normalize $\{\mathbf{p}^{(n)}, n=1,2,...,N-1\}$ and $\{\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)}, n=1,2,...,N\}$..
12. Use the updated vectors and Lemma 9 to compute the stationary distribution function (equation (5.29)), density function (equation (5.28)), and the mean fluid level (equation (5.31)).

We have tested Algorithm I extensively. Next, we present one example with all parameters. On the other hand, an *MMFF* process usually has many parameters. So, we present two more examples without parameters.

Example 5.1 (Example 3.1 continued) A sample path of the example are shown in Figure 1(b). Density function of the fluid level is shown in Figure 5(a). We have calculated the mean fluid level, which is $\mathbb{E}[X(t)]=2.7278$. For this three-layer *MMFF* process, the density function changes drastically at the two Borders $l_1=0$ and $l_2=3$. Within each layer, the density function looks like the exponential function. This is not surprising given the matrix-exponential form of the density function in equation (5.28).

Example 5.2 We also plot the density functions of the fluid level in Figure 5 for two more examples to show the variety of the density functions that can be generated by multi-layer *MMFF* processes. Figure 5(b) is for a three-layer *MMFF* process with $l_1=0$ and $l_2=3$ and mean fluid level $\mathbb{E}[X(t)]=1.8069..$ Figure 5(c) is for a five-layer *MMFF* process with $l_1=0,\ l_2=5,\ l_3=9,$ and $l_4=14.5$ and mean fluid level $\mathbb{E}[X(t)]=10.4758.$



(a) (b) (c)

Figure 5. The density functions of three multi-layer *MMFF* processes.

The set of multi-layer *MMFF* processes is a rich class of stochastic processes for which the density functions take many interesting shapes. Although the density function may look like the exponential function in individual layers, the shape of the entire function seems versatile. The algorithm works well for small/moderate size problems in terms of the number of states of the underlying Markov chain. For large size problems, the algorithm has to be modified in order to reduce the size of state space. For instance, computation of border probabilities $\{\mathbf{p}^{(n)}, n=1,2,...,N-1\}$ can face the dimensionality issue since the matrix $Q_{\mathbf{p}}$ can be too big for numerical evaluation. On the other hand, the state space used in computation can be drastically reduced for many cases by taking advantages of special structures of the multi-layer *MMFF* processes. In the second half of this paper, we use a queueing example to demonstrate how Algorithm I can be applied to analyze stochastic systems and how the state space can be reduced to make the algorithm numerically more efficient.

6. The *MAP/PH/K+GI* Queue

In this section, we apply the theory on multi-layer *MMFF* processes to analyze a queueing system with customer abandonment. In Subsection 6.1, we introduce the queueing model explicitly. In Subsection 6.2, we introduce a Markov process associated with the age of the customer at the head of the waiting queue, to be called *the age process*. Based on the age process, we introduce a multi-layer *MMFF* process in Subsection 6.3. Subsection 6.4 presents an algorithm for the stationary distribution of the age process. In Subsection 6.5, computational procedures are developed for a number of queueing quantities. Numerical examples are presented in Subsection 6.6.

6.1. Definitions of the MAP/PH/K+GI queue

We consider a multi-server queueing model with customer abandonment. Upon arrival, all customers join a single queue and are served on a first-come-first-in basis. There are K identical servers. When the waiting time of a customer reaches (random) time τ, the customer leaves the system without service.

(i) Customers arrive to the queueing system according to a continuous time Markovian arrival process (*MAP*) (D_0, D_1), where D_0 and D_1 are square matrices of order m_a. Intuitively, D_0 contains the transition rates without an arrival and D_1 contains the transition rates with one arrival. The underlying Markov chain of the arrival process $\{I_a(t), t \geq 0\}$ has an irreducible infinitesimal generator $D = D_0 + D_1$. The stationary distribution $\boldsymbol{\theta}_a$ of the underlying Markov chain satisfies $\boldsymbol{\theta}_a D = 0$ and $\boldsymbol{\theta}_a \mathbf{e} = 1$. The (average) customer arrival rate is given by $\lambda = \boldsymbol{\theta}_a D_1 \mathbf{e}$. See He [28] and Neuts [38] for more details on *MAP*s.

(ii) All customers join a single queue waiting for service and are served on a first-come-first-in basis. If a customer's waiting time reaches random time τ, the customer leaves

the system immediately without service. The abandonment time τ has a discrete distribution: $P\{\tau=l_n\}=\eta_n$, for $n=1,2,...,N$, where $l_1=0<l_2<...<l_{N-1}<l_N=\infty$.

(iii) There are *K* identical servers. When a server becomes available, the customer at the head of the queue (if there is any) enters the server for service. If an arriving customer finds an idle server, the customer enters the server for service upon arrival.

(iv) The service time of each customer has a phase-type distribution with *PH*-representation $(\boldsymbol{\beta},T)$ of order m_s. We assume that $\boldsymbol{\beta}\mathbf{e}=1$, i.e., the service time of a customer is always positive. The mean service time is given by $-\boldsymbol{\beta}T^{-1}\mathbf{e}$. Let $\mu_s=1/(-\boldsymbol{\beta}T^{-1}\mathbf{e})$, which is the service rate. See Neuts [39] for more about phase-type distributions.

(v) Define $\rho=\lambda/(K\mu_s)$. We assume $\eta_N\rho<1$ to ensure the stability of the queueing system. Since $\eta_N\lambda$ is the number of customers who arrive per unit time, who will not leave the system until service is done, and $K\mu_s$ is the number of customer that can be served per unit time, $\eta_N\rho<1$ ensures that all customers are either served or abandon the system in finite time. Consequently, the system is stable. If $\eta_N=0$, the system is automatically stable.

6.2. The Age Process

To obtain performance measures for the queueing model, we utilize a Markov process associated with the age of the customer at the head of the queue. The *age* of a customer is defined as the time elapsed since the customer enters the system. Since customers arrive according to an *MAP* and service times are of phase-type, tracking the age of the customer at the head of the queue, state of the arrival process, and states of the service processes of individual servers, provides enough information to describe the dynamics of the queueing system. Define

- $a(t)$: the age of the customer waiting at the head of the queue at time *t*, if the (waiting) queue is not empty; otherwise, $a(t)=0$ (See the top figure in Figure 6). If $l_n<a(t)<l_{n+1}$, for $n=1,2,...,N-1$, $a(t)$ increases linearly at rate one if there is no service completion; Otherwise (i.e., there is service completion), $a(t+0)=\max\{0,a(t)-u\}$, where u is the sum of the interarrival times between the customer at the head of the queue and the customer just behind it in the queue. If $a(t)=l_n$, for $n=2,3,...,N-1$, $a(t)$ continues to increase linearly at rate one with probability $1-\eta_n/(\eta_n+...+\eta_N)$; Otherwise, $a(t+0)=\max\{0,l_n-u\}$, where u is the interarrival time between the departing customer (since its waiting time reaches l_n) and the customer just behind it in the queue. The interarrival time u is the sum of interarrival times between those two consecutive customers at the head of the queue and all lost customers, if they exist, between them. By this definition, if $a(t)=0$, there is no customer waiting for service.

- $I_{(a)}(t)$: If $a(t)>0$, $I_{(a)}(t)$ is the state of the customer arrival process right after the arrival of the customer at the head of the queue; and if $a(t)=0$, the state of the arrival process at time t (i.e., $I_{(a)}(t)=I_a(t).$) By this definition, $I_{(a)}(t)$ is piece-wise constant and its value changes only when $a(t)$ drops down.

- $n_i(t)$: the number of servers whose service state is i at time t, for $i=1,2,...,m_s$.

It turns out that $\{(a(t),I_{(a)}(t),n_1(t),\ldots,n_{m_s}(t)),t\geq 0\}$, is a continuous time Markov chain whose infinitesimal generator can be constructed as in He *et al.* [30] by using the *CSFP* approach (See He and Alfa [29]). Based on the total number of working servers, the state space of $(n_1(t),\ldots,n_{m_s}(t))$ can be organized as $\Omega(0)\cup\Omega(1)\cup\ldots\cup\Omega(K)$, where, for $k=0,1,...,K$,

$$\Omega(k)=\left\{\mathbf{n}=(n_1,...,n_{m_s}):n_i\geq 0,n_i \text{ integer},i=1,...,m_s,\sum_{i=1}^{m_s}n_i=k\right\}. \qquad (6.1)$$

The set $\Omega(k)$ consists of all states such that there are exactly k customers in service (or k working servers), for $k=0,1,...,K$. The number of states in $\Omega(k)$ is given by $(k+m_s-1)!/(k!(m_s-1)!)$. Then the state space of the Markov process can be written as

$$\left\{\{0\}\times\{1,...,m_a\}\times\{\cup_{k=0}^{K}\Omega(k)\}\right\}\cup\left\{(0,\infty)\times\{1,...,m_a\}\times\Omega(K)\right\}. \qquad (6.2)$$

Remark. Instead of using $\{n_1(t),...,n_{m_s}(t)\}$, a more straightforward and simple way to model the service process is to keep track of the service process for each server (called *TPFS* in the literature) (See He *et al.* [30] or He and Alfa [29]). However, the number of states required by that approach to track the service status of the K servers (i.e., the *CSFP* method) is $O(m_s^K)$, which is significantly greater than $O\left(\binom{K+m_s-1}{m_s-1}\right)$, the number of states required by the approach used in this paper.

If $a(t)>0$, the state of the customer arrival process is frozen (i.e., constant) except for down-jump epochs. On the other hand, the states of the service processes are changing according to rate matrices $Q(K,m_s)$ for no service completion and $Q^{-}(K,m_s)P^{+}(K-1,m_s)$ for service completion. If $a(t)=0$, the arrival state and service states are all changing according to rate matrix

$$Q_{bb}^{(1)}=\begin{pmatrix} A_{0,0} & A_{0,1} & & & \\ A_{1,0} & A_{1,1} & A_{1,2} & & \\ & \ddots & \ddots & \ddots & \\ & & A_{K-1,K-2} & A_{K-1,K-1} & A_{K-1,K} \\ & & & A_{K,K-1} & (D_0+\eta_1 D_1)\otimes I+I\otimes Q(K,m_s) \end{pmatrix}. \qquad (6.3)$$

Construction of matrices $Q(K,m_s)$, $Q^{-}(K,m_s)$, $P^{+}(K-1,m_s)$, $Q_{bb}^{(1)}$, and the blocks within $Q_{bb}^{(1)}$ is complicated. An algorithm has been given in He and Alfa [29] for that purpose. Details are omitted.

Special cases of the age process have been analyzed by using an analytic method introduced in Choi *et al.* [18] (Also see Kim and Kim [33] and He *et al.* [30]). However, that approach seems not suitable for the analysis of the multi-server queue with *MAP* arrivals due to a matrix commutability issue. In this paper, based on the age process, we introduce a multi-layer *MMFF* process to solve the problem.

6.3. A multi-layer MMFF process

Now, based on the age process, we define a multi-layer *MMFF* process $\{(X(t),\phi(t)),t\geq 0\}$. The idea is conventional and is to change the down jumps of the age process into periods of decreasing fluid, keep the increasing periods of the age process for the periods of increasing fluid, and keep the periods with $a(t)=0$ for the periods with zero fluid (See the two figures in Figure 6). More specifically, we have

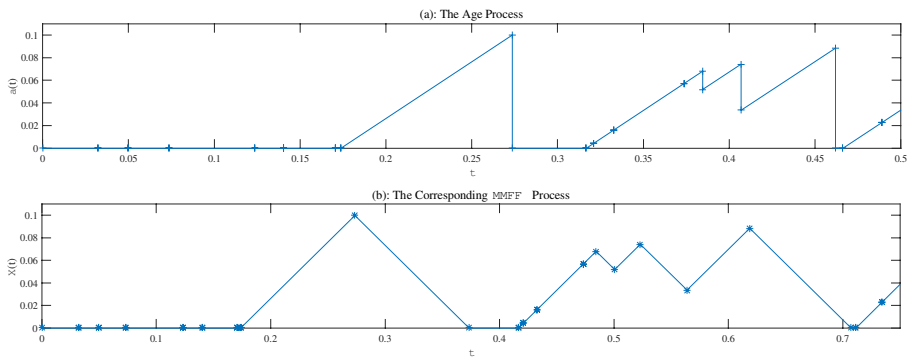


Figure 6. A sample path of the age process (top) and its corresponding *MMFF* process (bottom).

1. There are N layers with Borders l_n, for $n=1,2,...,N$. Layer 1 is empty (i.e., $\mathcal{S}^{(1)}=\varnothing$).

2. For Layer $n\geq 2$, the state space for $\phi(t)$ is:

$$\mathcal{S}_{+}^{(n)}=\{+\}\times\{1,...,m_a\}\times\Omega(K),\ \mathcal{S}_{-}^{(n)}=\{-\}\times\{1,...,m_a\}\times\Omega(K),\ \text{and}\ \mathcal{S}_{0}^{(n)}=\varnothing. \tag{6.4}$$

The Q-matrix $Q^{(n)}$ of the underlying Markov chain is:

$$Q^{(n)}=\begin{matrix}\mathcal{S}_{+}^{(n)}\\ \mathcal{S}_{-}^{(n)}\end{matrix}\begin{pmatrix} I\otimes Q(K,m_s) & I\otimes Q^{-}(K,m_s)P^{+}(K-1,m_s)\\ (\eta_n+...+\eta_N)D_1\otimes I & (\eta_1+...+\eta_{n-1})D_1\otimes I+D_0\otimes I\end{pmatrix}. \tag{6.5}$$

The fluid flow rates are all 1 or -1, i.e., $C_+^{(n)} = C_-^{(n)} = I$.

3. Within Border l_1 (i.e., $l_1 = 0$), the transition rates of the underlying Markov chain are given by equation (6.3) for $Q_{bb}^{(1)}$ and

$$Q_{b+}^{(1)} = \begin{pmatrix} 0 \\ \\ (\eta_2 + ... + \eta_N) D_1 \otimes I \end{pmatrix}; \quad Q_{b-}^{(1)} = 0. \tag{6.6}$$

4. The transition probabilities entering Border l_1 are given by (Remark: There is no Layer 1.)

$$P_{-b+}^{(1)} = 0;\ P_{-b-}^{(1)} = 0;\ P_{-bb}^{(1)} = (0, ..., 0, I). \tag{6.7}$$

When entering from Layer 2 to Border l_1, the underlying process $\phi(t)$ enters the set $\{0\} \times \{1, ..., m_a\} \times \Omega(K)$.

5. All other borders ($n > 1$) have no state. The probabilities of approaching Border l_n, for $2 \le n \le N-1$, from below are given by

$$P_{+b-}^{(n)} = \frac{\eta_n}{\eta_n + ... + \eta_N} I;\ P_{+b+}^{(n)} = \frac{\eta_{n+1} + ... + \eta_N}{\eta_n + \eta_{n+1} + ... + \eta_N} I;\ P_{+bb}^{(n)} = 0. \tag{6.8}$$

The probabilities of approaching Border l_n, for $2 \le n \le N-1$, from above are given by

$$P_{-b-}^{(n)} = I;\ P_{-b+}^{(n)} = 0;\ P_{-bb}^{(n)} = 0. \tag{6.9}$$

The joint stationary distribution of the multi-layer *MMFF* process can be obtained by using Algorithm I.

6.4. Joint stationary distribution of the age process and Algorithm II

We would like to point out that, if $\phi(t) \in \bigcup_{n=2}^N \mathcal{S}_+^{(n)}$, the service process evolves and the state of the arrival process is frozen in the multi-layer *MMFF* process, and, if $\phi(t) \in \bigcup_{n=2}^N \mathcal{S}_-^{(n)}$, the states of the service processes are frozen and the arrival process evolves. With a brief reflection of the definitions of the age process and the multi-layer *MMFF* process, it is easy to see that the age process can be obtained by censoring out states in $\bigcup_{n=2}^N \mathcal{S}_-^{(n)}$. Computations can be done using Algorithm I. However, the state space required for Algorithm I can be too large, unnecessarily, if *K* is big. Using certain special structure of the multi-layer *MMFF* process, we can modify Algorithm I and reduce the required state space for its implementation.

(i) **Border Probabilities:** Since all borders, except Border l_1, are empty, we have $\mathbf{p}^{(n)} = 0$, for $n = 2, 3, ..., N-1$. Therefore, we only have to compute $\mathbf{p}^{(1)}$, which satisfies $\mathbf{p}^{(1)} Q_{\mathbf{p}}^{(1)} = 0$, where

$$Q_{\mathbf{p}}^{(1)} = Q_{bb}^{(1)} + Q_{b+}^{(1)} \mathcal{P}_{\text{above}}^{(1)} P_{-bb}^{(1)}, \tag{6.10}$$

where $\mathcal{P}_{\text{above}}^{(1)}$ contains the first passage probabilities from the set above Border l_1 (up) to return to Border l_1 (from above), which can be computed recursively as follows. We define $\mathcal{P}_{\text{above}}^{(n)}$ the state transition probabilities that the process goes up leaving Border l_n and returns to Border l_n (from above) for the first time (i.e., starting in $\mathcal{S}_+^{(n+1)}$ and ending in $\mathcal{S}_-^{(n+1)}$). We also define $\mathcal{P}_{\text{above}}^{(n)}$ the state transition probabilities that the process goes up towards Border l_n and goes below Border l_n from above or reflecting on Border l_n for the first time (i.e., starting in $\mathcal{S}_+^{(n)}$ and ending in $\mathcal{S}_-^{(n)}$). Immediately, we have $\mathcal{P}_{\text{above}}^{(N-1)} = \Psi^{(N)}$, and for $n = 2,3,...,N-1$,

$$\begin{aligned} \mathcal{P}_{\text{below}}^{(n)} &= P_{+b-}^{(n)} + P_{+b+}^{(n)} \mathcal{P}_{\text{above}}^{(n)} \left(I - P_{-b+}^{(n)} \mathcal{P}_{\text{above}}^{(n)} \right)^{-1} P_{-b-}^{(n)}, \\ \mathcal{P}_{\text{above}}^{(n-1)} &= \Psi_{+-}^{(l_n - l_{n-1})} + \Lambda_{++}^{(l_n - l_{n-1})} \mathcal{P}_{\text{below}}^{(n)} \left(I - \widehat{\Psi}_{-+}^{(l_n - l_{n-1})} \mathcal{P}_{\text{below}}^{(n)} \right)^{-1} \widehat{\Lambda}_{--}^{(n, l_n - l_{n-1})}. \end{aligned} \tag{6.11}$$

(ii) **Vector $\mathbf{p}^{(1)}$:** Due to the special structure of $Q_{b+}^{(1)}$ and $P_{-bb}^{(1)}$, we obtain

$$Q_{\mathbf{p}}^{(1)} = \begin{pmatrix} A_{0,0} & A_{0,1} & & & \\ A_{1,0} & A_{1,1} & A_{1,2} & & \\ & \ddots & \ddots & \ddots & \\ & & A_{K-1,K-2} & A_{K-1,K-1} & A_{K-1,K} \\ & & & A_{K,K-1} & \widetilde{A}_{K,K} \end{pmatrix}, \tag{6.12}$$

where $\widetilde{A}_{K,K} = (D_0 + \eta_1 D_1) \otimes I + I \otimes Q(K, m_s) + ((\eta_2 + ... + \eta_N) D_1 \otimes I) \mathcal{P}_{\text{above}}^{(1)}$. We can explore the quasi birth-and-death (QBD) structure in $Q_{\mathbf{p}}^{(1)}$ to reduce the state space required for computing $\mathbf{p}^{(1)}$ as follows. Define

$$\begin{aligned} &B_1 = A_{1,0}(-A_{0,0})^{-1}; \\ &B_k = A_{k,k-1}(-A_{k-1,k-1} - B_{k-1} A_{k-2,k-1})^{-1}, \text{ for } k = 2,3,...,K; \end{aligned} \tag{6.13}$$

Define $Q_{\mathbf{p},K}^{(1)} = (D_0 + \eta_1 D_1) \otimes I + I \otimes Q(K, m_s) + ((\eta_2 + ... + \eta_N) D_1 \otimes I) \mathcal{P}_{\text{above}}^{(1)} + B_K A_{K-1,K}$. We also divide $\mathbf{p}^{(1)}$ according to the number of busy servers into $(\mathbf{p}_0^{(1)}, \mathbf{p}_1^{(1)}, ..., \mathbf{p}_K^{(1)})$. Then $\mathbf{p}_K^{(1)}$ satisfies $\mathbf{p}_K^{(1)} Q_{\mathbf{p},K}^{(1)} = 0$, and $\mathbf{p}_{k-1}^{(1)} = \mathbf{p}_k^{(1)} B_k$, for $k = K, K-1, ..., 1$. In the computation, we set $\mathbf{p}_K^{(1)} \mathbf{e} = 1$ and normalize the vectors later.

Denote by $\mathbf{p}_{K+1}(x)$ the joint stationary density function of the age process. Let $\mathbf{p}_{K+1}^{(n)}(x) = \mathbf{p}_{K+1}(x)$, if $l_{n-1} < x < l_n$. By Theorem 2 and censoring out $\bigcup_{n=2}^{N} \mathcal{S}_-^{(n)}$, we obtain

Theorem 3. *We assume that $\eta_N \rho < 1$ and $(\eta_n + ... + \eta_N)\rho \neq 1$ for $n = 2,3,...,N-1$. Then the steady state distribution of the age process exists and its density function is given by*

$$P\{a(t)=0\}=\sum_{k=0}^{K}\mathbf{p}_k^{(1)}\mathbf{e};$$
$$\mathbf{p}_{K+1}^{(n)}(x)=\mathbf{u}_+^{(n)}e^{\mathcal{K}^{(n)}(x-l_{n-1})}+\mathbf{u}_-^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)},\ \text{for } l_{n-1}\le x<l_n, n=2,...,N. \tag{6.14}$$

The normalization factor is (Remark: $\mathbf{u}_-^{(N)}=0$ *)*

$$\hat{c}_{norm}=\sum_{k=0}^{K}\mathbf{p}_k^{(1)}\mathbf{e}+\sum_{n=2}^{N}\int_{l_{n-1}}^{l_n}\left(\mathbf{u}_+^{(n)}e^{\mathcal{K}^{(n)}(y-l_{n-1})}+\mathbf{u}_-^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n-y)}\widehat{\Psi}^{(n)}\right)e\mathrm{d}y. \tag{6.15}$$

Proof. For the existence of the stationary of the age process, we need to show that $\eta_N\rho<1$ if and only if $\mu_N<0$. To do so, we find $\boldsymbol{\theta}$ satisfying $\boldsymbol{\theta}Q^{(n)}=0$ and $\boldsymbol{\theta}\mathbf{e}=1$. We divide $\boldsymbol{\theta}$ into $(\boldsymbol{\theta}_+,\boldsymbol{\theta}_-)$ according to $\mathcal{S}_+^{(n)}$ and $\mathcal{S}_-^{(n)}$. By routine calculations, we obtain $\boldsymbol{\theta}_+=:(\eta_n+...+\eta_N)(\boldsymbol{\theta}_aD_1)\otimes\tilde{\boldsymbol{\theta}}_s/(\boldsymbol{\theta}_+\mathbf{e}+\boldsymbol{\theta}_-\mathbf{e})$ and $\boldsymbol{\theta}_-=:\boldsymbol{\theta}_a\otimes(\tilde{\boldsymbol{\theta}}_sQ^-(K,m_s)P^+(K-1,m_s))/(\boldsymbol{\theta}_+\mathbf{e}+\boldsymbol{\theta}_-\mathbf{e})$, where $\tilde{\boldsymbol{\theta}}_s$ satisfies $\tilde{\boldsymbol{\theta}}_s(Q(K,m_s)+Q^-(K,m_s)P^+(K-1,m_s))=0$ and $\tilde{\boldsymbol{\theta}}_s\mathbf{e}=1$. It has been shown in He *et al.* [30] that $\tilde{\boldsymbol{\theta}}_sQ^-(K,m_s)P^+(K-1,m_s)\mathbf{e}=K\mu_s$ (i.e., the total service rate). Consequently, we obtain $\mu_n=\boldsymbol{\theta}_+\mathbf{e}-\boldsymbol{\theta}_-\mathbf{e}=((\eta_n+...+\eta_N)\lambda-K\mu_s)/(\boldsymbol{\theta}_+\mathbf{e}+\boldsymbol{\theta}_-\mathbf{e})$, which leads to the condition of the existence of the stationary distribution. Also, the relationship shows that $(\eta_n+...+\eta_N)\lambda-K\mu_s=0$ if and only if $\mu_n=0$. Thus, all assumptions in Theorem 2 are satisfied. The closed form solution of the density function of the age process is obtained from that of the multi-layer *MMFF* process by censoring.

Again, evaluations of integrals in the above equation can be done by applying Lemma 9. Next, we modify Algorithm I to compute the joint stationary distribution of the age process.

Algorithm II

1. Input Parameters: K, N, $\{l_1=0,l_2,...,l_N=\infty\}$, $\{\eta_1,\eta_2,...,\eta_N\}$, $\{m_a,D_0,D_1\}$, and $(\boldsymbol{\beta},T)$;
2. Construct $\{Q(K,m_s),Q^-(K,m_s),P^+(K-1,m_s),Q_{bb}^{(1)}\}$ by applying the algorithm in He and Alfa [29];
3. Construct transition blocks for the multi-layer *MMFF* process: $\{l_0=-\infty,l_1,...,l_{N-1},l_N=\infty\}$, $\{Q^{(n)},C_+^{(n)},C_-^{(n)},n=1,2,...,N\}$, and $\{P_{+b+}^{(n)}, P_{+bb}^{(n)}, P_{+b-}^{(n)}, P_{-b+}^{(n)}, P_{-bb}^{(n)}, P_{-b-}^{(n)}, Q_{bb}^{(n)}, Q_{b+}^{(n)}, Q_{b-}^{(n)}, n=1,2,...,N-1\}$, according to Subsection 6.3;
4. Similar to Steps 2 and 3 in Algorithm I, compute $\{\Psi^{(n)},\mathcal{K}^{(n)},\mathcal{U}^{(n)},\widehat{\Psi}^{(n)},\widehat{\mathcal{K}}^{(n)},\widehat{\mathcal{U}}^{(n)}\}$ for $\{Q^{(n)},C^{(n)}\}$; Compute $\{\Psi_{+-}^{(l_n-l_{n-1})},\widehat{\Psi}_{-+}^{(l_n-l_{n-1})},\Lambda_{++}^{(l_n-l_{n-1})},\widehat{\Lambda}_{--}^{(l_n-l_{n-1})}\}$ for $\{Q^{(n)},C_+^{(n)},C_-^{(n)}\}$, for $n=1,2,...,N-1;$;
5. Compute $\mathcal{P}_{\text{above}}^{(1)}$ using equation (6.11); Construct $Q_{\mathbf{p},K}^{(1)}$ using (6.13); and solve $\mathbf{p}_K^{(1)}Q_{\mathbf{p},K}^{(1)}=0$ and $\mathbf{p}_K^{(1)}\mathbf{e}=1,$, and Compute $\mathbf{p}^{(1)}$;

6. Compute $\{M_i^{(n)}, i=1,2,...,8, n=1,2,...,N\}$, $\{\mathbf{w}_+^{(n)}, \mathbf{w}_-^{(n)}, n=1,2,...,N\}$; and $\{\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)}, n=1,2,...,N\}$ by using Algorithm I;
7. Compute $\hat{c}_{norm}$ by using equation (6.15), and use $\hat{c}_{norm}$ to normalize $\{\mathbf{p}_k^{(1)}, n=0,1,...,K\}$ and $\{\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)}, n=1,2,...,N\}$;
8. Use the updated vectors and equation (6.14) to compute the density function of the age process.

We would like to point out that the above computation process can be simplified further. For example, there is no need to do Step 3. In all subsequent computations, matrices constructed in Step 2 can be used directly. The set $\mathcal{S}_0^{(n)}$ is empty. Thus, there is no need to consider them in computations. Since $C_+^{(n)}$ and $C_-^{(n)}$ are identity matrices, there is no need to construct and use them in computation.

Two particular issues related to the implementation of Algorithm II are worth mentioning. The first issue is to determine the left and right eigenvectors, corresponding to eigenvalue zero, of $\mathcal{K}^{(n)}$ and $\widehat{\mathcal{K}}^{(n)}$. The second one is about solving the Sylvester equation $AX + XB = C$. It is recommended to use Schur decomposition in combination with back-substitution, instead of Kronecker product.

6.5. Queueing quantities

Based on the joint stationary distribution of the age process, we find three sets of queueing quantities: (i) Customer abandonment/loss probabilities; (ii) Waiting times; and (iii) Queue lengths. We assume that conditions stated in Theorem 3 hold throughout this section.

Proposition 1. *The probability that a customer will eventually receive service is given by*

$$p_S = \frac{1}{\lambda}\sum_{k=0}^{K-1}\mathbf{p}_k^{(1)}(D_1 \otimes I)\mathbf{e} + \frac{1}{\lambda}\sum_{n=2}^{N}\left(\mathbf{u}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{u}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)(I \otimes Q^-(K,m_s)P^+(K-1,m_s))\mathbf{e}, \quad (6.16)$$

where $\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}$ *and* $\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}$ *are defined in Lemma 9. Then the customer abandonment probability is* $p_L = 1 - p_S$. *We decompose* p_L *into two parts: (i) loss probability* $p_{L,1}$ *of customers at the head of the waiting queue (including those customers who see no waiting queue and no available servers, and abandon the system); and ii) loss probability* $p_{L,>1}$ *of customers before reaching the head of the waiting queue. Then we obtain* $p_{L,>1} = p_L - p_{L,1}$, *and*

$$p_{L,1} = \frac{\mathbf{p}_k^{(1)}((\eta_1 D_1) \otimes I)\mathbf{e}}{\lambda} + \frac{1}{\lambda}\sum_{n=2}^{N-1}\left(\mathbf{u}_+^{(n)}e^{\mathcal{K}^{(n)}(l_n - l_{n-1})}\mathbf{e} + \mathbf{u}_-^{(n)}\widehat{\Psi}^{(n)}\mathbf{e}\right)\frac{\eta_n}{\sum_{m=n}^{N}\eta_m}. \quad (6.17)$$

Proof. By definitions, we have

$$p_S = \frac{1}{\lambda}\left(\sum_{k=0}^{K-1}\mathbf{p}_k^{(1)}(D_1 \otimes I)\mathbf{e} + \int_0^{\infty}\mathbf{p}_{K+1}(x)(I \otimes Q^-(K,m_s)P^+(K-1,m_s))\mathbf{e}dx\right). \tag{6.18}$$

We note that the numerator in equation (6.18) is the sum of transition rates that a customer enters a server for service, and the denominator in equation (6.18) is the arrival rate. Then the ratio is the percentage of customers who receive service, which is also the probability that a customer will eventually receive service. The desired expression is obtained by combining equation (6.18) and Lemma 9.

The probability for a customer who sees no waiting queue and no server available, and abandons the queue is $\mathbf{p}_k^{(1)}((\eta_1 D_1)\otimes I)\mathbf{e}/\lambda$. For a customer at the head of the queue to abandon the queue, its age must reach l_n for some $n=2,3,...,N-1$. If its age reaches l_n, its age must be greater than l_{n-1}, which occurs with probability $\eta_n+...+\eta_N$. Then the probability that it abandons the queue is $\eta_n/(\eta_n+...+\eta_N)$. Combining with the transition rate for the age to reach l_n, which is $\mathbf{p}_{K+1}(l_n)\mathbf{e}$, we obtain

$$p_{L,1} = \frac{\mathbf{p}_k^{(1)}((\eta_1 D_1)\otimes I)\mathbf{e}}{\lambda} + \frac{1}{\lambda}\sum_{n=2}^{N-1}\mathbf{p}_{K+1}(l_n)\mathbf{e}\,\frac{\eta_n}{\sum_{m=n}^{N}\eta_m}, \tag{6.19}$$

which leads to the desired result.

Proposition 2. *The distribution of waiting time W_S of customers who receive service is*

$$P\{W_S = 0\} = \frac{1}{p_S\lambda}\sum_{k=0}^{K-1}\mathbf{p}_k^{(1)}(D_1\otimes I)\mathbf{e};$$

$$\frac{dP\{W_S < x\}}{dx} = \frac{1}{p_S\lambda}\left(\mathbf{u}_+^{(n)}e^{\tilde{\kappa}^{(n)}(x-l_{n-1})} + \mathbf{u}_-^{(n)}e^{\tilde{\bar{\kappa}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)}\right)(I\otimes Q^-(K,m_s)P^+(K-1,m_s))\mathbf{e}, \tag{6.20}$$

$$\text{for } l_{n-1} \le x < l_n, n = 2,3,...,N.$$

The distribution of abandonment time $W_{L,1}$ of customers lost at the head of the waiting queue is given by

$$P\{W_{L,1} = l_n\} = \begin{cases} \dfrac{\eta_1\mathbf{p}_k^{(1)}(D_1\otimes I)\mathbf{e}}{p_{L,1}\lambda}, & \text{for } n=1; \\ \left(\dfrac{\eta_n}{\eta_n+...+\eta_N}\right)\dfrac{\mathbf{p}_{K+1}^{(n)}(l_n)\mathbf{e}}{p_{L,1}\lambda}, & \text{for } n=2,3,...,N-1. \end{cases} \tag{6.21}$$

The abandonment time $W_{L,>1}$ of a customer that abandons the queue before reaching the head of the queue, we have, for $k=1,2,...,N-1$,

$$P\{W_{L,>1}=l_k\}=\left(\sum_{n=k+1}^{N}\left(\mathbf{u}_+^{(n)}\mathcal{L}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n}\Psi^{(n)}+\mathbf{u}_-^{(n)}\widetilde{\mathcal{L}}^{\widetilde{\mathcal{K}}^{(n)}}_{l_{n-1},l_n}\right)(D_1\otimes I)\mathbf{e}\right)\frac{\eta_k}{p_{L,>1}\lambda}. \quad (6.22)$$

Proof. First, we note that $W_S=0$ occurs if a server is available when a customer arrives, which leads to the expression for $P\{W_S=0\}$. Similar to the proof of Proposition 1, we use the transition rate ratio to derive

$$\frac{\mathrm{d}P\{W_S<x\}}{\mathrm{d}x}=\frac{1}{p_S\lambda}\mathbf{p}_{K+1}(x)(I\otimes Q^-(K,m_s)P^+(K-1,m_s))\mathbf{e},\ \text{for } x>0, \quad (6.23)$$

which leads to the desired result.

Second, we note that $W_{L,1}=l_n$ if $a(t)$ reaches l_n from below and an abandonment occurs for $n=2,3,...,N-1$. The probability for $W_{L,1}$ to reach l_n is $\mathbf{p}^{(n)}_{K+1}(l_n)\mathbf{e}/(p_{L,1}\lambda)$. The probability for the abandonment to occur is $\eta_n/(\eta_n+...+\eta_N)$. For n = 1 (i.e., for Border l_1 = 0), the conditional probability for a customer who sees no waiting queue and no server available, and abandons the queue is $\mathbf{p}^{(1)}_K((\eta_1 D_1)\otimes I)\mathbf{e}/(p_{L,1}\lambda)$. Then expression (6.21) can be obtained easily.

We use the joint stationary distribution of the multi-layer *MMFF* process to find the distribution of $W_{L,>1}$. When the multi-layer *MMFF* process is in $\mathcal{S}_-^{(n)}$ and there is an arrival, the arriving customer will abandon the queue in the future with probability $\eta_2+...+\eta_{n-1}$ if $l_{n-1}<x<l_n$. Since customer arrivals take place only when the fluid level of the *MMFF* process is decreasing, we censor out the periods of time in which the fluid level is increasing. Using the censored process, we obtain, for $k=1,2,...,N-1$,

$$P\{W_{L,>1}=l_k\}=\frac{c_{norm}}{\hat{\hat{c}}_{norm}p_{L,>1}\lambda}\left(\sum_{n=k}^{N-1}\int_{l_n}^{l_{n+1}}\boldsymbol{\pi}_-^{(n+1)}(x)\mathrm{d}x(D_1\otimes I)\right)\mathbf{e}\eta_k, \quad (6.24)$$

where

$$\hat{\hat{c}}_{norm}=\sum_{k=0}^{K}\mathbf{p}_k^{(1)}\mathbf{e}+\sum_{n=2}^{N}\int_{l_{n-1}}^{l_n}\left(\mathbf{u}_+^{(n)}e^{\mathcal{K}^{(n)}(y-l_{n-1})}\Psi^{(n)}+\mathbf{u}_-^{(n)}e^{\widetilde{\mathcal{K}}^{(n)}(l_n-y)}\right)\mathbf{e}\mathrm{d}y. \quad (6.25)$$

(Remark: Vectors $\mathbf{u}_+^{(n)}$ and $\mathbf{u}_-^{(n)}$ in equation (6.25) are not normalized.) In the multi-layer *MMFF* process, the fluid level increases and decreases both at rate 1. If the process is ergodic, probabilities that the process is increasing or decreasing at an arbitrary time are equal. Thus, we must have $\hat{c}_{norm}=\hat{\hat{c}}_{norm}$, which leads to the desired result in equation (6.22).

According to the law of total probability, we must have $P\{W_S<\infty\}=1$ and $\sum_{n=1}^{N-1}P\{W_{L,1}=l_n\}=1$, which can be used to check computation accuracy. The law of total probability $\sum_{n=1}^{N-1}P\{W_{L,>1}=l_n\}=1$ can also be used to check computation accuracy. The mean

waiting time $\mathbb{E}[W_S]$ can be calculated by:

$$\mathbb{E}[W_S]=\frac{1}{p_S\lambda}\sum_{n=2}^{N}\left(\mathbf{u}_+^{(n)}\mathcal{M}_{l_{n-1},l_n}^{\vec{\kappa}^{(n)}}+\mathbf{u}_-^{(n)}\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\vec{\kappa}^{(n)}}\widehat{\Psi}^{(n)}\right)(I\otimes Q^-(K,m_s)P^+(K-1,m_s))\mathbf{e}, \quad (6.26)$$

where $\mathcal{M}_{l_{n-1},l_n}^{\vec{\kappa}^{(n)}}$ and $\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\vec{\kappa}^{(n)}}$ are defined in Lemma 9. The distribution of the waiting time W of an arbitrary customer can be found from that of W_S, $W_{L,1}$, and $W_{L,>1}$. The mean waiting time can be found by

$$\mathbb{E}[W]=p_S\mathbb{E}[W_S]+p_{L,1}\mathbb{E}[W_{L,1}]+p_{L,>1}\mathbb{E}[W_{L,>1}]. \quad (6.27)$$

Let $q_S(t)$ be the number of customers in service (or busy servers) and $q_W(t)$ the waiting queue length at an arbitrary time t. The distribution of $q_S(t)$ can be found directly from the border probability vector $\mathbf{p}^{(1)}$. The z-transform of $q_W(t)$ can be derived based on the joint distribution of the age process. If the $a(t)=x$ at an arbitrary time t, the waiting queue length consists of the customer at the head of the queue and all customers arrived after that customer (i.e., in the period $(t-x,t)$) who have not abandoned the queue yet. To identify who are still waiting in queue and who have abandoned the queue, we divide the interval $(t-x,t)$ into $(t-l_2,t)$, $(t-l_3,t-l_2)$, ..., $(t-x,t-l_{n-1})$, if $l_{n-1}<x<l_n$. Customers who arrived in $(t-l_2,t)$ are still in the system at time t with probability $1-\eta_1$. The conditional probability generating function of the number of such customers is $\exp\{(D_0+(\eta_1+(1-\eta_1)z)D_1)l_2\}$. (see He [28]). For customers arrived in $(t-l_3,t-l_2)$, they abandon the queue before t with probability $\eta_1+\eta_2$ and are still in the queue at time t with probability $1-\eta_1-\eta_2$. The conditional probability generating function is $\exp\{(D_0+(\eta_1+\eta_2+(1-\eta_1-\eta_2)z)D_1)(l_3-l_2)\}$. In general, for customers arrived in $(t-l_m,t-l_{m-1})$, they abandon the queue before t with probability $1-\hat{\eta}_m$ and are still in the queue at time t with probability $\hat{\eta}_m$, where $\hat{\eta}_m=\eta_m+\eta_{m+1}+...+\eta_N$. The conditional probability generating function is given by $\exp\{(D_0+(1-\hat{\eta}_m+\hat{\eta}_m z)D_1)(l_m-l_{m-1})\}$. Denote by $P^*(\eta,z,x)=\exp\{(D_0+(1-\eta+\eta z)D_1)x\}\otimes I$. Conditioning on $a(t)$ at an arbitrary time t, the probability generating function of $q_W(t)$ can be found as follows.

Lemma 10.

$$\mathbb{E}[z^{q_W(t)}]=\mathbf{p}^{(1)}\mathbf{e}+z\sum_{n=2}^{N}\int_{l_{n-1}}^{l_n}\mathbf{p}_{K+1}^{(n)}(x)P^*(\hat{\eta}_n,z,x-l_{n-1})\prod_{m=n-1}^{2}P^*(\hat{\eta}_m,z,b_m)\mathrm{d}x\mathbf{e}, \quad (6.28)$$

(Remark: $b_m=l_m-l_{m-1}$, for $m=2,3,...,N$.)

By Theorem 2.3.2 in He [28], we have

$$\left.\frac{\partial P^*(\eta,z,x)\mathbf{e}}{\partial z}\right|_{z=1} = \left(\eta\lambda x\mathbf{e}+(e^{Dx}-I)(D-\mathbf{e}\boldsymbol{\theta}_a)^{-1}\eta D_1\mathbf{e}\right)\otimes I. \tag{6.29}$$

Recall that $\lambda=\boldsymbol{\theta}_a D_1\mathbf{e}$ and $D=D_0+D_1$. Consequently, we obtain

Proposition 3. *The distribution of $q_S(t)$ is given by*

$$P\{q_S(t)=k\}=\begin{cases}\mathbf{p}_k^{(1)}\mathbf{e}, & \text{if } k=0,1,...,K-1;\\ 1-\sum_{k=0}^{K-1}\mathbf{p}_k^{(1)}\mathbf{e}, & \text{if } k=K.\end{cases} \tag{6.30}$$

The mean waiting queue length is given by

$$\begin{aligned}\mathbb{E}[q_W(t)] &= 1-\mathbf{p}^{(1)}\mathbf{e}\\ &+\sum_{n=2}^{N}\sum_{m=2}^{n-1}\int_{l_{n-1}}^{l_n}\mathbf{p}_{K+1}^{(n)}(x)\left(e^{D(x-l_m)}\otimes I\right)\mathrm{d}x\left(\hat{\eta}_m\lambda b_m I+(e^{Db_m}-I)(D-\mathbf{e}\boldsymbol{\theta}_a)^{-1}\hat{\eta}_m D_1\right)\mathbf{e}\otimes\mathbf{e}\\ &+\sum_{n=2}^{N}\int_{l_{n-1}}^{l_n}\mathbf{p}_{K+1}^{(n)}(x)\left(\hat{\eta}_n\lambda(x-l_{n-1})I+(e^{D(x-l_{n-1})}-I)(D-\mathbf{e}\boldsymbol{\theta}_a)^{-1}\hat{\eta}_n D_1\right)\mathbf{e}\otimes\mathbf{e}\mathrm{d}x.\end{aligned} \tag{6.31}$$

To calculate the mean queue length, we need to evaluate the integral, for $2\le m<n\le N$,

$$\int_{l_{n-1}}^{l_n}\left(\mathbf{u}_+^{(n)}e^{\bar{\kappa}^{(n)}(x-l_{n-1})}+\mathbf{u}_-^{(n)}e^{\bar{\kappa}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)}\right)\left(e^{D(x-l_{n-1})}\otimes I\right)\mathrm{d}x. \tag{6.32}$$

Define, for $a<b$ and matrix M,

$$\mathcal{L}_{a,b}^{(M,D)}=\int_a^b e^{M(x-a)}\left(e^{D(x-a)}\otimes I\right)dx;\ \text{and}\ \widetilde{\mathcal{L}}_{a,b}^{(M,D)}=\int_a^b e^{M(b-x)}\widehat{\Psi}^{(n)}\left(e^{D(x-a)}\otimes I\right)\mathrm{d}x. \tag{6.33}$$

Lemma 11. *If matrix M is invertible, $\mathcal{L}_{a,b}^{(M,D)}$ and $\widetilde{\mathcal{L}}_{a,b}^{(M,D)}$ satisfy the following Sylvester equations, respectively,*

$$\begin{aligned}&M\mathcal{L}_{a,b}^{(M,D)}+\mathcal{L}_{a,b}^{(M,D)}(D\otimes I)=e^{M(b-a)}\left(e^{D(b-a)}\otimes I\right)-I;\\ &M\widetilde{\mathcal{L}}_{a,b}^{(M,D)}-\widetilde{\mathcal{L}}_{a,b}^{(M,D)}(D\otimes I)=e^{M(b-a)}\widehat{\Psi}^{(n)}-\widehat{\Psi}^{(n)}\left(e^{D(b-a)}\otimes I\right).\end{aligned} \tag{6.34}$$

If M is non-invertible, let $\mathbf{v}_L$ and $\mathbf{v}_R$ the left and right eigenvectors of M, satisfying $\mathbf{v}_L\mathbf{v}_R=1$ and $\mathbf{v}_L\mathbf{e}=1$. Then $\mathcal{L}_{a,b}^{(M,D)}$ and $\widetilde{\mathcal{L}}_{a,b}^{(M,D)}$ satisfy the following Sylvester equations, respectively,

$$\begin{aligned}&(M-\mathbf{v}_R\mathbf{v}_L)\mathcal{L}_{a,b}^{(M,D)}+\mathcal{L}_{a,b}^{(M,D)}(D\otimes I)=e^{M(b-a)}\left(e^{D(b-a)}\otimes I\right)-I-\mathbf{v}_R\mathbf{v}_L L_1;\\ &(M-\mathbf{v}_R\mathbf{v}_L)\widetilde{\mathcal{L}}_{a,b}^{(M,D)}-\widetilde{\mathcal{L}}_{a,b}^{(M,D)}(D\otimes I)=e^{M(b-a)}\widehat{\Psi}^{(n)}-\widehat{\Psi}^{(n)}\left(e^{D(b-a)}\otimes I\right)-\mathbf{v}_R\mathbf{v}_L\widehat{\Psi}^{(n)}L_1,\end{aligned} \tag{6.35}$$

$$where \quad L_1 = \left(\int_a^b e^{D(x-a)} \mathrm{d}x \right) \otimes I = \left((e^{D(b-a)} - I - (b-a)\mathbf{e}\boldsymbol{\theta}_a)(D - \mathbf{e}\boldsymbol{\theta}_a)^{-1} \right) \otimes I.$$

Proof. The lemma can be proved similar to that of Lemma 9. Details are omitted.

Remark. It is well-known that a Sylvester equation of the type $AX + XB = C$ has a unique solution if matrices A and B have no common eigenvalues. That is why we present the second part of Lemma 11. The existence of a solution is still not guaranteed, though.

Combining Proposition 3 and Lemma 11, we obtain

$$\begin{aligned}
&\mathbb{E}[q_W(t)] = 1 - \mathbf{p}^{(1)}\mathbf{e} \\
&+ \sum_{n=2}^{N} \left(\mathbf{u}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{(\vec{\kappa}^{(n)},D)} + \mathbf{u}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{(\vec{\kappa}^{(n)},D)} \right) \left(\sum_{m=2}^{n-1} e^{D(l_{n-1}-l_m)} \left(\hat{\eta}_m \lambda b_m I + (e^{Db_m} - I)(D - \mathbf{e}\boldsymbol{\theta}_a)^{-1} \hat{\eta}_m D_1 \right) \otimes I \right) \mathbf{e} \\
&+ \sum_{n=2}^{N} \hat{\eta}_n \lambda \left(\mathbf{u}_+^{(n)} \left(M_{l_{n-1},l_n}^{\vec{\kappa}^{(n)}} - l_{n-1} \mathcal{L}_{l_{n-1},l_n}^{\vec{\kappa}^{(n)}} \right) + \mathbf{u}_-^{(n)} \left(\widetilde{M}_{l_{n-1},l_n}^{\vec{\kappa}^{(n)}} - l_{n-1} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\vec{\kappa}^{(n)}} \right) \widehat{\Psi}^{(n)} \right) \mathbf{e} \\
&+ \sum_{n=2}^{N} \left(\mathbf{u}_+^{(n)} \left(\mathcal{L}_{l_{n-1},l_n}^{(\vec{\kappa}^{(n)},D)} - \mathcal{L}_{l_{n-1},l_n}^{\vec{\kappa}^{(n)}} \right) + \mathbf{u}_-^{(n)} \left(\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{(\vec{\kappa}^{(n)},D)} - \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\vec{\kappa}^{(n)}} \widehat{\Psi}^{(n)} \right) \right) \left((D - \mathbf{e}\boldsymbol{\theta}_a)^{-1} \hat{\eta}_n D_1 \otimes I \right) \mathbf{e}.
\end{aligned} \tag{6.36}$$

Note that Lemma 9 is used in the above expression.

Let $q_{tot}(t)$ be the total number of customers in the queueing system at an arbitrary time t. Then the probability generating function and the mean of $q_{tot}(t)$ can be found as

$$\begin{aligned}
&\mathbb{E}[z^{q_{tot}(t)}] = \sum_{k=0}^{K} z^k \mathbf{p}_k^{(1)} \mathbf{e} + z^K \mathbb{E}[z^{q_W(t)}]; \\
&\mathbb{E}[q_{tot}(t)] = \sum_{k=0}^{K} k \mathbf{p}_k^{(1)} \mathbf{e} + K(1 - \mathbf{p}^{(1)}\mathbf{e}) + \mathbb{E}[q_W(t)].
\end{aligned} \tag{6.37}$$

The queueing quantities are connected to each other by the well-known Little's law:

- $\mathbb{E}[q_W(t)] = \lambda \mathbb{E}[W]$ for the number of waiting customers and the actual waiting times of customers;
- $\mathbb{E}[q_S(t)] = \lambda p_S \boldsymbol{\beta}(-T)^{-1}\mathbf{e}$ for the number of customers in service and service times; and
- $\mathbb{E}[q_{tot}(t)] = \lambda \mathbb{E}[W] + \lambda p_S \boldsymbol{\beta}(-T)^{-1}\mathbf{e}$ for the total number of customers in the queueing system and the sojourn times of customers.

The relationships provide insight on the quantities and queueing system of interest, and can be used for checking computation accuracy.

6.6. Numerical examples

In this subsection, we present five examples to gain insight on the abandonment probabilities, waiting times, and queue lengths. We also use the examples to discuss the dimensionality issues of our algorithm. We apply our algorithm to queueing models

investigated in Dai and He [21] and Whitt [44].

Example 6.1. We consider an *MAP/PH/K+GI* queue with $K=50$, $N=7$, $(l_1,l_2,l_3,l_4,l_5,l_6,l_7)=(0,1,2,3,4,6,\infty)$, $\boldsymbol{\eta}=(0,0.1,0.2,0.1,0.2,0.2,0.2)$,

$$D_0=\begin{pmatrix}-66 & 16\\ 10 & -88\end{pmatrix},\ D_1=\begin{pmatrix}20 & 30\\ 20 & 58\end{pmatrix};\ \boldsymbol{\beta}=(0.4,0.6),\ T=\begin{pmatrix}-2.0 & 1.0\\ 0.6 & -1.0\end{pmatrix}. \quad (6.38)$$

Table 2 shows basic quantities for the arrival process, the service time distribution, and the abandonment time distribution for those customers that are not infinitely patient, that is, for $\tau|\tau<\infty$. We see that $\rho=\lambda/(\mu K)=2.257,$ indicating that the system receives more than double the traffic volume that it can serve, and therefore we expect that a large proportion of customers will abandon the queue. The abandonment time distribution for the 80% of customers who are not infinitely patient is bimodal, which illustrates the flexibility of our approach

Table 2. Basic Quantities of the Input Parameters for Example 6.1.

	MAP	*PH-Dist.*	$\tau\|\tau<\infty$
Rate	$\lambda=66.9474$	$\mu=0.5932$	$\frac{1}{\mathbb{E}[\tau\|\tau<\infty]}=0.2857$
Mean	0.0154	1.6857	3.5
SCV	1.0475	1.0396	0.245

Applying Algorithm II, a number of queueing quantities can be obtained. First, we plot the stationary density functions of the age of the customer at the head of the queue and the waiting time of an arbitrary served customer in Figure 7. It seems that most of the customers have to wait in the queue for service, yet the mean waiting times are mostly less than $l_6=6.$ Thus, the densities are concentrated around $l_5=4.$ Second, we present the (conditional) distributions of the waiting times of customers abandoned the queue in Table 3. While the possibility of customers abandoning the queue varies significantly before they reach the head of the queue, most of them abandon the queue at $l_3=2$ (if they do abandon the queue). If a customer reaches the head of the waiting queue, it has a big chance to enter service before their waiting time (or age) reaches $l_6=6.$ Lastly, we summarize other queueing quantities in Table 4.

Table 3. Conditional distributions of waiting times of customers abandoned the queue for Example 6.1.

	l_1	l_2	l_3	l_4	l_5	l_6	l_7
$P\{W_{L,1}=l_n\}$	0	0.0	0.0	0.0	0.9962	0.0038	0.0
$P\{W_{L,>1}=l_n\}$	0	0.1845	0.3690	0.1845	0.2619	0.0001	0.0
$P\{W_L=l_n\}$	0	0.1795	0.3591	0.1795	0.2816	0.0002	0.0

Table 4. Summary of queueing quantities for Example 6.1.

$\mathbb{E}[a(t)]$	p_S	p_L	$p_{L,1}$	$p_{L,>1}$	$p_{q,0}$	$\mathbb{E}[W_S]$
4.2200	0.4431	0.5569	0.0149	0.5419	0.0	4.2195
$\mathbb{E}[W_{L,1}]$	$\mathbb{E}[W_{L,>1}]$	$\mathbb{E}[W_L]$	$\mathbb{E}[W]$	$\mathbb{E}[q_S]$	$\mathbb{E}[q_W]$	$\mathbb{E}[q_{tot}]$
4.008	2.524	2.564	3.297	50.00	220.76	270.76

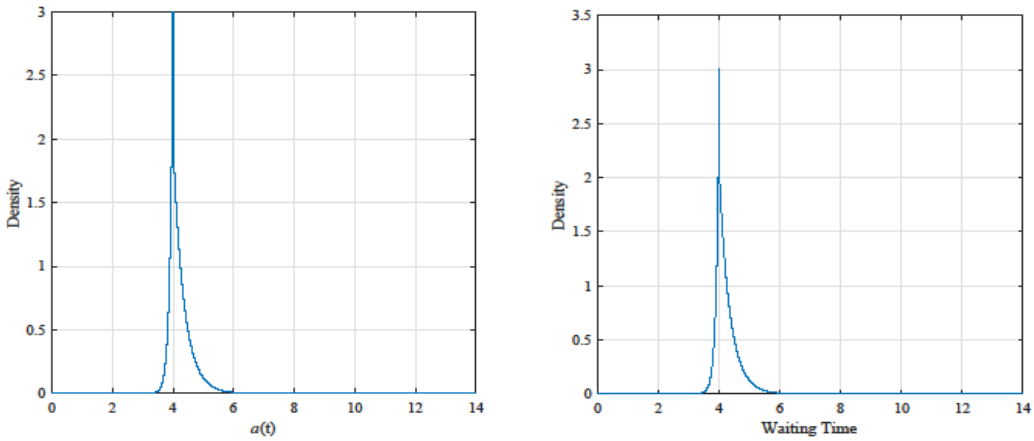
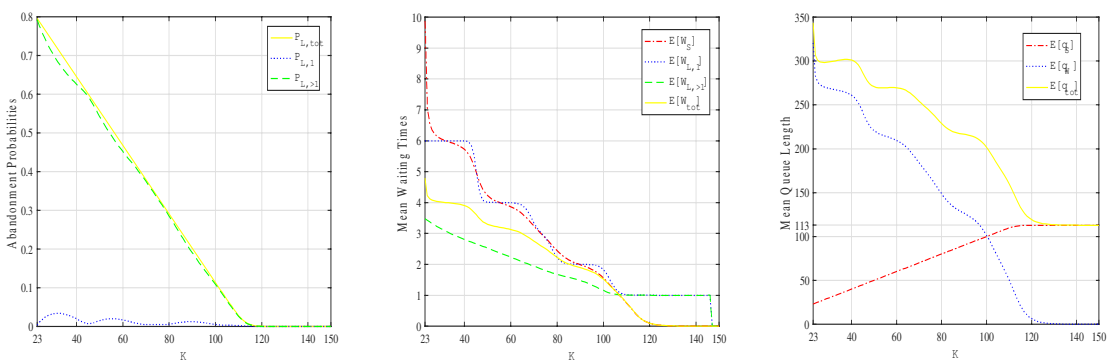


Figure 7. The stationary density functions of $a(t)$ and W_S for Example 6.1.

Example 6.2. (Example 6.1 continued) We extend Example 6.1 by varying the number of servers from $K=23$ (corresponding to $\rho=4.514$) to $K=150$ (corresponding to $\rho=0.7524$), and compute queueing quantities for those queueing systems. The results are divided into three groups $\{p_L, p_{L,1}, p_{L,>1}\}$, $\{\mathbb{E}[W_S], \mathbb{E}[W_{L,1}], \mathbb{E}[W_{L,>1}], \mathbb{E}[W]\}$, and $\{\mathbb{E}[q_S], \mathbb{E}[q_W], \mathbb{E}[q_{tot}]\}$. The results are plotted in Figure 8.



(a) Abandonment Probabilities (b) Mean Waiting Times (c) Mean Queue Length

Figure 8. Summary of queueing quantities for Example 6.2.

From Figure 8, it is interesting to see that (i) The abandonment probability $p_{L,1}$ can

go up and down as K increases; (ii) The mean waiting times are all decreasing (which is intuitive); and (iii) The mean total queue length can increase when K increases, which is due to more customers in service.

We also plot the density function of the waiting time of served customers for K = 25, 50, 100, 120 in Figure 9. It is interesting to see how the waiting time distribution shifts as K changes. One thing particularly interesting is the impact of the abandonment epochs on the waiting time distribution, which becomes less significant as K increases. Intuitively, it happens because fewer customers are forced to abandon as the number of servers increases.

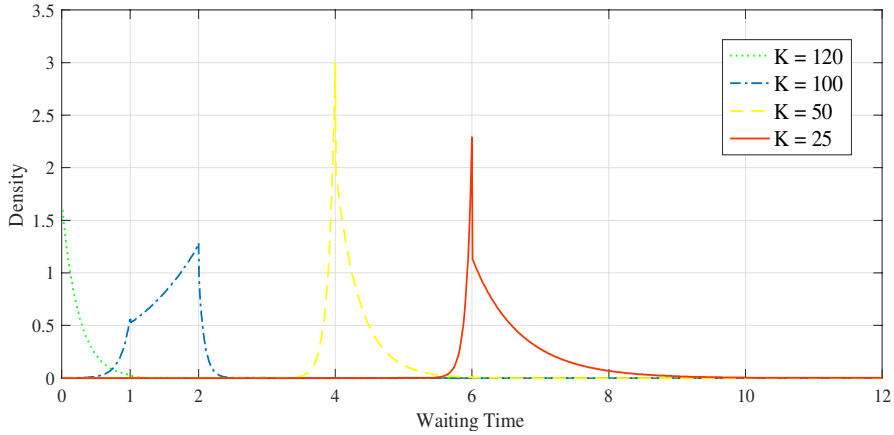
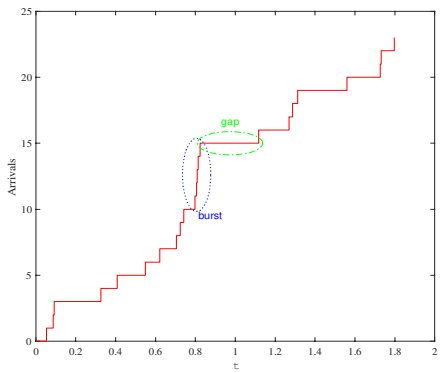


Figure 9. The stationary density functions of W_S for $K=25,50,100,120$ for Example 6.2.

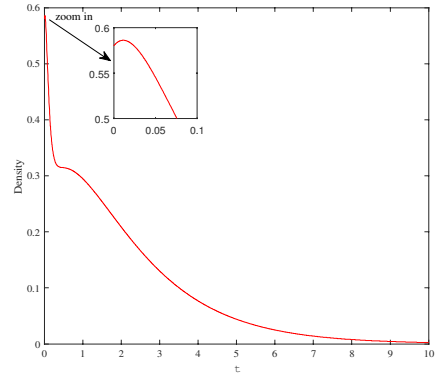
Example 6.3. In this example, we consider a queueing system with a bursty arrival process and service times with a big variation. We assume $N=5$, $l_1=0$, $l_2=1$, $l_3=3$, $l_4=5$, $l_5=\infty$, $\boldsymbol{\eta}=(0,0.2,0.3,0.4,0.1)$,

$$m_a=4,\ D_0=\begin{pmatrix}-15 & 0 & 2 & 2\\ 20 & -45 & 2 & 2\\ 1 & 2 & -25 & 5\\ 1 & 0 & 2 & -15\end{pmatrix},\ D_1=\begin{pmatrix}5 & 5 & 1 & 0\\ 10 & 5 & 1 & 5\\ 1 & 6 & 5 & 5\\ 5 & 1 & 1 & 5\end{pmatrix};$$
$$m_s=4,\ \boldsymbol{\beta}=(0.1,0.1,0.7,0.1),\ T=\begin{pmatrix}-17 & 0 & 0 & 12\\ 17 & -17 & 0 & 0\\ 0 & 0.4 & -0.8 & 0.4\\ 0.1 & 0 & 0.1 & -1\end{pmatrix}. \tag{6.39}$$

This example is special since the arrival process is bursty and the service times have a special distribution as shown in Figure 10, although Table 5 seems to indicate a less variable queue than that in Examples 6.1 and 6.2. We use this example to demonstrate that (i) Algorithm II can be used for analyzing queueing systems with matrix-dimensions greater than two; and (ii) The algorithm faces the matrix-dimensionality challenge, but it is applicable to models with specially featured arrival and service processes.



(a) Sample Path of Bursty Arrival Process.



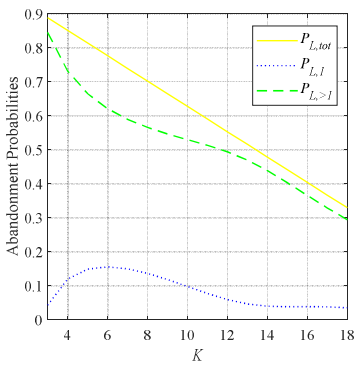
(b) Density of Service Times.

Figure 10. Burstiness of the arrival process and density function of the service times for Example 6.3.

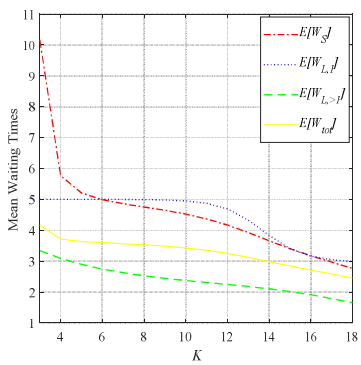
Table 5. Basic Quantities of the Input Parameters for Example 6.3.

	MAP	*PH-Dist.*	$\tau \mid \tau < \infty$
Rate	$\lambda = 13.0132$	$\mu = 0.4848$	$\frac{1}{\mathbb{E}[\tau \mid \tau < \infty]} = 0.2903$
Mean	0.0801	2.0627	3.4444
SCV	1.0194	0.8171	0.2081

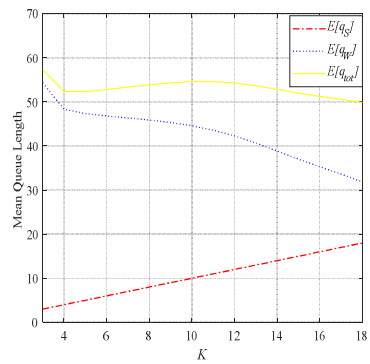
We vary K from 3 to 18. We compute queueing quantities for Example 6.3. Results related to customer abandonment, waiting times and queue lengths are plotted in Figure 11. As expected, the queue length seems big, the waiting time seems long, and abandonment of customers seems significant, even when $K = 18$.



(a) Abandonment Probabilities



(b) Mean Waiting Times



(c) Mean Queue Length

Figure 11. Summary of queueing quantities for Example 6.3.

One issue related to the analysis of complicated stochastic systems is state space explosion. Specifically, for our *MAP*/*PH*/*K*+*GI* queue, the number of states in $\Omega(K)$ can be

very big. For Examples 6.1 and 6.3, the number of states for each layer is given by $m_a\begin{pmatrix} K+m_s-1 \\ m_s-1 \end{pmatrix}$. We present the number of states as a function of K in Table 6.

Table 6. Number of states in $S_+^{(n)} \cup S_-^{(n)}$ for Examples 6.2 and 6.3.

K	1	5	8	10	12	14	15	18	50	100
Example 6.2	4	12	18	22	26	30	32	38	102	202
$m_a=4$, $m_s=3$	12	84	180	264	364	480	544	760	5304	20604
Example 6.3	16	224	660	1144	1820	2720	3264	5320	93704	707404

It is shown that, if m_a and m_s are small, Algorithm II can be applied for computing queueing quantities for K up to 50 or even over 100. Since one can generate all kinds of arrival processes and service times even for small m_a and m_s (e.g., Examples 6.1 and cases with $m_s=3$), the method developed in this paper can be useful for researchers and practitioners.

Next, we use our algorithm to address the performance insensitivity to abandonment time distributions, an issue examined in Dai and He [24].

Example 6.4. We use the example in Section 6 in Dai and He [24]. We consider an $M/M/100+GI$ queue with Poisson arrival process $\{D_0=-105, D_1=105\}$ and exponential service time $\{\boldsymbol{\beta}=1, T=-1\}$. The distribution of the abandonment time τ can be (i) an exponential distribution with parameter α, denoted as exp, (ii) a uniform distribution on $[0,1/\alpha]$, denoted as $Unif$, or (iii) a phase-type distribution with $\{\boldsymbol{\beta}_\tau=(0.7,0.3)$ and $T_\tau=\begin{pmatrix} -0.3\alpha & 0 \\ 0 & -79\alpha/30 \end{pmatrix}\}$, denoted as H_2, which is the well-known Hyperexponential distribution, where α is a positive constant.

To use Algorithm II, we discretize the above three abandonment distributions with $N=1000$, which gives satisfactory approximation results to the continuous case (as compared to results in Dai and He [24]. Specifically, for abandonment time τ with an exponential or H_2 distribution, the interval $[0,3\mathbb{E}[\tau]]$ is divided into $N-1$ identical intervals of length $\delta=3\mathbb{E}[\tau]/(N-1)$. Then we define $\eta_1=0$, $\eta_n=P\{(n-1)\delta\le\tau<n\delta\}$, for $n=2,3,...,N-1$, and $\eta_N=P\{\tau\ge N\delta\}$. For τ with an uniform distribution, the interval $[0,2\mathbb{E}[\tau]]$ is divided into $N-1$ identical intervals of length $\delta=2\mathbb{E}[\tau]/(N-1)$. Then we define $\eta_1=0, \eta_n=1/(N-1)$, for $n=2,3,...,N-1$, and $\eta_N=0$.

Dai and He [24] observes that the performance of the queue is insensitive to abandonment time distributions. Specifically, through simulation, they have observed that the queue with those three abandonment time distributions perform similarly, even though, for given α, the three abandonment times have different means and variances. Results presented in Table 7 indicates that queueing performance, with respect to more queueing quantities than those in Dai and He [24], is insensitive to abandonment time distributions,

which is consistent with the conclusion in Dai and He [24].

Table 7. Summary of queueing quantities for Example 6.4: Part I.

	$\mathbb{E}[a(t)]$			p_L			$p_{L,1}$			$p_{L,>1}$		
α	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2
0.1	0.5176	0.5010	0.5562	0.0496	0.0497	0.0493	0.0009	0.0010	0.0009	0.0487	0.0487	0.0484
0.5	0.1216	0.1154	0.1319	0.0601	0.0605	0.0593	0.0037	0.0040	0.0034	0.0564	0.0565	0.0559
1	0.0660	0.0614	0.0728	0.0668	0.0674	0.0658	0.0063	0.0069	0.0057	0.0605	0.0605	0.0601
2	0.0354	0.0319	0.0402	0.0738	0.0747	0.0726	0.0103	0.0116	0.0091	0.0635	0.0631	0.0635
10	0.0074	0.0056	0.0099	0.0886	0.0901	0.0868	0.0276	0.0340	0.0225	0.0609	0.0561	0.0643
	$\mathbb{E}[W_S]$			$\mathbb{E}[W_{L,1}]$			$\mathbb{E}[W_{L,>1}]$			$\mathbb{E}[W]$		
α	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2
0.1	0.5187	0.5021	0.5572	0.5390	0.5306	0.5701	0.3460	0.3414	0.3731	0.5103	0.4943	0.5483
0.5	0.1232	0.1170	0.1336	0.1566	0.1556	0.1621	0.1113	0.1100	0.1172	0.1227	0.1167	0.1327
1	0.0673	0.0627	0.0742	0.0995	0.0994	0.1019	0.0720	0.0710	0.0752	0.0678	0.0635	0.0744
2	0.0364	0.0329	0.0413	0.0651	0.0654	0.0659	0.0474	0.0465	0.0492	0.0373	0.0341	0.0420
10	0.0078	0.0059	0.0104	0.0257	0.0264	0.0255	0.0182	0.0169	0.0192	0.0089	0.0072	0.0113
	$p_{q,0}$			$\mathbb{E}[q_S]$			$\mathbb{E}[q_W]$			$\mathbb{E}[q_{tot}]$		
α	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2
0.1	0.0340	0.0355	0.0287	99.794	99.784	99.826	53.582	51.904	57.57	153.38	151.69	157.39
0.5	0.2165	0.2238	0.2027	96.686	98.642	98.770	12.880	12.250	13.94	111.57	110.90	112.71
1	0.3316	0.3425	0.3144	97.988	97.922	98.092	7.122	6.667	7.816	105.11	104.59	105.91
2	0.4532	0.4684	0.4323	97.250	97.158	97.377	3.921	3.581	4.410	101.17	100.74	101.79
10	0.7089	0.7356	0.6774	95.699	95.537	95.890	0.936	0.758	1.185	96.63	96.30	97.07

The observation seems to hold for queueing systems with a Poisson arrival process and exponential service times. However, it may not hold, even approximately, for queueing systems with a non-Poisson arrival process. Now, we change the customer arrival process from Poisson to *MAP* with

$$D_0 = \begin{pmatrix} -1 & 0.2 \\ 1 & -310 \end{pmatrix} \quad D_1 = \begin{pmatrix} 0.1 & 0.7 \\ 1 & 308 \end{pmatrix}. \tag{6.40}$$

The average arrival rate is 96.4483. The arrival process is bursty since the arrival rates in the two states of the underlying Markov chain are drastically different. Quantities in Table 7 are reproduced and presented in Table 8. Table 8 demonstrates that some quantities can be significantly different for the three abandonment times (e.g., $p_{L,1}$ and $\mathbb{E}[q_W]$ for $\alpha \geq 2$), which indicates that the queueing performance is no longer insensitive to the abandonment time distributions.

To end this section, we analyze the $M/E_2/100+E_2$ queue and compare our results to that in Whitt [44].

Example 6.5. We consider the example in Section 2 in Whitt [44]. Instead of limiting the waiting spaces to 200 in the original example (i.e., $M/E_2/100/200+E_2$ with 200 extra waiting spaces), we assume that the queue has unlimited waiting space (i.e., $M/E_2/100+E_2$). The arrival process and service time follow a Poisson arrival process $\{D_0=-102, D_1=102\}$ and Erlang 2 (E_2) service time distribution $\{\boldsymbol{\beta}=[1,0], T=\begin{pmatrix}-2 & 2\\ 0 & -2\end{pmatrix}\}$ respectively. The abandonment time τ has a Erlang distribution with phase-type representation $\{\boldsymbol{\beta}_\tau=[1,0], T_\tau=\begin{pmatrix}-2 & 2\\ 0 & -2\end{pmatrix}\}$. Similar to Example 6.4, we discretize the above Erlang distribution with $N=1000$.

For the queueing model, the customer arrival rate is $\lambda=102$ and service rate of a server is $\mu_s=1$. Then $\rho=1.02$. Since η_N is almost zero, $\eta_N\rho$ is nearly zero and the queueing system is stable. Due to customer abandonments, the (waiting) queue length rarely reaches 200. Thus, the performance of the $M/E_2/100/200+E_2$ queue and the $M/E_2/100+E_2$ (*discretized*) queue is very close. Results are presented in Table 9.

Table 8. Summary of queueing quantities for Example 6.4: Part II.

	$\mathbb{E}[a(t)]$			p_L			$p_{L,1}$			$p_{L,>1}$		
α	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2
0.1	1.2385	1.1090	1.4095	0.1470	0.1550	0.1376	0.0007	0.0008	0.0006	0.1463	0.1542	0.1371
0.5	0.3686	0.2842	0.4968	0.2525	0.2719	0.2302	0.0026	0.0038	0.0018	0.2500	0.2681	0.2284
1	0.2020	0.1432	0.3027	0.2994	0.3219	0.2715	0.0043	0.0071	0.0028	0.2951	0.3148	0.2687
2	0.1062	0.0694	0.1783	0.3413	0.3629	0.3105	0.0072	0.0137	0.0042	0.3341	0.3493	0.3063
10	0.0213	0.0117	0.0449	0.4031	0.4134	0.3821	0.0247	0.0617	0.0111	0.3785	0.3517	0.3710
	$\mathbb{E}[W_S]$			$\mathbb{E}[W_{L,1}]$			$\mathbb{E}[W_{L,>1}]$			$\mathbb{E}[W]$		
	Exp	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2
0.1	1.5054	1.3608	1.6947	1.8257	1.8502	1.8126	1.3504	1.3316	1.3851	1.4829	1.3567	1.6523
0.5	0.5113	0.4048	0.6691	0.7401	0.7240	0.7535	0.4933	0.4590	0.5296	0.5074	0.4205	0.6374
1	0.2990	0.2190	0.4307	0.4842	0.4518	0.5151	0.3015	0.2672	0.3401	0.3005	0.2358	0.4066
2	0.1671	0.1130	0.2681	0.3077	0.2666	0.3530	0.1763	0.1472	0.2139	0.1712	0.1270	0.2518
10	0.0370	0.0207	0.0753	0.0893	0.0630	0.1383	0.0427	0.0306	0.0640	0.0404	0.0268	0.0718
	$p_{q,0}$			$\mathbb{E}[q_S]$			$\mathbb{E}[q_W]$			$\mathbb{E}[q_{tot}]$		
	Exp	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2	*Exp*	*Unif*	H_2
0.1	0.3182	0.3321	0.3020	82.269	81.498	83.172	143.03	130.85	159.36	225.30	212.35	242.53
0.5	0.5009	0.5344	0.4622	72.091	70.226	74.247	48.94	40.56	61.47	121.03	110.78	135.72
1	0.5821	0.6210	0.5337	67.567	65.400	70.266	28.99	22.74	39.22	96.55	88.14	109.48
2	0.6546	0.6921	0.6013	63.530	61.443	66.501	16.51	12.25	24.29	80.04	73.69	90.79
10	0.7618	0.7796	0.7254	57.567	56.580	59.597	3.90	2.58	6.93	61.46	59.16	66.52

Table 9. Summary of queueing quantities for Example 6.5.

Performance Measure	Simulation (Whitt)	Approximation (Whitt)	*MMFF*
$P\{W=0\}$	0.217 ± 0.0021	0.250	0.2153
p_L	0.0351 ± 0.00029	0.0381	0.0350
$\mathbb{E}[q_W]$	11.52 ± 0.075	11.41	11.620
$\mathbb{E}[q_{tot}]$	109.9 ± 0.092	109.5	110.05
$\mathbb{E}[W_S]$	0.1115 ± 0.00071	0.1102	0.1125
$\mathbb{E}[W_L]$	0.1508 ± 0.00042	0.1521	0.1524

We note that the half width of 95% confidence intervals are shown in the column for simulation results. Table 9 shows that our numerical results are fairly close to simulation results. Some of our results are not in the 95% intervals of corresponding quantities since their model has finite waiting space while our model has infinite waiting space. In addition, the following two reasons may contribute to the difference in the numerical results: (i) There is always a chance that the actual quantity is outside of the confidence interval; and (ii) The abandonment time distributions are different for our and their models.

7. Conclusions

In this paper, we reviewed and extended the basic theory on the joint stationary distribution for multi-layer *MMFF* processes. We applied the basic theory to the *MAP*/*PH*/*K*+*GI* queue and developed computational methods for queueing quantities such as the customer abandonment probabilities, distributions of waiting times, and the mean queue lengths.

As aforementioned in Section 2, the method developed in this paper can be applied to the *MAP*/*PH*/*K*+*GI* in which the customer arrival process and/or service times depends on the age of the customer at the head of the waiting queue. There are also a number of issues for future research: (i) Computational details if $\mu_n = 0$ for some $n=2,...,N-1$ for multi-layer *MMFF* processes; (ii) The queue length distribution for the *MAP*/*PH*/*K*+*GI* queue; (iii) The *MAP*/*PH*/*K*+*GI* queue in which customers make their abandonment decisions at specific waiting time epochs $\{l_2, l_3, ..., l_{N-1}\}$; (iv) The *MMAP*[*L*]/*PH*[*L*]/*K*+*GI* queue, a queueing model with multiple-types of customers; and (v) The *MMAP*[*L*]/*PH*[*L*]/*K* queue with customer priorities. Those issues/models are currently under investigation.

Acknowledgements

We would like to thank two anonymous referees for their valuable comments and suggestions on the paper that improved the quality of the paper significantly.

References

[1] Ahn, S., Badescu, A. L., & Ramaswami, V. (2007). Time dependent analysis of finite buffer fluid flows and risk models with a dividend barrier. *Queueing Systems*, 55, 207-222.

[2] Ahn, S., & Ramaswami, V. (2003). Fluid flow models and queues- A connection by stochastic coupling. *Stochastic Models*, 19, 325-348.

[3] Ahn, S., & Ramaswami, V. (2004). Transient analysis of fluid flow models via stochastic coupling to a queue. *Stochastic Models*, 20, 71-101.

[4] Ahn, S., & Ramaswami, V. (2011). Duality results for Markov-modulated fluid flow models. *Journal of Applied Probability*, 48, 309-318.

[5] Anick, D., Mitra, D., & Sondhi, M. M. (1982). Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61, 1871-1894.

[6] Asmussen, S. (1995). Stationary distributions for fluid flow models with or without Brownian noise. *Communications in Statistics. Stochastic Models,* 11, 21-49.

[7] Asmussen, S. (2014) Lévy processes, phase-type distributions, and martingales. *Stochastic Models*, 30, 443-468.

[8] Avram, F., & Usabel, M. (2004). Ruin probabilities and deficit for the renewal risk model with phase-type interarrival times. *Astin Bulletin*, 34, 315-332.

[9] Badescu, A., Breuer, L., Soares, A. D. S., Latouche, G., Remiche, M. A., & Stanford, D. (2005). Risk processes analyzed as fluid queues. *Scandinavian Actuarial Journal*, 105, 127-141.

[10] Badescu, A., Drekic, S., & Landriault, D. (2007a). Analysis of a threshold dividend strategy for a *MAP* risk model. *Scandinavian Actuarial Journal*, 4, 227-247.

[11] Badescu, A., Drekic, S., & Landriault, D. (2007b). On the analysis of a multi-threshold Markovian risk model. *Scandinavian Actuarial Journal,* 4, 248-260.

[12] Badescu, A., & Landriault, D. (2008). Recursive calculation of the dividend moments in a multi-threshold risk model. *North American Actuarial Journal*, 12, 74-88.

[13] Badescu, A., & Landriault, D. (2009). Applications of fluid flow matrix analytic methods in ruin theory - a review. Series A: *Matemticas de la Revista de la Real Academia de Ciencias Exactas, Fsicas y Naturales*, 103, 353-372.

[14] Bean, N., & O'Reilly, M. (2008). Performance measures of a multi-layer Markovian fluid model. *Annals of Operations Research*, 160, 99-120.

[15] Bean, N., & O'Reilly, M. (2013). A stochastic two-dimensional fluid model. *Stochastic Models*, 29, 31-63.

[16] Bean, N., O'Reilly, M., & Taylor, P. (2005). Hitting probabilities and hitting times for stochastic fluid flows. *Stochastic Processes and their Applications,* 115, 1530-1556.

[17] Bean, N., O'Reilly, M., & Taylor, P. (2009). Hitting probabilities and hitting times for stochastic fluid flows: the bounded model. *Probability in the Engineering and Information Sciences*, 23, 121-147.

[18] Choi, B., Kim, B., & Zhu, D. (2004). *MAP/M/c* queue with constant impatient time. *Mathematics of Operations Research*, 29, 309-325.

[19] da Silva Soares, A., & Latouche, G. (2002). Further results on the similarity between fluid queues and QBDs. In G. Latouche and P. Taylor, editors, *Proceedings of the 4th International Conference on Matrix-Analytic Methods*, 89-106.

[20] da Silva Soares, A., & Latouche, G. (2005) A matrix-analytic approach to fluid queues with feedback control. *International Journal of Simulation: Systems, Science and Technology*, 6, 1-2.

[21] da Silva Soares, A., & Latouche, G. (2006). Matrix-analytic methods for fluid queues with finite buffers. *Performance Evaluation*, 63, 295-314.

[22] da Silva Soares, A., & Latouche, G. (2009). Fluid queues with level dependent evolution. *European Journal of Operational Research*, 196, 1041-1048.

[23] Dai, J., & He, S. (2010). Customer abandonment in many-server queues. *Mathematics of Operations Research*, 35, 347-362.

[24] Dai, J., & He, S. (2011) Queues in service systems: Customer abandonment and diffusion approximations. *Transforming Research into Action*, 36-59 (INFORMS).

[25] Dai, J., He, S., & Tezcan, T., (2010). Many-server diffusion limits for *G/Ph/n + GI* queues. *The Annals of Applied Probability*, 20, 1854-1890.

[26] Guo, C. H. (2001). Nonsymmetric algebraic Riccati equations and Wiener-Hopf factorization for M-matrices. *SIAM Journal on Matrix Analysis and Applications*, 23, 225-242.

[27] Guo, C. H. (2002). A note on the minimal nonnegative solution of a nonsymmetric algebraic Riccati equation. *Linear Algebra and its Applications*, 357, 299-302.

[28] He, Q. M. (2014). Fundamentals of Matrix-Analytic Methods, Volume 365 (Springer).

[29] He, Q. M., & Alfa, A. S. (2017). Space reduction for a class of multidimensional Markov chains: A summary and some applications. *INFORMS Journal on Computing*, 30, 1-10.

[30] He, Q. M., Zhang, H., & Ye, Q. (2018). An *M/PH/K* queue with constant impatient time. *Mathematical Methods of Operations Research*, 87, 139-168.

[31] Horváth, G. (2015). Efficient analysis of the *MMAP[K]/PH[K]*/ 1 priority queue. *European Journal of Operational Research*, 246, 128-139.

[32] Horváth, G., & Van Houdt, B. (2012). A multi-layer fluid queue with boundary phase transitions and its application to the analysis of multi-type queues with general customer impatience. *Quantitative Evaluation of Systems (QEST), the Ninth International Conference on Quantitative Evaluation of Systems*, 23-32.

[33] Kim, B., & Kim, J. (2015). A single server queue with Markov modulated service rates and impatient customers. *Performance Evaluation*, 83, 1-15.

[34] Latouche, G., & Nguyen, G. (2018). Analysis of fluid flow models. *Queueing Models and Service Management*, 1, 30-43.

[35] Latouche, G., & Ramaswami, V. (2011). Introduction to Matrix Analytic Methods in Stochastic Modeling (ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia PA, 1999. Second printing).

[36] Loynes, R. (1962). A continuous-time treatment of certain queues and infinite dams. *Journal of the Australian Mathematical Society*, 2, 484-498.

[37] Meini, B. (2013). On the numerical solution of a structured nonsymmetric algebraic Riccati equation. *Performance Evaluation*, 70, 682-690.

[38] Neuts, M. (1979). A versatile Markovian point process. *Journal of Applied Probability*, 16, 764-779.

[39] Neuts, M. (1981). Matrix-Geometric Solution in Stochastic Model: An Algorithmic Approach, The Johns Hopkins University Press, Baltimore, MD.

[40] Ramaswami, V. (1985). Independent Markov processes in parallel. *Stochastic Models*, 1, 419-432.

[41] Ramaswami, V. (1999). Matrix analytic methods for stochastic fluid flows. in: D. Smith, P. Hey (Eds.), T*eletraffic Engineering in a Competitive World (Proceedings of the 16th International Teletraffic Congress)*, 1019-1030.

[42] Rogers, L. (1994). Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *Annals of Applied Probability*, 4, 390-413.

[43] Van Houdt, B. (2012). Analysis of the adaptive *MMAP[K]/PH[K]*/1 queue: A multi-type queue with adaptive arrivals and general impatience. *European Journal of Operational Research*, 220, 695-704.

[44] Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Science*, 51, 221-235.