SIAM J. Appl. Math.
Vol. 74, No. 4, pp. 1238–1265

GRAM CHARLIER EXPANSION FOR TIME VARYING MULTISERVER QUEUES WITH ABANDONMENT[*]

JAMOL PENDER[†]

Abstract. In this paper, we introduce a new approximation for estimating the dynamics of multiserver queues with abandonment. The approximation involves a four-dimensional dynamical system that uses the skewness and kurtosis of the queueing distribution via the Gram Charlier expansion. We show that the additional information captured in the skewness and kurtosis allows us to estimate the dynamics of the mean and variance much better than fluid and diffusion limit theorems or other methods that use only mean and variance behavior. Lastly, our approach also yields accurate approximations for the probability of delay, which is an important metric for quality of service.

Key words. multiserver queues, Gram Charlier expansion, Edgeworth expansion, asymptotics, abandonment, dynamical systems, time varying rates, Hermite polynomials, fluid and diffusion limits, skewness, kurtosis, cumulant moments, Gaussian, Berry–Esseen

AMS subject classifications. Primary, 60K25; Secondary, 60F17, 90B22

DOI. 10.1137/120896815

1. Introduction.

Motivated by the need for better approximations for the performance of small and medium sized service systems, such as emergency care centers and small data centers, we introduce a new, four-dimensional dynamical system approximation for queueing systems using the mean, variance, skewness, and kurtosis of these of dynamic rate Markov processes. Better approximations for these queueing models are needed as they help managers to optimally staff and accurately maintain quality of service metrics imposed by service level agreements. Since real service systems like call centers experience time varying behavior and large arrivals of customers and have multiple agents ready to deliver service, Markovian service networks are the class of time inhomogeneous stochastic processes that capture all these dynamics.

Our canonical queueing model assumes the customer arrival process is a nonhomogeneous Poisson process. We also have $c(t)$ servers at time t with i.i.d. (independent identically distributed) service times that are exponentially distributed with time dependent rate $\mu(t)$. Finally, all the customers have i.i.d. abandonment times that are also exponentially distributed with time varying rate $\beta(t)$. This model is known as the $M_t/M_t/C_t + M_t$ queueing model, where the $+M$ is included for abandonment. Using the functional strong law of large numbers (FSLLN) developed for our family of Markovian service networks in Mandelbaum, Massey, and Reiman [8], one can show that the limiting behavior of Markovian service networks can be described by a deterministic nonlinear ordinary differential equation. A more refined functional central limit theorem (FCLT), also developed in [8], yields that the behavior of the network can be described by a Gaussian diffusion that solves a stochastic differential equation.

The Gaussian diffusion from the FCLT relies on the fact that the amount of time that the mean number of customers is equal to the number of servers is of measure zero. However, when the mean number of customers lingers around the number of

[*] Received by the editors October 29, 2012; accepted for publication (in revised form) May 16, 2014; published electronically August 28, 2014.
http://www.siam.org/journals/siap/74-4/89681.html

[†] School of Operations Research and Information Engineering, Cornell University, West Windsor, NJ 08550 (jjp274@cornell.edu).

agents, the limiting diffusion is not as good an approximation of the unscaled queueing process. Numerical examples from Mandelbaum et al. [9] support these limit theorems for when the number of arrivals and the number of servers are large. However, the authors of [9] note that the fluid and diffusion limits do not approximate the simulated queueing dynamics well in the critically loaded regime where the number of servers is equal to the mean number in the system, i.e., $q = c$ or when the numbers of arrivals and servers are small. Some might think that the condition $q = c$ might have measure zero; however, if one considers optimal control of queueing systems via fluid limits, such as in the work of Hampshire and coworkers [3, 4], the optimal staffing policy $c^* = q$ forces the dynamics of the queueing system in the critical region. Thus, it is important to have accurate dynamics of the queueing process in order to yield reliable staffing procedures and policies.

To address some of these concerns Ko and Gautam [7] developed a Gaussian smoothing technique to better approximate the mean and variance of the original queueing system. Since the rate functions of the queueing system are not smooth, they propose mollifying them with a Gaussian density. Using this approach they show that they can improve the approximation of the mean behavior; however, they indicate that the variance still needs some improvement in the critically loaded regions. Thus, the method of [7] implicitly assumes that the distribution is symmetric. However, since the queueing process distribution lies on the positive real line and is unbounded, the distribution should be asymmetric. This would imply that the distribution has nonzero skewness and kurtosis values, which a Gaussian distribution cannot replicate.

The present work summarizes and extends [7] as follows. First, we summarize their results using the moment-forward equations for the actual queueing process. Second, we impose a new distribution on the queue length, which we believe to be quite natural given the work of [7]. In fact, our approximate distribution for the queue length is constructed by using a Gram Charlier expansion with a Gaussian reference density. We choose to use only the skewness and kurtosis terms as refinements as they are the most relevant to the queueing process behavior. We then use the skewness and kurtosis to correct the estimates of the mean and variance of the queue system in the critical regions where it is not approximated very well. Our simulation experiments show that our method outperforms the method of [7] significantly. We should also mention that recent work by Massey and Pender [10, 11] uses a Hermite polynomial expansion for the queueing process to model the non-Gaussian behavior of the queueing process. However, our work is different and complementary. Massey and Pender [11] provide a truncated Hermite polynomial L^2 expansion of the queueing process, while we provide an L^2 expansion of the queue length density. Using the Gram Charlier expansion allows us to explore higher cumulant moments using standard asymptotic techniques and yields much simpler approximations than the method of [11], which relies on the computation of polynomial roots. Moreover, our correction terms are linear in the skewness and kurtosis parameters, while they are nonlinear in [11]. Lastly, we explore the kurtosis of the queueing process, which is not explored in any of the previous literature.

In addition to estimating the time varying moments, we are able to estimate the probability of delay. To this end, there has been an explosion of research that is dedicated to refining square root staffing procedures or analyzing asymptotic expansions for Poisson processes; see, for instance, Janssen, van Leeuwaarden, and Zwart [5, 6] and Zhang et al. [16]. However, much of the research has focused on time homogeneous queueing models. This is because in the time homogeneous models, one can exploit rigorous asymptotic expansions for the Poisson process and use these asymptotic expansions for refining the probability of delay for the queueing models. For time inhomogeneous models, these asymptotic expansions are not warranted, and new approaches are needed to address the staffing issues. To address this we show that, using the Gram Charlier expansions with skewness and kurtosis corrections, we are better able to estimate the probability of delay in our time varying queueing systems. Finally, these approximations help us save substantial computational time since it is now unnecessary to simulate the queueing processes to obtain approximate performance measures. Instead, one can numerically integrate four differential equations and obtain accurate information regarding the stochastic behavior of our queueing model in much less time.

1.1. Contributions.

To the best of our knowledge our contributions in this work are the following:

- We obtain very accurate estimates for the mean and variance of the $M_t/M_t/C_t$ $+M_t$ queue in critical regions.
- We show how higher order moments of the $M_t/M_t/C_t + M_t$ queue can add valuable information for the mean and variance behavior and show that in some cases it is not sufficient to use fluid and diffusion limits without adjustments.
- We give explicit approximations for the mean, variance, skewness, kurtosis, and probability of delay for the $M_t/M_t/C_t + M_t$ queue via Gram Charlier expansions and reduce much of the stochastic dynamics to the numerical integration of four differential equations.

1.2. Organization of the paper.

The rest of the paper continues as follows. In section 2, we review our queueing model and the associated fluid and diffusion limits derived in [8]. We also provide expressions for the functional Kolmogorov forward equations for our queueing model. In section 3, we give a summary of previous methods that analyze the dynamics of mean and variance using these methods. In section 4, we give a summary of the Gram Charlier expansion and give insight into why we use it. In section 5, we illustrate how the skewness helps in estimating the mean and variance behavior of our queueing process. In section 6, we analyze how the kurtosis affects the estimation of the queueing process dynamics. In section 7, we show how to use the skewness and kurtosis corrections to estimate the probability of delay. In section 8, we give more numerical examples to show that our method works in a variety of parameter settings. In section 9, we give concluding remarks and some simple extensions to other types of Markovian queueing models that are important in the literature. Lastly, in Appendix A we provide the proofs of our main theorems and lemmas that are needed to construct the approximations of the paper.

2. Analysis of time varying Erlang-A model.

Motivated by small and medium sized health care centers and data centers, we study methods for approximating the transient and time varying behavior of a multiserver queue with abandonment. Mandelbaum, Massey, and Reiman [8] showed that the $M_t/M_t/C_t + M_t$ queueing system process $\{Q(t)|t \geq 0\}$ is represented by the following equation:

$$\begin{aligned} Q(t) = Q(0) + \Pi_1\left(\int_0^t \lambda(s)ds\right) - \Pi_2\left(\int_0^t \mu \cdot (Q(s) \wedge c(s))ds\right) \\ -\Pi_3\left(\int_0^t \beta \cdot (Q(s) - c(s))^+ ds\right), \end{aligned}$$

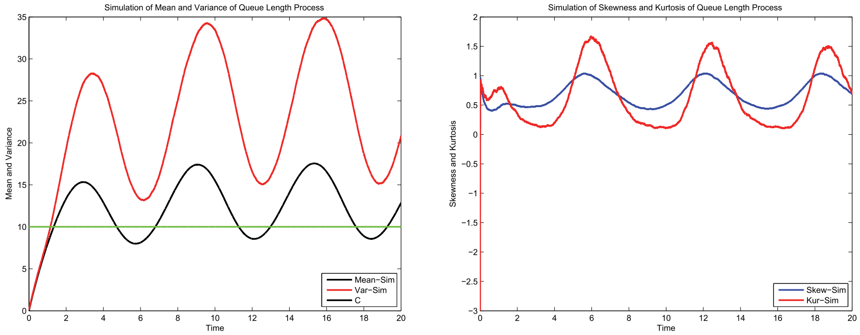

FIG. 1. *Left: Simulation of mean and variance of the queueing process. Right: Simulation of skewness and kurtosis of the queueing process.*

where $\mathit{\Pi}_i \equiv \{\mathit{\Pi}_i(t) | t \geq 0\}$ for $i = 1, 2, 3$ are i.i.d. standard (rate 1) Poisson processes. A deterministic time change for $\mathit{\Pi}_1$ transforms it into a nonhomogeneous Poisson arrival process with rate $\lambda(t)$. Thus, $\mathit{\Pi}_1(\int_0^t \lambda(s) ds)$ serves to count the number of customers that have arrived to the queue within the interval $(0, t]$. Moreover, if we subject $\mathit{\Pi}_2$ to a random time change, $\mathit{\Pi}_2$ counts the number of service departures from $c(s)$ servers and an exponentially distributed service times function of rate $\mu(t)$. Lastly, if we subject $\mathit{\Pi}_3$ to a random time change, then $\mathit{\Pi}_3$ represents the number of abandonments from $c(t)$ servers and exponentially distributed abandonment times of rate $\beta(t)$. Since our process is a linear combination of Poisson random measures and is Markovian, the $M_t/M_t/C_t + M_t$ queueing model is an example of a Markovian service network, which was studied extensively in [8]. The primary numerical example that we study in this paper to demonstrate the usefulness of our approximation methods has an arrival rate of $\lambda(t) = 10 + 5\sin(t)$, a service rate of $\mu = 1$, an abandonment rate of $\beta = .05$, and $c = 10$ servers. Moreover, we simulate our queueing model over the time interval (0, 20] for 10^5 independent sample paths.

In Figure 1(left) is a plot of the simulated mean $E[Q(t)]$ and variance $\text{Var}[Q(t)]$ of our queueing process. In Figure 1(right) we plot the simulated values of the skewness $\text{Skew}[Q(t)]$ and kurtosis $\text{Kur}[Q(t)]$ of the queueing system. The skewness and kurtosis are related to the third and fourth cumulant moments and are given by the formulas

(2.1)
$$\text{Skew}[Q(t)] = \frac{E[(Q(t) - E[Q(t)])^3]}{\text{Var}[Q(t)]^{3/2}} \quad \text{and} \quad \text{Kur}[Q(t)] = \frac{E[(Q(t) - E[Q(t)])^4]}{\text{Var}[Q(t)]^2} - 3.$$

As Figure 1(right) shows, the skewness and kurtosis are nonzero quantities. Since the skewness and kurtosis for a Gaussian random variable are defined to be zero, Figure 1(right) gives us supporting evidence that the queueing process distribution is non-Gaussian. However, one also observes from Figure 1 that while the skewness and kurtosis are nonzero, they are not extremely large quantities. Since they are not large, this gives us some confidence that using asymptotic expansions around a Gaussian distribution might be reasonable. Moreover, the skewness and kurtosis have the potential to give us valuable information about the properties of our queueing distribution. In fact, compared to a Gaussian distribution, the skewness can tell us whether the median of the queueing distribution is to the left or right of the mean of the distribution, and the kurtosis can provide information on the peakedness of the distribution. The skewness is especially important since the real queueing process is nonnegative, unbounded, and asymmetric, while the Gaussian distribution can realize negative values and is symmetric around the mean. Thus, the skewness is critical in capturing asymmetries of the queueing distributions. Although the skewness and kurtosis are important statistical and mathematical quantities, they also have some practical value because they can help managers adjust or refine the staffing levels appropriately according to the information about the values of the skewness and kurtosis. In fact when the skewness and kurtosis are near zero, they validate the use of the Gaussian approximations. However, when they are away from zero, they can serve to refine Gaussian behavior predicted from rigorous limit theorems.

To understand the time varying dynamics of our queueing model it is necessary to study its cumulant moment behavior. In this paper, we study the time derivatives of the mean, the variance, the third cumulant moment, and the fourth cumulant moment, which in turn gives us important information about the average number of customers in the queue, the deviations from the mean, whether the queueing distribution is skewed to the left or right of the mean, and the peakedness of the queueing distribution, respectively. To study the behavior of the cumulant moments, we choose to use the functional version of the Kolmogorov forward equations for the $M_t/M_t/C_t + M_t$ queue, which are of the form

$$\begin{aligned}\overset{\bullet}{E}[f(Q)] &= \lambda \cdot E[f(Q+1) - f(Q)] + \mu \cdot E[(Q \wedge c) \cdot (f(Q-1) - f(Q))] \\ &\quad + \beta \cdot E[(Q-c)^+ \cdot (f(Q-1) - f(Q))].\end{aligned}$$

For the special cases of the mean, variance, third cumulant moment, and fourth cumulant moment, we obtain the following set of differential equations:

$$\begin{aligned}
\overset{\bullet}{E}[Q] &= \lambda - \mu \cdot E[Q \wedge c] - \beta \cdot E[(Q-c)^+], \\
\overset{\bullet}{\mathrm{Var}}[Q] &= \lambda + \mu \cdot E[Q \wedge c] + \beta \cdot E[(Q-c)^+] - 2\left(\mu \cdot \mathrm{Cov}[Q, Q \wedge c] + \beta \cdot \mathrm{Cov}[Q, (Q-c)^+]\right), \\
\overset{\bullet}{\mathrm{C}}^{[3]}[Q] &= \lambda - \mu \cdot E[Q \wedge c] - \beta \cdot E[(Q-c)^+] + 3\left(\mu \cdot \mathrm{Cov}[Q, Q \wedge c] + \beta \cdot \mathrm{Cov}[Q, (Q-c)^+]\right) \\
&\quad - 3\left(\mu \cdot \mathrm{Cov}\left[\overline{Q}^2, Q \wedge c\right] + \beta \cdot \mathrm{Cov}\left[\overline{Q}^2, (Q-c)^+\right]\right), \\
\overset{\bullet}{C}^{[4]}[Q] &= \lambda + \mu \cdot E\left[Q \wedge c\right] + \beta \cdot E\left[(Q-c)^+\right] \\
&\quad - 4 \cdot \left(\mu \cdot \mathrm{Cov}\left[Q, Q \wedge c\right] + \beta \cdot \mathrm{Cov}\left[Q, (Q-c)^+\right]\right) \\
&\quad + 6 \cdot \left(\mu \cdot \mathrm{Cov}\left[\overline{Q}^2, Q \wedge c\right] + \beta \cdot \mathrm{Cov}\left[\overline{Q}^2, (Q-c)^+\right]\right) \\
&\quad - 4 \cdot \left(\mu \cdot \mathrm{Cov}\left[\overline{Q}^3, Q \wedge c\right] + \beta \cdot \mathrm{Cov}\left[\overline{Q}^3, (Q-c)^+\right]\right) \\
&\quad + 12 \cdot \left(\mu \cdot \mathrm{Var}[Q] \cdot \mathrm{Cov}\left[Q, Q \wedge c\right] + \beta \cdot \mathrm{Var}[Q] \cdot \mathrm{Cov}\left[Q, (Q-c)^+\right]\right),
\end{aligned}$$

where $\overline{Q} = Q - E[Q]$.

In addition to having an expression for the dynamics for the cumulant moments for the queueing process, one can also derive several asymptotic limits for the Markovian multiserver queue with abandonment. In the seminal paper [8] it was shown that the Markovian multiserver queue with abandonment has fluid and diffusion limits, i.e.,

$$\frac{1}{\eta} Q^{\eta} \;=\; q \;\text{ a.s. u.o.c.} \quad \text{and} \quad \sqrt{\eta} \cdot \left(\frac{1}{\eta} Q^{\eta} - q \right) \overset{d}{\Rightarrow} \hat{Q}, \tag{2.2}$$

where the fluid mean q is governed by the one-dimensional dynamical system

$$\overset{\bullet}{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q-c)^+. \tag{2.3}$$

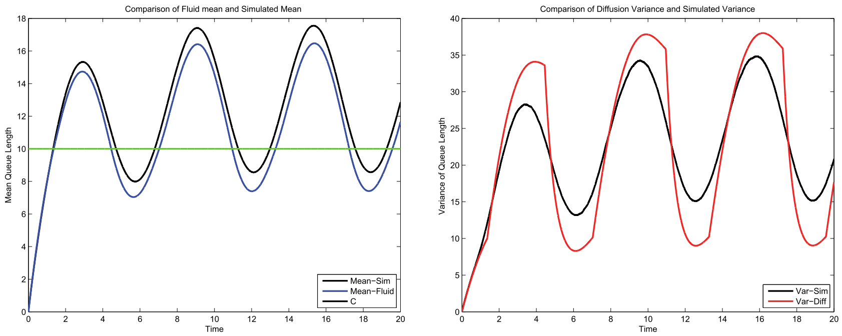

FIG. 2. *Left: Comparisons of simulated mean and its fluid limit. Right: Comparisons of simulated variance and diffusion limit.*

It can be shown that the fluid mean q has three basic modes of operation. When the mean number of the system represented by q is less than the number of servers c, we say that the system is *underloaded*. Conversely, when q is greater than the number of servers c, we say that the system is *overloaded*. Lastly, when $q = c$, we say the system is *critically loaded*. Moreover, if we make the assumption that the set $\{t | q(t) = c\}$ has measure zero, then $\hat{Q}$ is a Gaussian diffusion whose variance $v \equiv \mathrm{Var}[\hat{Q}]$ combines with the fluid mean to form a two-dimensional dynamical system given by (2.3) and

$$\overset{\bullet}{v} = \lambda + \mu \cdot (q \wedge c) + \beta \cdot (q - c)^{+} - 2 \cdot v \cdot (\mu \cdot \{q < c\} + \beta \cdot \{q \geq c\}) \,, \tag{2.4}$$

where $\{q < c\}$ denotes an indicator function equaling one if $q < c$ and zero otherwise.

Remark 2.1. It is important to note that the fluid and diffusion limits are a partially coupled system. Although the variance is a function of the mean, the mean or fluid limit equation (2.3) is independent of the variance behavior. Thus, for small η, where it might be useful for the mean dynamics to have some information about the distributional behavior of the queueing process, the fluid limit does not depend on any information other than the mean behavior of the queueing process, which is quite limiting.

Remark 2.2. When our queueing system is extremely underloaded it behaves like an infinite server queue, which has a Poisson transient distribution. A Poisson distribution is characterized by all of its cumulant moments equaling its mean, which forces the fluid mean and the diffusion variance to be equal, i.e., $q = v$. In the context of dispersion theory, the underloaded system is neither overdispersed or underdispersed. However, in other regions where the queue does not behave like an infinite server queue, we do not in general have that $q = v$. Thus, when $q > v$, we have that the queue is underdispersed, and when $q < v$, the queue length distribution is overdispersed.

In the left and right of Figure 2, we compare simulations of the mean and variance of the queueing process to its fluid limit and diffusion limit, respectively. We use the fluid limit as an approximation for the mean and the diffusion limit variance as an approximation to the variance. In this example, the fluid limit is a good approximation to the dynamics of the mean, but seems to underestimate the queue length most of the time. The diffusion variance does not work quite as well, especially the *kinked* areas where the simulated mean is equal to the number of servers. Thus, for small values of η, the fluid and diffusion limits may not be the best approximations for estimating the dynamics of the queueing system as they were intended for large values of η. This inaccuracy in the small η case motivates the rest of the paper.

3. Summary of previous methods.

Now armed with the Kolmogorov forward equations, we are ready to use them to model the stochastic behavior of our queueing model. However, upon inspection of the forward equations, one notices that they are not an autonomous system unless $c = \infty$. This means that we cannot compute the expectations or covariance terms that define the dynamics. Thus, we are forced to assume an underlying distribution for our stochastic queueing process. If we assume a particular distribution, we are able to compute the expectation and covariance terms and close the system, thus making it autonomous. This closure approximation is not new in the queueing theory literature; see, for instance, Rothkopf and Oren [13] as well as Taaffe and Ong [15]. However, one reason our closure approximation method is different is that it is motivated from the fluid and diffusion limits. Moreover, we use a continuous, rather than a discrete, distribution for our underlying queueing distribution. This is also motivated from the fluid and diffusion limits, as the limiting distribution is Gaussian under appropriate scalings. Similar to [11], we will motivate our estimation procedures via a series of approximations that are special cases of our general method.

3.1. Deterministic mean approximation.

For the first approximation of our queueing process, we define the *deterministic mean approximation* (DMA) by assuming $\{q(t)|t \geq 0\}$ is a deterministic process that approximates the queueing process. In this approximation, we assume that $Q \approx q$, and we replace Q in the Kolmogorov forward equation by the mean of Q, which is q. Thus, the mean solves the following autonomous, one-dimensional, dynamical system:

$$\overset{\bullet}{q} = \lambda - \mu \cdot (q \wedge c) - \beta \cdot (q - c)^{+}. \tag{3.1}$$

Since we approximate the stochastic process by a deterministic one, we implicitly assume that the higher cumulant moments are identically zero. Moreover, one should note that the deterministic mean approximation is also equivalent to the fluid limit given by (2.3). Although the DMA gives us information about the predictable variation of the stochastic queueing process, it does not give us any information about the variations of our queueing process since it implicitly assumes that the variations are zero.

3.2. Gaussian variance approximation.

Unlike the DMA, we now assume that our queueing model follows a Gaussian distribution. This approximation, i.e., the *Gaussian variance approximation* (GVA), was first developed by Ko and Gautam [7], and further expanded by Massey and Pender [11], although we should mention that [7] did not use the functional forward equations to derive their approximation. In [11], it was shown that the functional forward equations approach is equivalent to the Gaussian mollifier approach used by [7]. The Gaussian distribution assumption on the queue length is equivalent to

$$Q(t) \overset{d}{=} q(t) + X \cdot \sqrt{v(t)} \tag{3.2}$$

for all $t \geq 0$, where $\{q(t), v(t)|t \geq 0\}$ is some two-dimensional dynamical system where the v process is always positive and X is a standard Gaussian random variable. The functional forward equations for the mean and variance of the queue length process Q are

$$\overset{\bullet}{E}[Q] = \lambda - \left(\mu \cdot E[Q \wedge c] + \beta \cdot E[(Q - c)^{+}]\right), \tag{3.3}$$

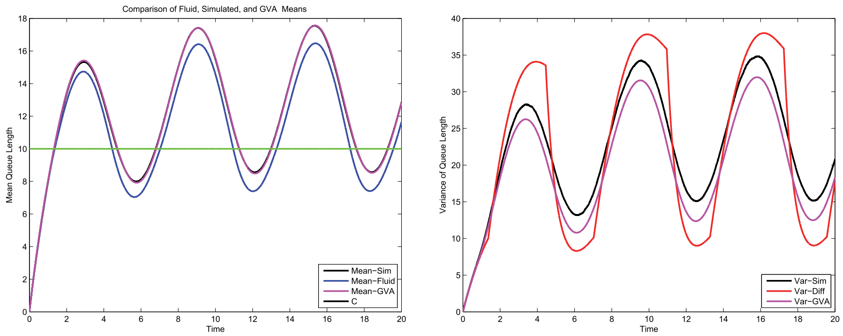

FIG. 3. *Left: Comparison of simulation, fluid, and GVA means. Right: Comparison of simulation, diffusion, and GVA variances.*

$$\overset{\bullet}{\mathrm{Var}}[Q] = \lambda + \mu \cdot E[Q \wedge c] + \beta \cdot E[(Q-c)^+] - 2\left(\mu \cdot \mathrm{Cov}[Q, Q\wedge c] + \beta \cdot \mathrm{Cov}[Q,(Q-c)^+]\right). \tag{3.4}$$

Now if we define the variable $\chi = \frac{c-q}{\sqrt{v}}$, we get the following differential equations for the mean and variance of the queueing process under the distributional assumptions of GVA:

$$\overset{\bullet}{E}[Q] = \lambda - \left(\mu \cdot \sqrt{v} \cdot E[X \wedge \chi] + \beta \cdot \sqrt{v} \cdot E[(X-\chi)^+]\right),$$

$$\begin{aligned}\overset{\bullet}{\mathrm{Var}}[Q] = {} & \lambda + \mu \cdot \sqrt{v} \cdot E[X \wedge \chi] + \beta \cdot \sqrt{v} \cdot E[(X-\chi)^+] \\ & - 2\left(\mu \cdot v \cdot \mathrm{Cov}[X, X \wedge \chi] + \beta \cdot v \cdot \mathrm{Cov}[X,(X-\chi)^+]\right).\end{aligned}$$

Thus, in order to understand the dynamics via numerical integration, it only remains to compute closed form expressions for the expectations and covariance terms in the cumulant moment forward equations. Like [11], we resort to using Stein's lemma to derive the expectations and covariance terms. The use of Stein's lemma [14] yields the following equations for the mean and variance dynamics of our queueing process:

$$\begin{aligned}\overset{\bullet}{E}[Q] &= \lambda - \mu \cdot q + \sqrt{v} \cdot (\mu - \beta) \cdot \left(\phi(\chi) - \chi \cdot \overline{\Phi}(\chi)\right), \\ \overset{\bullet}{\mathrm{Var}}[Q] &= \lambda + \mu \cdot q - \sqrt{v} \cdot (\mu - \beta) \cdot \left(\phi(\chi) - \chi \cdot \overline{\Phi}(\chi)\right) - 2 \cdot v \cdot \left(\mu \cdot \Phi(\chi) + \beta \cdot \overline{\Phi}(\chi)\right).\end{aligned}$$

Unlike the DMA, the GVA derives equations that are different from the fluid and diffusion limits. Not only are the mean and variance equations different, but also they are fully coupled to one another. This implies that the mean is incorporating information from the variance. Thus, we expect that the mean dynamics of the GVA should be different and better than the fluid limits or the DMA.

In Figure 3, we compare the fluid and diffusion limits derived in [8] with the GVA method, and we see that GVA does a better job of estimating the mean and variance dynamics of the true queueing process. It improves the variance and does a better job at approximating the mean behavior. This confirms the fact that the mean equation is obtaining more information by using the distributional information of the queueing model. Although GVA does a much better job of approximating the true behavior of our queueing model than the fluid and diffusion limits, it still needs some improvement for its estimate for the variance. The inaccuracy of the variance motivates the next section which introduces our method, the Gram Charlier series expansion.

4. Gram Charlier expansion technique.

In this section, we introduce our new approximation method for estimating the time varying dynamics of our queueing process. We do not know the true density of the time varying queueing process, but our idea is to use a statistical series expansion of the queueing distribution to derive an approximation of the true queueing process distribution. Suppose that the true distribution of our queueing process is a continuous function $\tau(x)$. Using this distribution, we can calculate the moment generating function as

$$M(t) = \int_{-\infty}^{\infty} \tau(x) e^{tx} dx. \tag{4.1}$$

Moreover, the moment generating function has a Taylor series expansion of the form

$$M(t) = \sum_{n=0}^{\infty} m_n \cdot \frac{t^n}{n!}, \tag{4.2}$$

where the moments m_n are defined as

$$m_n = \int_{-\infty}^{\infty} x^n \cdot \tau(x) dx. \tag{4.3}$$

Although the moments are important quantities, it is often convenient to use the cumulant generating function $K(t)$, which is the *logarithm* of the moment generating function, i.e.,

$$K(t) = \log\left(M(t)\right). \tag{4.4}$$

Like the moment generating function, the cumulant generating function also has a Taylor series expansion, which is of the form

$$K(t) = \sum_{n=1}^{\infty} \kappa_n \cdot \frac{t^n}{n!}. \tag{4.5}$$

In fact, by differentiating in the time variable and setting $t = 0$, we see that moments are related to the cumulant moments by the following expression:

$$m_{n+1} = \sum_{j=0}^{n} \binom{n}{j} m_{n-j} \cdot \kappa_{j+1}. \tag{4.6}$$

Now that we have an understanding of the cumulant generating function, we can use it to construct the Gram Charlier expansion using a finite number of terms. Using Fourier transform techniques, we can write the true density as

$$\begin{aligned} \tau(x) &= \mathcal{F}^{-1}\left(M(it)\right) && (4.7)\\ &= \frac{1}{2\cdot\pi}\int_{-\infty}^{\infty} M(it)\cdot e^{-itx} dt, && (4.8)\end{aligned}$$

where $\mathcal{F}^{-1}$ is defined as the inverse Fourier transform and where we now define $M(it)$ as

$$M(it) = \int_{-\infty}^{\infty} \tau(x) e^{itx} dx. \tag{4.9}$$

Now if we use the cumulant moment representation for $M(it)$ and note that the true probability distribution is a real function, then we have that the true distribution has the representation

$$\begin{aligned}
\tau(x) &= \frac{1}{2\cdot\pi}\int_{-\infty}^{\infty} M(it)\cdot e^{-itx}dt \\
&= \frac{1}{2\cdot\pi}\int_{-\infty}^{\infty} \exp\left(\sum_{n=1}^{\infty}\kappa_n\cdot\frac{(it)^n}{n!}\right)\cdot e^{-itx}dt \\
&= \frac{1}{2\cdot\pi}\int_{-\infty}^{\infty} \exp\left(\sum_{j=1}^{\infty}(-1)^j\cdot\kappa_{2j}\cdot\frac{t^{2j}}{(2j)!}\right)\cdot\cos\left(xt+\sum_{j=1}^{\infty}(-1)^j\cdot\kappa_{2j-1}\cdot\frac{t^{2j-1}}{(2j-1)!}\right)dt \\
&= \frac{1}{\pi}\int_{0}^{\infty} \exp\left(\sum_{j=1}^{\infty}(-1)^j\cdot\kappa_{2j}\cdot\frac{t^{2j}}{(2j)!}\right)\cdot\cos\left(xt+\sum_{j=1}^{\infty}(-1)^j\cdot\kappa_{2j-1}\cdot\frac{t^{2j-1}}{(2j-1)!}\right)dt.
\end{aligned}$$

Now suppose that only the first two cumulant moments are nonzero. This gives us the following integral representation for the true distribution as

$$\begin{aligned}
\tau(x) &= \frac{1}{\pi}\int_{0}^{\infty} \exp\left(-\frac{1}{2}\kappa_2\cdot t^2\right)\cdot\cos\left(xt-\kappa_1\cdot t\right)dt \\
&= \frac{1}{\sqrt{2\pi\cdot\kappa_2}}\exp\left(-\frac{1}{2}\left(\frac{x-\kappa_1}{\sqrt{\kappa_2}}\right)^2\right),
\end{aligned}$$

which is the Gaussian distribution with mean κ_1 and standard deviation κ_2.

However, it is often the case that two cumulant moments are not sufficient to accurately approximate a distribution; thus if we add more cumulant moments, we get the following expression for the true density $\tau(x)$ by using the properties of the exponential function and Euler's formula:

$$\begin{aligned}
\tau(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(\sum_{j=1}^{\infty}(-1)^j\cdot\kappa_{2j}\cdot\frac{t^{2j}}{(2j)!}\right)\cdot\cos\left(xt-\sum_{j=1}^{\infty}(-1)^j\cdot\kappa_{2j-1}\cdot\frac{t^{2j-1}}{(2j-1)!}\right)dt \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-\frac{\kappa_2 t^2}{2}\right)\cdot\cos\left(xt-\kappa_1\cdot t\right)\cdot\exp\left(\sum_{j=2}^{\infty}(-1)^j\cdot\kappa_{2j}\cdot\frac{t^{2j}}{(2j)!}\right) \\
&\quad \cdot\cos\left(\sum_{j=2}^{\infty}(-1)^j\cdot\kappa_{2j-1}\cdot\frac{t^{2j-1}}{(2j-1)!}\right)dt \\
&\quad -\frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-\frac{\kappa_2 t^2}{2}\right)\cdot\sin\left(xt-\kappa_1\cdot t\right)\cdot\exp\left(\sum_{j=2}^{\infty}(-1)^j\cdot\kappa_{2j}\cdot\frac{t^{2j}}{(2j)!}\right) \\
&\quad \cdot\sin\left(\sum_{j=2}^{\infty}(-1)^j\cdot\kappa_{2j-1}\cdot\frac{t^{2j-1}}{(2j-1)!}\right)dt.
\end{aligned}$$

Now to derive the Gram Charlier expansion, we must expand the functions

$$\exp\left(\sum_{j=2}^{\infty}(-1)^j\cdot\kappa_{2j}\cdot\frac{t^{2j}}{(2j)!}\right)\cdot\cos\left(\sum_{j=2}^{\infty}(-1)^j\cdot\kappa_{2j-1}\cdot\frac{t^{2j-1}}{(2j-1)!}\right) \tag{4.10}$$

and

$$\exp\left(\sum_{j=2}^{\infty}(-1)^j\cdot\kappa_{2j}\cdot\frac{t^{2j}}{(2j)!}\right)\cdot\sin\left(\sum_{j=2}^{\infty}(-1)^j\cdot\kappa_{2j-1}\cdot\frac{t^{2j-1}}{(2j-1)!}\right) \tag{4.11}$$

using a Taylor series expansion around $t = 0$. By using this expansion and setting to zero all cumulant moment terms greater than four, we derive the Gram Charlier expansion

$$\tau(x)\approx\phi(x)+\frac{\kappa_3}{3!\cdot\kappa_2^{1.5}}\cdot h_3(x)\cdot\phi(x)+\frac{\kappa_4}{4!\cdot\kappa_2^2}\cdot h_4(x)\cdot\phi(x), \tag{4.12}$$

where $h_j(x)$ is the jth Hermite polynomial and ϕ is the Gaussian density. A brief introduction to the Hermite polynomials is given in Appendix A.

4.1. Convergence of the Gram Charlier expansion. It is important to note that the error of the expansion in (4.12) does not necessarily converge uniformly to zero as we add more terms to the series. This implies that (4.12) is not a true asymptotic expansion. However, there exists a rigorous asymptotic series expansion where the error term does converge uniformly to zero as we add more terms. This expansion is known as the Edgeworth series expansion and can be expressed as

$$\begin{aligned}\tau(x)=\phi(x)&+\frac{\kappa_3}{3!\cdot\kappa_2^{1.5}}\cdot h_3(x)\cdot\phi(x)+\frac{\kappa_4}{4!\cdot\kappa_2^2}\cdot h_4(x)\cdot\phi(x)\\&+\frac{10\cdot\kappa_3^2}{6!\cdot\kappa_2^3}\cdot h_6(x)\cdot\phi(x)+\epsilon(x).\end{aligned} \tag{4.13}$$

Although the two expansions given in (4.12) and (4.13) differ by one term, neither expansion has theoretical superiority over the other due to the different moment assumptions needed to derive the expansion. However, for the remainder of the paper, we will consider (4.12) as the approximation to our true density.

5. Gram Charlier skewness approximation. In this section, we introduce our first new method for modeling the dynamics of our queueing process called the *Gram Charlier skewness approximation* (GCS). The Gram Charlier expansion method has been used successfully in the mathematical finance literature to accurately price stock options using the Black–Scholes model as a base model; see, for example, Corrado and Su [1], who use the skewness seen in financial returns to correctly price option prices seen in the current market. In the same spirit our method is a natural extension of the GVA method because we expand the true density around the Gaussian density. Our goal is to use the skewness of the queueing process to correct the mean and variance behavior of the queueing dynamics. Using the same methodology as [1], we assume that our queueing process under the GCS method has the following approximate density:

$$\phi_{Skew}(x)=\phi(x)\cdot\left(1+\frac{\kappa_3}{3!\cdot\sqrt{v^3}}\cdot h_3(x)\right)=\phi_{GVA}(x)+\phi_{GCS}(x), \tag{5.1}$$

where $\{q, v, \kappa_3\}$ are the mean, variance, and third cumulant moment of the queueing process and $h_3(x)$ is a Hermite polynomial of order 3. This representation of the

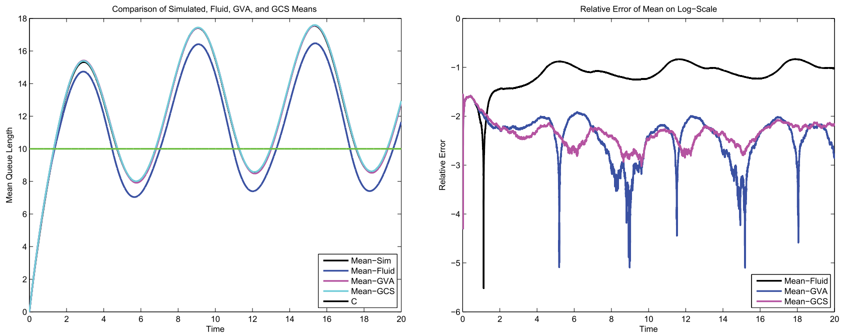

FIG. 4. *Left: Comparison of simulated, GVA, and GCS means. Right: Relative error of means using GVA and GCS approximations.*

density also shows that the GCS method is a perturbation of the GVA method, which adds the skewness of the queueing system. We will show that the skewness will allow us to better estimate the mean and variance dynamics of the queueing system with our first main theorem.

THEOREM 5.1. *Using the density given in* (5.1) *for the distribution of our queueing system, we have the following equations for the mean, variance, and third cumulant moment of our multiserver queue with abandonment:*

$$\overset{\bullet}{E}[Q] = \lambda - \mu \cdot E[Q \wedge c] - \beta \cdot E[(Q-c)^+],$$

$$\overset{\bullet}{\mathrm{Var}}[Q] = \lambda + \mu \cdot E[Q \wedge c] + \beta \cdot E[(Q-c)^+] - 2\left(\mu \cdot \mathrm{Cov}[Q, Q \wedge c] + \beta \cdot \mathrm{Cov}[Q, (Q-c)^+]\right),$$

$$\overset{\bullet}{C}^{[3]}[Q] = \lambda - \mu \cdot E[Q \wedge c] - \beta \cdot E[(Q-c)^+] + 3\left(\mu \cdot \mathrm{Cov}[Q, Q \wedge c] + \beta \cdot \mathrm{Cov}[Q, (Q-c)^+]\right) \\ -3\left(\mu \cdot \mathrm{Cov}\left[\overline{Q}^2, Q \wedge c\right] + \beta \cdot \mathrm{Cov}\left[\overline{Q}^2, (Q-c)^+\right]\right),$$

where we have the following expressions for the unknown expectations and covariances:

$$E\left[(Q \wedge c)\right] = q\sqrt{v} \cdot \phi(\chi) + \chi \cdot \sqrt{v} \cdot \overline{\Phi}(\chi) - \frac{\chi \cdot \phi(\chi) \cdot \kappa_3}{6 \cdot v},$$

$$E\left[(Q-c)^+\right] = \sqrt{v} \cdot \phi(\chi) - \chi \cdot \sqrt{v} \cdot \overline{\Phi}(\chi) + \frac{\chi \cdot \phi(\chi) \cdot \kappa_3}{6 \cdot v},$$

$$\mathrm{Cov}\left[Q, (Q-c)^+\right] = v \cdot \overline{\Phi}(\chi) + \frac{(\chi^2+2) \cdot \phi(\chi) \cdot \kappa_3}{6\sqrt{v}},$$

$$\mathrm{Cov}\left[Q, (Q \wedge c)\right] = v \cdot \Phi(\chi) - \frac{(\chi^2+2) \cdot \phi(\chi) \cdot \kappa_3}{6\sqrt{v}},$$

$$\mathrm{Cov}\left[\overline{Q}^2, (Q-c)^+\right] = \sqrt{v^3} \cdot \phi(\chi) + \frac{\kappa_3}{6} \cdot \left[(\chi^3 + 4 \cdot \chi) \cdot \phi(\chi) + 6 \cdot \overline{\Phi}(\chi)\right],$$

$$\mathrm{Cov}\left[\overline{Q}^2, (Q \wedge c)\right] = \kappa_3 - \sqrt{v^3} \cdot \phi(\chi) - \frac{\kappa_3}{6} \cdot \left[(\chi^3 + 4 \cdot \chi) \cdot \phi(\chi) + 6 \cdot \overline{\Phi}(\chi)\right].$$

Proof. See Appendix A. □

5.1. GCS numerical results. In Figure 4(left), we compare the simulated mean with its estimates using GVA and GCS methods. Our new GCS method does

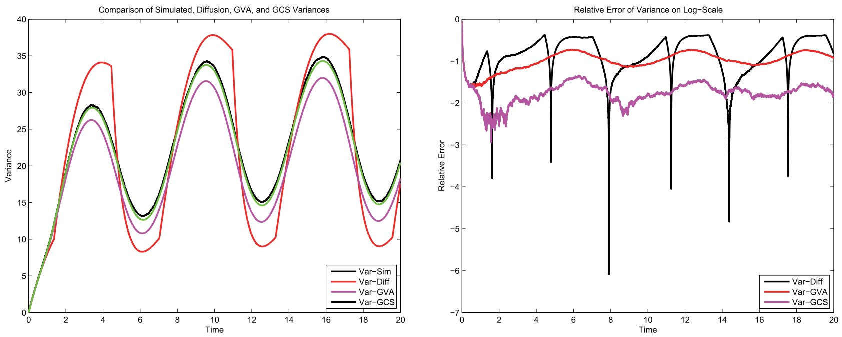

FIG. 5. *Left: Comparison of simulated, GVA, and GCS variances. Right: Relative error of variances using GVA and GCS approximations.*

a slightly better job at estimating the mean behavior of the queueing process. This slight improvement is easier to see with the graph in Figure 4(right) which plots over time the logarithm (base 10) of the relative errors of both GVA and GCS versus the simulated mean.

In Figure 5(left), we also compare the simulated variance with its estimates using the GVA and GCS methods. Unlike the mean behavior, we can even see an improvement of the estimates of the variance without resorting to the relative error. However, we also show in Figure 5(right) that the GCS method does a significantly better job of approximating the variance dynamics of the queueing system when we use the relative error comparison. These results show that the skewness provides valuable information about the mean and variance of our stochastic process and should be taken into consideration more often.

Lastly, in Figure 6 we compute the L_1 norm of the error over time. It is easy to see that the GCS method does a better job than GVA of reducing the error from approximating the real system over longer periods of time. It is a slight improvement for the mean behavior, but it is a significant improvement for the variance. It is apparent that the skewness can add valuable information to the understanding of the dynamics of our queueing system. We expect that the kurtosis will add a smaller amount of information about our queueing system, but how small is it?

6. Gram Charlier kurtosis approximation.

For this section, we again add another term to our Gram Charlier expansion to capture the kurtosis of our queueing system. We call this new approximation the *Gram Charlier kurtosis approximation* (GCK). Similarly to the GCS method, we hope that adding another term will further refine our approximations for the mean, variance, and skewness of the queueing model. This will help us attain even better estimates for the mean, variance, and skewness, which can be used for better staffing and optimization purposes. For the GCK method, we assume that our queueing process has the following approximate density:

$$\begin{aligned} \phi_{Kur}(x) &= \phi(x) \cdot \left(1 + \frac{\kappa_3}{3! \cdot \sqrt{v^3}} \cdot h_3(x) + \frac{\kappa_4}{4! \cdot v^2} \cdot h_4(x) \right) \\ &= \phi_{GVA}(x) + \phi_{GCS}(x) + \phi_{GCK}(x). \end{aligned} \tag{6.1}$$

Using the GCK approximation as the model for our queueing dynamics allows us to give our next main approximation result.

THEOREM 6.1. *Using the approximate density given in* (6.1), *we have the following equations for the mean, variance, third cumulant moment, and fourth moment of* *our multiserver queue with abandonment:*

$$
\begin{aligned}
\overset{\bullet}{E}[Q] &= \lambda - \mu \cdot E[Q \wedge c] - \beta \cdot E[(Q-c)^+], \\
\overset{\bullet}{\text{Var}}[Q] &= \lambda + \mu \cdot E[Q \wedge c] + \beta \cdot E[(Q-c)^+] - 2\left(\mu \cdot \text{Cov}[Q, Q \wedge c] + \beta \cdot \text{Cov}[Q, (Q-c)^+]\right), \\
\overset{\bullet}{C}^{[3]}[Q] &= \lambda - \mu \cdot E[Q \wedge c] - \beta \cdot E[(Q-c)^+] + 3\left(\mu \cdot \text{Cov}[Q, Q \wedge c] + \beta \cdot \text{Cov}[Q, (Q-c)^+]\right) \\
&\quad - 3\left(\mu \cdot \text{Cov}\left[\overline{Q}^2, Q \wedge c\right] + \beta \cdot \text{Cov}\left[\overline{Q}^2, (Q-c)^+\right]\right), \\
\overset{\bullet}{C}^{[4]}[Q] &= \lambda + \mu \cdot E\left[Q \wedge c\right] + \beta \cdot E\left[(Q-c)^+\right] \\
&\quad - 4 \cdot \left(\mu \cdot \text{Cov}\left[Q, Q \wedge c\right] + \beta \cdot \text{Cov}\left[Q, (Q-c)^+\right]\right) \\
&\quad + 6 \cdot \left(\mu \cdot \text{Cov}\left[\overline{Q}^2, Q \wedge c\right] + \beta \cdot \text{Cov}\left[\overline{Q}^2, (Q-c)^+\right]\right) \\
&\quad - 4 \cdot \left(\mu \cdot \text{Cov}\left[\overline{Q}^3, Q \wedge c\right] + \beta \cdot \text{Cov}\left[\overline{Q}^3, (Q-c)^+\right]\right) \\
&\quad + 12 \cdot \left(\mu \cdot \text{Var}[Q] \cdot \text{Cov}\left[Q, Q \wedge c\right] + \beta \cdot \text{Var}[Q] \cdot \text{Cov}\left[Q, (Q-c)^+\right]\right),
\end{aligned}
$$

where we have the following expressions for the unknown expectations and covariances:

$$
\begin{aligned}
E\left[(Q \wedge c)\right] &= q - \sqrt{v} \cdot \phi(\chi) + \chi \cdot \sqrt{v} \cdot \overline{\Phi}(\chi) - \frac{\chi \cdot \phi(\chi) \cdot \kappa_3}{6 \cdot v} - \frac{(\chi^2 - 1) \cdot \phi(\chi) \cdot \kappa_4}{6 \cdot \sqrt{v^3}}, \\
E\left[(Q-c)^+\right] &= \sqrt{v} \cdot \phi(\chi) - \chi \cdot \sqrt{v} \cdot \overline{\Phi}(\chi) + \frac{\chi \cdot \phi(\chi) \cdot \kappa_3}{6 \cdot v} + \frac{(\chi^2 - 1) \cdot \phi(\chi) \cdot \kappa_4}{6 \cdot \sqrt{v^3}}, \\
\text{Cov}\left[Q, (Q-c)^+\right] &= v \cdot \overline{\Phi}(\chi) + \frac{(\chi^2 + 2) \cdot \phi(\chi) \cdot \kappa_3}{6 \cdot \sqrt{v}} + \frac{(\chi^3 + \chi) \cdot \phi(\chi) \cdot \kappa_4}{24 \cdot v}, \\
\text{Cov}\left[Q, (Q \wedge c)\right] &= v \cdot \Phi(\chi) - \frac{(\chi^2 + 2) \cdot \phi(\chi) \cdot \kappa_3}{6 \cdot \sqrt{v}} - \frac{(\chi^3 + \chi) \cdot \phi(\chi) \cdot \kappa_4}{24 \cdot v}, \\
\text{Cov}\left[\overline{Q}^2, (Q-c)^+\right] &= \sqrt{v^3} \cdot \phi(\chi) + \frac{\kappa_3}{6} \cdot \left[(\chi^3 + 4 \cdot \chi) \cdot \phi(\chi) + 6 \cdot \overline{\Phi}(\chi)\right] \\
&\quad + \frac{\kappa_4}{24 \cdot \sqrt{v}} \cdot (\chi^4 + 3 \cdot \chi^2 + 6) \cdot \phi(\chi), \\
\text{Cov}\left[\overline{Q}^2, (Q \wedge c)\right] &= \kappa_3 - \sqrt{v^3} \cdot \phi(\chi) - \frac{\kappa_3}{6} \cdot \left[(\chi^3 + 4 \cdot \chi) \cdot \phi(\chi) + 6 \cdot \overline{\Phi}(\chi)\right] \\
&\quad - \frac{\kappa_4}{24 \cdot \sqrt{v}} \cdot (\chi^4 + 3 \cdot \chi^2 + 6) \cdot \phi(\chi), \\
\text{Cov}\left[\overline{Q}^3, (Q-c)^+\right] &= v^2 \cdot \left((\chi^2 + 1) \cdot \phi(\chi)\right) + 3 \cdot v^2 \cdot \overline{\Phi}(\chi) \\
&\quad + \frac{\kappa_3 \cdot \sqrt{v}}{6} \cdot \left((h_4(\chi) + 12 \cdot h_2(\chi) + 27) \cdot \phi(\chi)\right) \\
&\quad + \frac{\kappa_4}{24 \cdot v^2} \cdot \sqrt{v^3} \cdot \left((h_5(\chi) + 15 \cdot h_3(\chi) + 48 \cdot h_1(\chi)) \cdot \phi(\chi) + 24 \cdot \overline{\Phi}(\chi)\right), \\
\text{Cov}\left[\overline{Q}^3, (Q \wedge c)\right] &= 3 \cdot v^2 + \kappa_4 - \text{Cov}\left[\overline{Q}^3, (Q-c)^+\right].
\end{aligned}
$$

Proof. See Appendix A. ∎

6.1. GCK numerical results. In Figure 7(left), we compare the simulated mean with its estimates using GCS and GCK methods. Our new GCK method does a slightly better job at estimating the mean behavior of the queueing process.

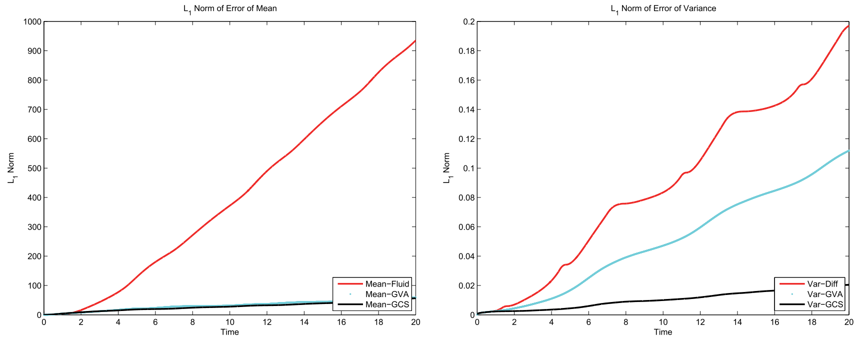

Fig. 6. *Left: Comparisons of L_1 norm of fluid, GVA, and GCS versus simulation means. Right: Comparisons of L_1 norm of diffusion, GVA, and GCS versus simulation variances.*

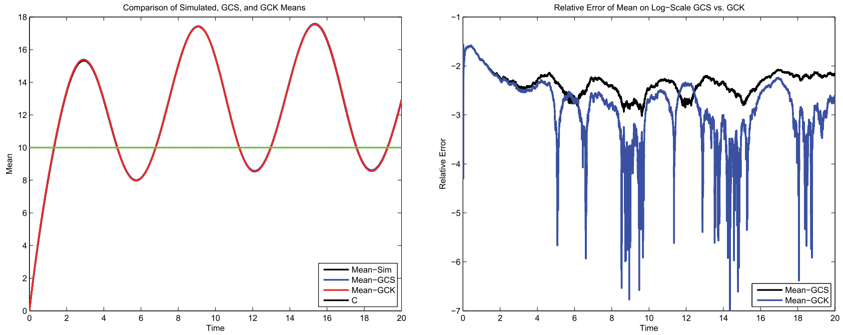

Fig. 7. *Left: Comparison of GCS and GCK means. Right: Relative error of means using GCS and GCK approximations.*

This slight improvement is easier to see in Figure 7(right), which plots over time the logarithm of the relative errors of both GCS and GCK versus the simulated mean.

In Figure 8(left), we also compare the simulated variance with its estimates using the GCS and GCK methods. Since the GCS did a decent job of approximating the variance, we have to resort to looking at the relative error in Figure 8(right). By adding the kurtosis, we can see that the GCK method is approximating the variance behavior better than the GCS. These results indicate that the kurtosis provides valuable information about the variance of our stochastic process and, like the skewness, should be analyzed more often in the literature.

In Figure 9, we compute the L_1 norm of the error over time of the GCS and GCK methods. It is easy to see that the GCK method does a better job than GCS of reducing the error from approximating the real system over longer periods of time. This is because we are adding more stochastic behavior to the model via the kurtosis. This also shows that our method is more stable over time.

In Figure 10, we compare the skewness generated by our simulation, GCS, and GCK. It is clear from the left and right plots that the GCK method is doing a better job of approximating the dynamics of the skewness. This is important because we use the skewness in the GCS method, so if we have better estimates for the skewness, we can get better corrections for the mean and variance as well, which is what we saw in the previous plots. In Figure 10(right), we see that GCK is doing a good job of approximating the kurtosis as well. It does not do a good job in the initial transient phase; however, it does well as time progresses.

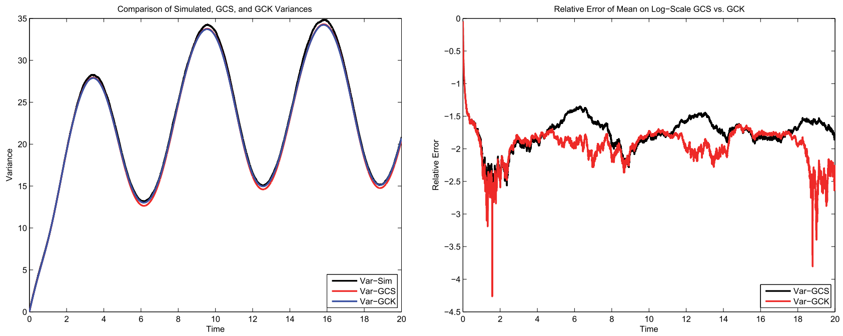

FIG. 8. *Left: Comparison of GCS and GCK variances. Right: Relative error of variances using GCS and GCK approximations.*

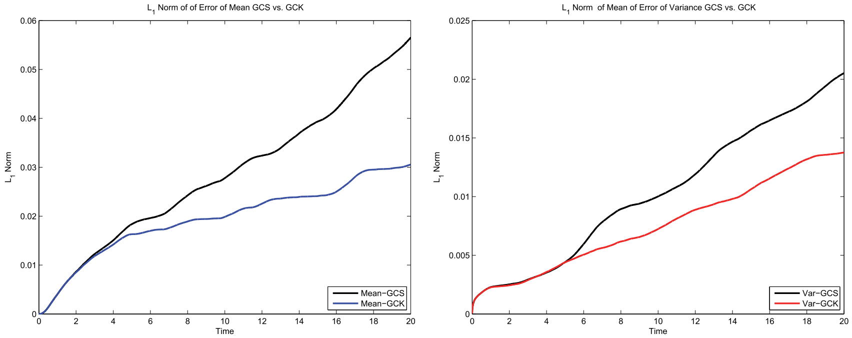

FIG. 9. *Left: Comparisons of L_1 norm GCS and GCK versus simulation means. Right: Comparisons of L_1 norm of GCS and GCK versus simulation variances.*

7. Performance measure approximations: Probability of delay.

7.0.1. Probability of delay for GVA, GCS, and GCK methods. By using the GVA, GCS, and GCK approximations for our queueing process we can also derive an approximate formula for the probability of delay or the probability that a customer who enters the queue at time t will have to wait for service. For GCK, the probability of delay is

$$\begin{aligned}\mathbb{P}(Q \geq c) &= E_{Kur}[\{X \geq \chi\}] \\ &= E_{GVA}[\{X \geq \chi\}] + E_{GCS}[\{X \geq \chi\}] + E_{GCK}[\{X \geq \chi\}] \\ &= \overline{\Phi}(\chi) + (\chi^2 - 1) \cdot \phi(\chi) \cdot \frac{\kappa_3}{6 \cdot \sqrt{v^3}} + (\chi^3 - 3 \cdot \chi) \cdot \phi(\chi) \cdot \frac{\kappa_4}{24 \cdot v^2}.\end{aligned}$$

Like the density, the GCK approximation for the probability of delay is a perturbation of the probability of delay of the GCS, which includes the kurtosis. Thus, if the kurtosis is set to be zero, we get back the GCS probability of delay and, if we also set the skewness to zero, we get the GVA probability of delay. In fact, the perturbations of the skewness and kurtosis terms correspond to the second and third order Edgeworth expansion terms, respectively.

Moreover, in Figure 11, we also simulated the probability of delay for the queueing system and compared it with the different approximations via our Edgeworth expansion. In Figure 11(right) via our L_1 norm comparison, we see that the GCK method outperforms the GCS and GVA methods when approximating this important probabilistic metric.

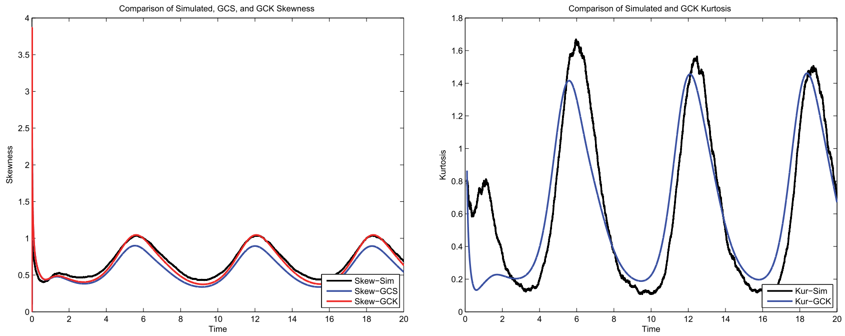

FIG. 10. *Left: Simulated skewness, GCS, and GCK approximations. Right: Simulated kurtosis and GCK approximation.*

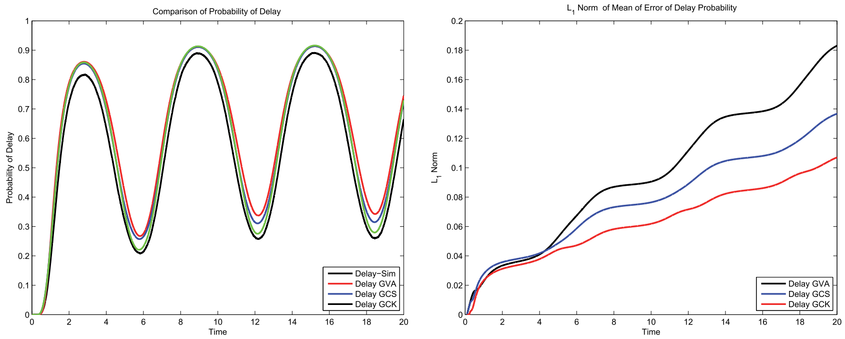

FIG. 11. *Left: Comparisons of the simulated probability of delay against GVA, GCS, and GCK. Right: Comparisons of the L_1 norm of the simulated probability of delay against GVA, GCS, and GCK.*

7.1. Special case $\boldsymbol{\mu = \beta}$ and constant arrival rate $\boldsymbol{\lambda}$. In the special case that $\mu = \beta$, our queueing process dynamics reduces to that of an infinite server queue. This is because we have the following relationship when $\mu = \beta$:

$$\mu \cdot E[Q \wedge c] + \beta \cdot E[(Q-c)^+] = \mu \cdot E[Q] = \beta \cdot E[Q]. \tag{7.1}$$

We know that the infinite server queue when initialized with zero customers has a Poisson distribution with rate $\frac{\lambda}{\mu}$ for fixed t. One property of the Poisson distribution is that all of the cumulant moments are equivalent to its mean. This implies that the mean is equivalent to the variance. Using this property one can show that the Berry–Esseen bound and central limit theorem imply that

$$\mathbb{P}\left(Q \leq c\right) = \Phi(\chi) + \mathcal{O}(\lambda^{-1/2}) \quad \text{as} \quad \lambda \ \rightarrow \ \infty. \tag{7.2}$$

In fact using the Edgeworth expansion for the Poisson distribution yields the following estimates for the probability of delay:

$$\mathbb{P}\left(Q \leq c\right) = \Phi(\chi) - \frac{\phi(\chi) \cdot (\chi^2 - 1)}{6 \cdot \sqrt{\lambda}} + \mathcal{O}\left(\frac{1}{\lambda}\right) \quad \text{as} \quad \lambda \ \rightarrow \ \infty. \tag{7.3}$$

In the constant rate case, these bounds for the probability of delay are proved rigorously; see [5] for an example and for a list of more references. Although we do not prove these bounds in the time varying case, these bounds give us hope that the Gram Charlier expansions we use in this paper are useful for approximating the moment dynamics and the probability of delay dynamics for time varying queues as well.

Table 1
Dynamic staffing parameters.

Parameter	Value
$\lambda(t)$	$10 + 5 \cdot \sin t$
μ	1
$c(t)$	$\lceil \lambda(t) \cdot \frac{3}{2} \rceil$
β	.25

Table 2
Two server parameters.

Parameter	Value
$\lambda(t)$	$2 + 1 \cdot \sin t$
μ	1
$c(t)$	2
β	.25

8. Additional numerical examples.

In this section, we give additional numerical examples of our methods with skewness and kurtosis corrections. These additional examples give support that our new methods work in a variety of parameter settings that are important in practice. Software to implement some of these methods is available on the author's website.

8.1. Dynamic staffing example. In Figure 12 we give an example of a queueing system, with the parameters given in Table 1, where the number of servers changes dynamically through time. On the top left of Figure 12 we see that the GCS and GCK are doing a better job of approximating the mean behavior of the queueing process than the fluid limit or the GVA. This is also confirmed in the middle left of Figure 12, where we see that GCS and GCK have lower log relative errors than GVA and the fluid limit. In the top right of Figure 12, we see a similar picture for the variance. Once again on the top right and the middle right of Figure 12, we see that the GCS and GCK are estimating the variance better than GVA or the diffusion limit. On the bottom left of Figure 12 we see that the GCS and GCK are estimating the skewness quite well. Lastly, on the bottom right of Figure 12, we see that GCS and GCK are doing well at mimicking the probability of delay. This example provides evidence that our methods work well when the number of servers dynamically changes throughout time and is not just a constant.

8.2. Two server example. In the following numerical example, with the parameters given in Table 2, the purpose is to illustrate that the GCS and GCK methods can give reasonable approximation for the smallest multiserver queue with abandonment, i.e., when $c = 2$. As in Figure 12 we see in the top left of Figure 13 that the GCS and GCK are doing a better job of estimating the mean behavior of the queueing process than the fluid limit. Moreover, in the top right of Figure 13 we also see the same improvements for the variance. Unlike in systems with more servers, the improvement of the skewness and kurtosis is easily seen. This is also further supported in the bottom right of Figure 13, where we see that the log relative errors for GCS and GCK are much lower than the DMA, GVA, or diffusion limit. Lastly, we see in the bottom left of Figure 13 that the GCS and GCK methods provide accurate estimates of the skewness of the queueing system; however, it seems apparent that the GCK method is providing a better estimate for the skewness in this example. It is also im-

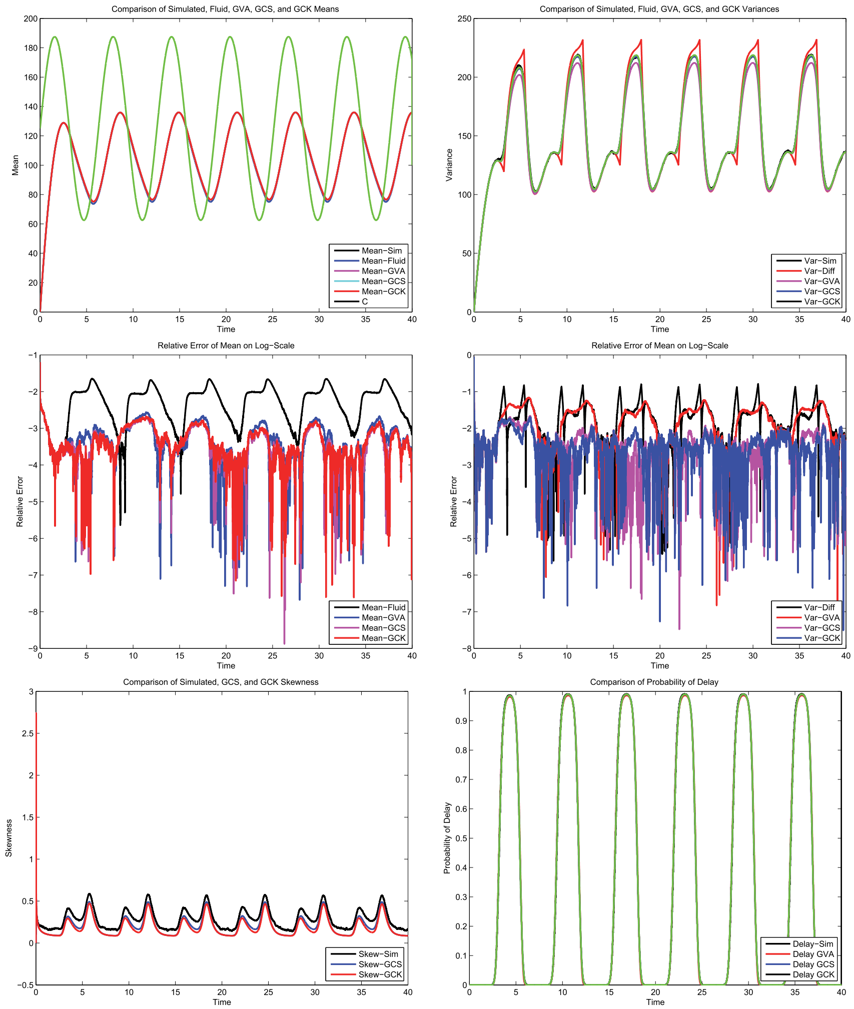

FIG. 12. *Sinusoidal arrival rate and sinusoidal staffing schedule.*

portant to note that the system has a larger skewness value than the other numerical examples, which should be expected since we are as far away from the many server Gaussian heavy traffic limits as we could be. Thus, this example gives evidence that the GCS and GCK can accurately estimate the dynamics of systems where there are a small number of servers and when the queueing process is clearly not in the many server heavy traffic regime.

8.3. High arrival rate and large number of servers example. In Figure 14, we give an example of the dynamics of a queueing system, with the parameters given in Table 3, with a high arrival rate and a large number of servers. We see in the top of Figure 14 that the mean and variance are approximated very well regardless of the method used. One reason is that we are very close to operating in the many server heavy traffic regime, and distribution is becoming more *Gaussian like.* In the middle of Figure 14 we see that the log relative error indicates that the GCS and GCK methods are doing the best at approximating the mean and variance dynamics. In the

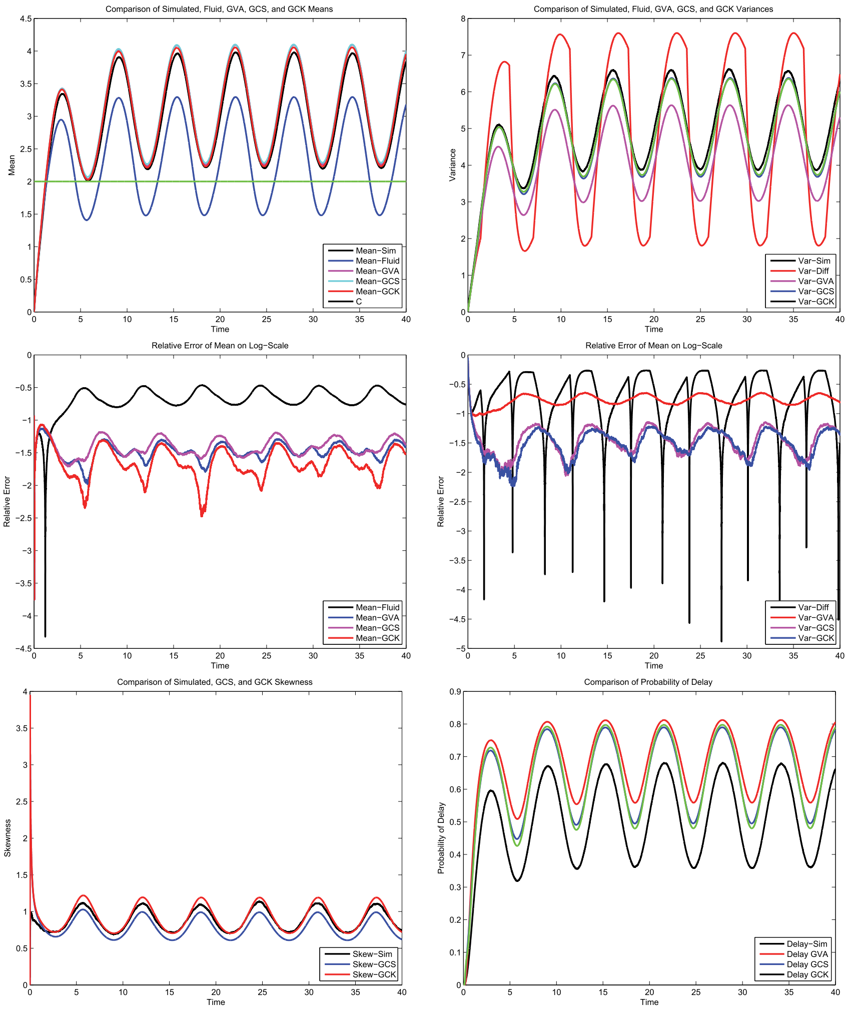

FIG. 13. *Sinusoidal arrival rate and constant staffing schedule* $c = 2$.

bottom left of Figure 14, we see that GCS is doing a better job of approximating the skewness of the queueing distribution, which might explain why the Gram Charlier expansion is not uniformly better as we add more terms. Lastly, on the bottom right of Figure 14, we see that GVA, GCS, and GCK are all doing a good job of estimating the probability of delay very well. With a high arrival rate and large number of servers, it is not necessary to use the skewness and kurtosis, as there is not much room for correcting the estimates of the mean and variance dynamics. One small thing to notice is that the skewness has a local maximum when the queueing process is critically loaded, i.e., $(q = c)$, where we expect the queueing system not to be Gaussian.

9. Conclusions and future work. We have shown in this paper that the skewness and kurtosis and higher moments can add valuable information to our understanding of queueing system dynamics. As the skewness and kurtosis of our queueing system are nonzero, it is apparent that it is not sufficient to study just mean and

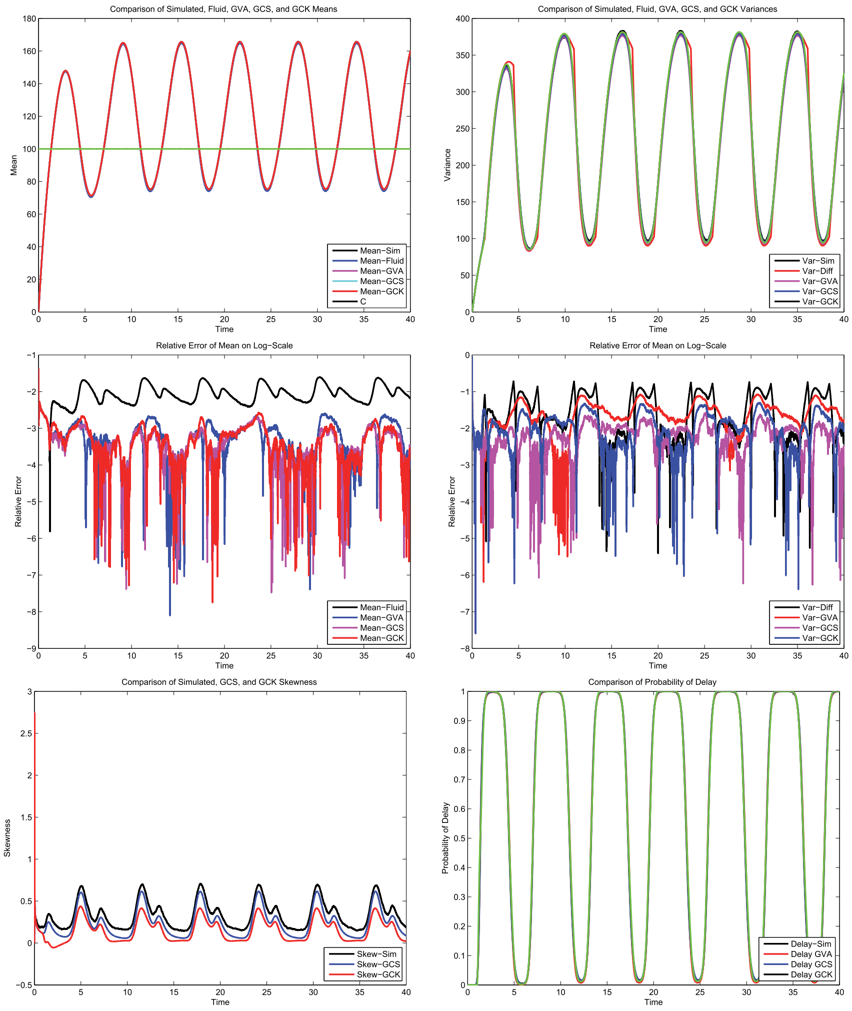

FIG. 14. *Sinusoidal arrival rate and constant staffing schedule $c = 100$.*

TABLE 3
High arrival rate parameters.

Parameter	Value
$\lambda(t)$	$100 + 50 \cdot \sin t$
μ	1
$c(t)$	100
β	.25

variance behavior of time dependent stochastic systems. By using the Gram Charlier expansion, we were able to add the skewness and kurtosis effects to the behavior of our queueing model and take the first step toward moving beyond mean and variance analysis, which is the main focus of the current literature. We can mimic very accurately the real behavior of the mean, variance, skewness, and kurtosis of our queueing system by using four simple differential equations. As a result, to understand important quantities such as the probability of delay for our queueing system, it suffices to integrate four differential equations instead of spending much time simulating the system. Moreover, we can show that the refinements of our approximations are small in regions where the fluid and diffusion limits are warranted, and are large when these approximations are not warranted. We see that the skewness adds significant information for the non-Gaussian behavior of the queueing process. The information gained by including the kurtosis is also useful, but not as influential on accurate dynamics as the skewness information. Thus, this analysis yields insight and information for managers who want to optimally staff their service systems.

We should also mention that our method easily extends to the case of a *fast abandonment* queueing model, which was introduced by Hampshire, Jennings, and Massey [3] as an approximation of a multiserver loss queueing process. This method might even be more enlightening and useful when the queue is overloaded and the distribution is skewed to the right where customers are blocked from accessing the system.

Although this paper considers only a one-dimensional Markovian service network example, there exists multivariate forms of the Gram Charlier expansion such as the following two-dimensional version:

$$\begin{aligned}\phi_{Skew}(x_1,x_2,\rho) &= \phi(x_1,x_2,\rho) + \phi(x_1)\cdot\phi(x_2)\cdot\left(\frac{\gamma_{3,0}}{6}\cdot h_3(x_1) + \frac{\gamma_{0,3}}{6}\cdot h_3(x_2)\right)\\ &\quad + \phi(x_1)\cdot\phi(x_2)\cdot\left(\frac{\gamma_{2,1}}{2}\cdot h_2(x_1)\cdot h_1(x_2) + \frac{\gamma_{1,2}}{2}\cdot h_1(x_1)\cdot h_2(x_2)\right).\end{aligned}$$

One would hope that using a multivariate Gram Charlier expansion might yield insight on approximating the skewness and kurtosis behavior in multidimensional Markovian stochastic networks. We expect that this multivariate approach yields better approximations for larger networks as well. We are currently pursuing this extension for other types of Markovian service networks. Our approach is not uniformly better as we add more terms. It would be interesting to find an asymptotic method that uniformly yields better results as we add more terms to the approximation. Lastly, we have expanded our queueing process around a Gaussian density, which is motivated from the diffusion limit being a Gaussian diffusion. However, in the infinite server case, we know that the queueing process has a Poisson distribution. It would be interesting to expand around a non-Gaussian distribution such as the Poisson distribution with Poisson–Charlier polynomials and analyze this problem as well.

Appendix A.

A.1. Hermite polynomials. In this section, we give a brief introduction to Hermite polynomials and their main properties. The building blocks of our approximation method or the expectations and covariance terms found in Theorem 5.1 can be computed using the Hermite polynomial calculus developed in [11]. The Hermite polynomials (probabilistic) are defined as

$$h_n(x) \equiv e^{x^2/2}\left(-\frac{d}{dx}\right)^n e^{-x^2/2}.$$

We give the first eight probabilistic Hermite polynomials for future reference, as they will be used throughout the rest of this section for the proofs and derivations of the unknown expectation and covariances terms:

$$h_0(x) = 1, \quad h_1(x) = x, \quad h_2(x) = x^2 - 1, \quad h_3(x) = x^3 - 3x,$$
$$h_4(x) = x^4 - 6x^2 + 3, \quad h_5(x) = x^5 - 10x^3 + 15x,$$
$$h_6(x) = x^6 - 15x^4 + 45x^2 - 15, \quad h_7(x) = x^7 - 21x^5 + 105x^3 - 105x.$$

Remark A.1. It is important to note that all of the Hermite polynomials evaluated at X, except for the first, have expectation zero. This will simplify the analysis of future computations to come since many expectations with respect to some of these polynomials will vanish.

Proposition A.2. *If X is a standard Gaussian random variable, then*

$$E[f(X) \cdot h_n(X)] = E[f^{(n)}(X)],$$

where f is any generalized function.

This follows easily from integration by parts since the Gaussian density is a smooth density. From this result follows the orthogonality property of Hermite polynomials.

Proposition A.3. *Hermite polynomials satisfy the following orthogonality properties:*

$$E[h_i(X)] = \begin{cases} 1 & \text{if } i = 0, \\ 0 & \text{if } i \neq 0. \end{cases}$$

$$E[h_i(X) \cdot h_j(X)] = \begin{cases} j! & \text{if } i \ = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Proposition A.4. *$(\frac{1}{\sqrt{n!}} H_n)$ is an orthonormal basis of $L^2(\mathbb{R}, \nu)$, where ν is the Gaussian measure.*

Proof. It suffices to show that the Hermite polynomials are orthogonal and that the set of Hermite polynomials $(H_n)_{n \geq 0}$ is complete in $L^2(\mathbb{R}, \nu)$; i.e., the set of linear combinations of Hermite polynomials is dense in $L^2(\mathbb{R}, \nu)$. The first part was proven in the previous theorem. The second part can be found in [12]. ▯

Proposition A.5. *Any L^2 function can be written as an infinite sum of Hermite polynomials of X, i.e.,*

$$f(X) \overset{L^2}{=} \sum_{n=0}^{\infty} \frac{1}{n!} E[f^{(n)}(X)] \cdot h_n(X)$$

and

$$E[f(X) \cdot g(X)] = \sum_{n=0}^{\infty} \frac{1}{n!} \cdot E[f^{(n)}(X)] \cdot E[g^{(n)}(X)]$$

and

$$\text{Cov}[f(X), g(X)] = \sum_{n=1}^{\infty} \frac{1}{n!} \cdot E[f^{(n)}(X)] \cdot E[g^{(n)}(X)].$$

A.2. Calculations of unknown expectations and covariance terms. In this section, we derive explicit formulas for the expectations and covariances needed to construct our dynamical system approximation for our queueing process.

We have the following expression of polynomials in terms of the Hermite polynomials. These lemmas will be extremely useful for calculating the Gram Charlier expectations.

Lemma A.6.

$$\begin{aligned} X &= h_1(X), \\ X^2 &= h_2(X) + h_0(X), \\ X^3 &= h_3(X) + 3 \cdot h_1(X), \\ X^4 &= h_4(X) + 6 \cdot h_2(X) + 3 \cdot h_0(X), \\ X^5 &= h_5(X) + 10 \cdot h_3(X) + 15 \cdot h_1(X), \\ X^6 &= h_6(X) + 15 \cdot h_4(X) + 45 \cdot h_2(X) + 15 \cdot h_0(X), \\ X^7 &= h_7(X) + 21 \cdot h_5(X) + 105 \cdot h_3(X) + 105 \cdot h_1(X). \end{aligned}$$

Lemma A.7.

$$\begin{aligned} X \cdot h_3(X) &= h_4(X) + 3 \cdot h_2(X), \\ X^2 \cdot h_3(X) &= h_5(X) + 7 \cdot h_3(X) + 6 \cdot h_1(X), \\ X^3 \cdot h_3(X) &= h_6(X) + 12 \cdot h_4(X) + 27 \cdot h_2(X) + 6 \cdot h_0(X), \\ X \cdot h_4(X) &= h_5(X) + 4 \cdot h_3(X), \\ X^2 \cdot h_4(X) &= h_6(X) + 9 \cdot h_4(X) + 12 \cdot h_2(X), \\ X^3 \cdot h_4(X) &= h_7(X) + 15 \cdot h_5(X) + 48 \cdot h_3(X) + 24 \cdot h_1(X). \end{aligned}$$

We give expressions for the variance, third cumulant moment, and fourth cumulant moment in terms of the raw moments of a random variable.

Lemma A.8.

$$\begin{aligned} \mathrm{Var}[Q] &= E[(Q - E[Q])^2] \\ &= E[Q^2] - E[Q]^2, \\ C^{[3]}[Q] &= E[(Q - E[Q])^3] \\ &= E[Q^3] - 3 \cdot E[Q^2] \cdot E[Q] + 2 \cdot E[Q]^3, \\ C^{[4]}[Q] &= E[(Q - E[Q])^4] - 3 \cdot \mathrm{Var}[Q]^2 \\ &= E[Q^4] - 4 \cdot E[Q^3] \cdot E[Q] - 3 \cdot E[Q^2]^2 + 12 \cdot E[Q^2] \cdot E[Q]^2 - 6 \cdot E[Q]^4. \end{aligned}$$

Now by taking the time derivative, we have the following expressions for the cumulant moments in terms of the moments and their time derivatives.

Lemma A.9.

$$\begin{aligned} \dot{\mathrm{Var}}[Q] &= \dot{E}[Q^2] - 2 \cdot \dot{E}[Q] \cdot E[Q], \\ \dot{C}^{[3]}[Q] &= \dot{E}[Q^3] - 3 \cdot \dot{E}[Q^2] \cdot E[Q] - 3 \cdot E[Q^2] \cdot \dot{E}[Q] + 6 \cdot \dot{E}[Q] \cdot E[Q]^2, \\ \dot{C}^{[4]}[Q] &= \dot{E}[Q^4] - 4 \cdot \dot{E}[Q^3] \cdot E[Q] - 4 \cdot E[Q^3] \cdot \dot{E}[Q] - 6 \cdot \dot{E}[Q^2] \cdot E[Q^2], \\ &\quad + 12 \cdot \dot{E}[Q^2] \cdot E[Q]^2 + 24 \cdot E[Q^2] \cdot E[Q] \cdot \dot{E}[Q] - 24 \cdot \dot{E}[Q] \cdot E[Q]^3. \end{aligned}$$

Now we compute all of the unknown expectations and covariance terms of Theorems 5.1 and 6.1. It suffices to prove the results for Theorem 6.1 as it is a generalization of Theorem 5.1. Moreover, it suffices to prove only the terms $(Q-c)^+$ since we have the relationship

$$Q \wedge c = Q - (Q-c)^+. \tag{1.1}$$

A.2.1. Computation of first order terms. Now we compute the first order terms, using the Hermite polynomial machinery explained above, which are important for deriving the approximations for the mean behavior. We compute only the terms for the GCK method since GCS can be obtained by setting κ_4 to zero. Moreover, GVA terms can be obtained by setting both κ_3 and κ_4 to zero. Thus, it suffices to compute only the expectation and covariance terms only for the case of the GCK:

$$\begin{aligned}
E\left[(Q-c)^+\right] &= \sqrt{v}\cdot E\left[(X-\chi)^+\right] \\
&= \sqrt{v}\cdot E_{GVA}\left[(X-\chi)^+\right] + \sqrt{v}\cdot E_{GCS}\left[(X-\chi)^+\right] + \sqrt{v}\cdot E_{GCK}\left[(X-\chi)^+\right] \\
&= \sqrt{v}\cdot E\left[(X-\chi)^+\right] + \frac{\kappa_3}{6\cdot v}\cdot E\left[h_3(X)\cdot(X-\chi)^+\right] \\
&\quad + \frac{\kappa_4}{24\cdot\sqrt{v^3}}\cdot E\left[h_4(X)\cdot(X-\chi)^+\right] \\
&= \sqrt{v}\cdot\phi(\chi) - \chi\cdot\sqrt{v}\cdot\overline{\Phi}(\chi) + \frac{\chi\cdot\phi(\chi)\cdot\kappa_3}{6\cdot v} + \frac{(\chi^2-1)\cdot\phi(\chi)\cdot\kappa_4}{6\cdot\sqrt{v^3}},
\end{aligned}$$

$$\begin{aligned}
E\left[Q\wedge c\right] &= E[Q] - E\left[(Q-c)^+\right] \\
&= q - \sqrt{v}\cdot\phi(\chi) + \chi\cdot\sqrt{v}\cdot\overline{\Phi}(\chi) - \frac{\chi\cdot\phi(\chi)\cdot\kappa_3}{6\cdot v} - \frac{(\chi^2-1)\cdot\phi(\chi)\cdot\kappa_4}{6\cdot\sqrt{v^3}}.
\end{aligned}$$

A.2.2. Computation of second order terms. Now we compute the second order terms, which are useful for deriving the variance approximations:

$$\begin{aligned}
\text{Cov}\left[Q,(Q-c)^+\right] &= \text{Cov}\left[q+\sqrt{v}\cdot X,(Q-c)^+\right] \\
&= \text{Cov}\left[\sqrt{v}\cdot X,(Q-c)^+\right] \\
&= v\cdot\text{Cov}\left[X,(X-\chi)^+\right] \\
&= v\cdot E\left[X\cdot(X-\chi)^+\right] \\
&= v\cdot E_{GVA}\left[X\cdot(X-\chi)^+\right] + v\cdot E_{GCS}\left[X\cdot(X-\chi)^+\right] + v\cdot E_{GCK}\left[X\cdot(X-\chi)^+\right] \\
&= v\cdot E\left[X\cdot(X-\chi)^+\right] + \frac{\kappa_3}{6\cdot\sqrt{v}}\cdot E\left[h_3(X)\cdot X\cdot(X-\chi)^+\right] \\
&\quad + \frac{\kappa_4}{24\cdot v}\cdot E\left[h_4(X)\cdot X\cdot(X-\chi)^+\right] \\
&= v\cdot E\left[X\cdot(X-\chi)^+\right] + \frac{\kappa_3}{6\cdot\sqrt{v}}\cdot E\left[(h_4(X)+3\cdot h_2(X))\cdot(X-\chi)^+\right] \\
&\quad + \frac{\kappa_4}{24\cdot v}\cdot E\left[(h_5(X)+4\cdot h_3(X))\cdot(X-\chi)^+\right] \\
&= v\cdot\overline{\Phi}(\chi) + \frac{\kappa_3}{6\cdot\sqrt{v}}\cdot(\chi^2+2)\cdot\phi(\chi) + \frac{\kappa_4}{24\cdot v}\cdot(\chi^3+\chi)\cdot\phi(\chi),
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}\left[Q,(Q\wedge c)\right] &= \mathrm{Cov}\left[Q,Q\right]-\mathrm{Cov}\left[Q,(Q-c)^{+}\right]\\
&= v-\mathrm{Cov}\left[Q,(Q-c)^{+}\right]\\
&= v\cdot\Phi(\chi)-\frac{\kappa_3}{6\cdot\sqrt{v}}\cdot(\chi^2+2)\cdot\phi(\chi)-\frac{\kappa_4}{24\cdot v}\cdot(\chi^3+\chi)\cdot\phi(\chi).
\end{aligned}$$

A.2.3. Computation of third order terms.

$$\begin{aligned}
\mathrm{Cov}\left[\overline{Q}^{2},(Q-c)^{+}\right] &= \mathrm{Cov}\left[v\cdot X^2,(Q-c)^{+}\right]\\
&= \sqrt{v^3}\cdot\mathrm{Cov}\left[X^2,(X-\chi)^{+}\right]\\
&= \sqrt{v^3}\cdot\mathrm{Cov}_{GVA}\left[X^2-1,(X-\chi)^{+}\right]+\sqrt{v^3}\cdot\mathrm{Cov}_{GCS}\left[X^2,(X-\chi)^{+}\right]\\
&\quad+\sqrt{v^3}\cdot\mathrm{Cov}_{GCK}\left[X^2,(X-\chi)^{+}\right]\\
&= \sqrt{v^3}\cdot E\left[\delta_{\chi}(X)\right]+\frac{\kappa_3}{6\cdot\sqrt{v^3}}\cdot\sqrt{v^3}\cdot\mathrm{Cov}\left[X^2\cdot h_3(X),(X-\chi)^{+}\right]\\
&\quad+\frac{\kappa_4}{24\cdot v^2}\cdot\sqrt{v^3}\cdot\mathrm{Cov}\left[X^2\cdot h_4(X),(X-\chi)^{+}\right]\\
&= \sqrt{v^3}\cdot\phi(\chi)+\sqrt{v^3}\cdot\frac{\kappa_3}{6\cdot\sqrt{v^3}}\cdot\left[(\chi^3+4\cdot\chi)\cdot\phi(\chi)+6\cdot\overline{\Phi}(\chi)\right]\\
&\quad+\frac{\kappa_4}{24\cdot v^2}\cdot\sqrt{v^3}\cdot(\chi^4+3\cdot\chi^2+6)\cdot\phi(\chi)\\
&= \sqrt{v^3}\cdot\phi(\chi)+\sqrt{v^3}\cdot\frac{\kappa_3}{6\cdot\sqrt{v^3}}\cdot\left[(\chi^3+4\cdot\chi)\cdot\phi(\chi)+6\cdot\overline{\Phi}(\chi)\right]\\
&\quad+\frac{\kappa_4}{24\cdot v^2}\cdot\sqrt{v^3}\cdot(\chi^4+3\cdot\chi^2+6)\cdot\phi(\chi)\\
&= \sqrt{v^3}\cdot\phi(\chi)+\frac{\kappa_3}{6}\cdot\left[(\chi^3+4\cdot\chi)\cdot\phi(\chi)+6\cdot\overline{\Phi}(\chi)\right]\\
&\quad+\frac{\kappa_4}{24\cdot\sqrt{v}}\cdot(\chi^4+3\cdot\chi^2+6)\cdot\phi(\chi),
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}\left[\overline{Q}^{2},(Q\wedge c)\right] &= \mathrm{Cov}\left[\overline{Q}^{2},Q\right]-\mathrm{Cov}\left[\overline{Q}^{2},(Q-c)^{+}\right]\\
&= \kappa_3-\mathrm{Cov}\left[\overline{Q}^{2},(Q-c)^{+}\right]\\
&= \kappa_3-\sqrt{v^3}\cdot\phi(\chi)-\frac{\kappa_3}{6}\cdot\left[(\chi^3+4\cdot\chi)\cdot\phi(\chi)+6\cdot\overline{\Phi}(\chi)\right]\\
&\quad-\frac{\kappa_4}{24\cdot\sqrt{v}}\cdot(\chi^4+3\cdot\chi^2+6)\cdot\phi(\chi).
\end{aligned}$$

A.2.4. Computation of fourth order terms.

$$
\begin{aligned}
\text{Cov}\left[\overline{Q}^3, (Q-c)^+\right] &= \text{Cov}\left[\sqrt{v^3}\cdot X^3, (Q-c)^+\right] \\
&= v^2 \cdot \text{Cov}\left[X^3, (X-\chi)^+\right] \\
&= v^2 \cdot \text{Cov}_{GVA}\left[X^3, (X-\chi)^+\right] + v^2 \cdot \text{Cov}_{GCS}\left[X^3, (X-\chi)^+\right] \\
&\quad + v^2 \cdot \text{Cov}_{GCK}\left[X^3, (X-\chi)^+\right] \\
&= v^2 \cdot \text{Cov}\left[X^3 - 3X, (X-\chi)^+\right] + v^2 \cdot \text{Cov}\left[3X, (X-\chi)^+\right] \\
&\quad + \frac{\kappa_3}{6\cdot\sqrt{v^3}} \cdot \sqrt{v^3} \cdot \text{Cov}\left[X^3 \cdot h_3(X), (X-\chi)^+\right] \\
&\quad + \frac{\kappa_4}{24\cdot v^2} \cdot \sqrt{v^3} \cdot \text{Cov}\left[X^3 \cdot h_4(X), (X-\chi)^+\right] \\
&= v^2 \cdot \left((\chi^2+1)\cdot\phi(\chi)\right) + 3\cdot v^2 \cdot \overline{\Phi}(\chi) \\
&\quad + \frac{\kappa_3}{6\cdot\sqrt{v^3}} \cdot v^2 \cdot \text{Cov}\left[h_6(X) + 12\cdot h_4(X) + 27\cdot h_2(X) + 6, (X-\chi)^+\right] \\
&\quad + \frac{\kappa_4}{24\cdot v^2} \cdot \sqrt{v^3} \cdot \text{Cov}\left[h_7(X) + 15\cdot h_5(X) + 48\cdot h_3(X) + 24\cdot h_1(X), (X-\chi)^+\right] \\
&= v^2 \cdot \left((\chi^2+1)\cdot\phi(\chi)\right) + 3\cdot v^2 \cdot \overline{\Phi}(\chi) \\
&\quad + \frac{\kappa_3\cdot\sqrt{v}}{6} \cdot \left((h_4(\chi) + 12\cdot h_2(\chi) + 27)\cdot\phi(\chi)\right) \\
&\quad + \frac{\kappa_4}{24\cdot v^2} \cdot \sqrt{v^3} \cdot \left((h_5(\chi) + 15\cdot h_3(\chi) + 48\cdot h_1(\chi))\cdot\phi(\chi) + 24\cdot\overline{\Phi}(\chi)\right),
\end{aligned}
$$

$$
\begin{aligned}
\text{Cov}\left[\overline{Q}^3, (Q\wedge c)\right] &= \text{Cov}\left[\overline{Q}^3, Q\right] - \text{Cov}\left[\overline{Q}^3, (Q-c)^+\right] \\
&= v^2 \cdot \text{Cov}\left[X^3, X\right] + \frac{\kappa_3\cdot\sqrt{v}}{6} \cdot \text{Cov}\left[h_3(X)\cdot X^3, X\right] \\
&\quad + v^2 \cdot \frac{\kappa_4 \cdot}{24\cdot v^2} \cdot \text{Cov}\left[h_4(X)\cdot X^3, X\right] - \text{Cov}\left[\overline{Q}^3, (Q-c)^+\right] \\
&= 3\cdot v^2 + \kappa_4 - \text{Cov}\left[\overline{Q}^3, (Q-c)^+\right].
\end{aligned}
$$

Acknowledgments. The author would like to thank the two referees and the area editor as their comments were very useful in improving this paper. The author would also like to thank the hospitality of Columbia University's IEOR department.

REFERENCES

[1] C. Corrado and T. Su, *Skewness and Kurtosis in S&P* 500 *index returns implied by option prices*, J. Financial Res., 19 (1996), pp. 175–192.

[2] S. Halfin and W. Whitt, *Heavy-traffic limit theorems for queues with many exponential servers*, Oper. Res., 29 (1981), pp. 567–588.

[3] R. C Hampshire, O. B Jennings, and W. A. Massey, *A time varying call center design with Lagrangian mechanics*, Probab. Engrg. Inform. Sci., 23 (2009), pp. 231–259.

[4] R. C. Hampshire and W. A. Massey, *Dynamic optimization with applications to dynamic rate queueing*, Tutorials Oper. Res., 7 (2010), pp. 208–247.

[5] A. Janssen, J. van Leeuwaarden, and B. Zwart, *Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula*, Adv. in Appl. Probab., 40 (2008), pp. 122–143.

[6] A. Janssen, J. van Leeuwaarden, and B. Zwart, *Refining square-root safety staffing by expanding Erlang C*, Oper. Res., 59 (2010), pp. 1512–1522.

[7] Y. M. Ko and N. Gautam, *Critically loaded time-varying multiserver queues: Computational challenges and approximations*, INFORMS J. Comput., 25 (2013), pp. 285–301.

[8] A. Mandelbaum, W. A. Massey, and M. Reiman, *Strong approximations for Markovian service networks*, Queueing Syst., 30 (1998), pp. 149–201.

[9] A. Mandelbaum, W. A. Massey, M. Reiman, B. Rider, and S. Stolyar, *Queue lengths and waiting times for multiserver queues with abandonment and retrials*, Telecommunication Systems, 21 (2002), pp. 149–171.

[10] W. A. Massey and J. Pender, *Poster: Skewness variance approximation for dynamic rate multi-server queues with abandonment*, ACM SIGMETRICS Performance Evaluation Review, 39 (2011), p. 74.

[11] W. A. Massey and J. Pender, *Gaussian skewness approximation for dynamic rate multi-server queues with abandonment*, Queueing Syst., 75 (2013), pp. 243–277.

[12] D. Nualart, *The Malliavin Calculus and Related Topics*, Springer, Berlin, 1995 .

[13] M. H. Rothkopf and S. S. Oren, *A closure approximation for the nonstationary M/M/s queue*, Management Sci., 25 (1979/80), pp. 522–534.

[14] C. M. Stein, *Approximate Computation of Expectations*, IMS Lecture Notes Monogr. Ser. 7, Institute of Mathematical Statistics, Hayward, CA, 1986.

[15] M. R. Taaffe and K. L. Ong, *Approximating nonstationary Ph(t)/M(t)/s/c queueing systems*, Ann. Oper. Res., 8 (1987), pp. 103–116.

[16] B. Zhang, A. Janssen, J. van Leeuwaarden, and B. Zwart, *Refined square-root staffing for call centers with impatient customers*, Oper. Res., 60 (2012), pp. 461–474.