Annals of Operations Research (2022) 310:595–620
https://doi.org/10.1007/s10479-020-03859-3

S.I. : QUEUESTOCHMOD2019

A queueing-inventory system with random order size policy and server vacations

Yuying Zhang[1] · **Dequan Yue**[2] · **Wuyi Yue**[3]

Accepted: 3 November 2020 / Published online: 16 November 2020

Abstract
In this paper, we consider a queueing-inventory system under continuous review with a random order size policy and lost sales. If the inventory is depleted after the service of a customer, a replenishment order is instantaneously triggered. The replenishment order size may be randomized according to a discrete probability distribution. Customers arrive in the system according to a Poisson process and require service from a server. The server takes multiple vacations once the inventory is depleted. The service time, the lead time, and the vacation time are all assumed to be distributed exponentially. We derive the stationary joint distribution of the queue length, the on-hand inventory level, and the status of the server in explicit product form. Furthermore, the conditional distributions of the on-hand inventory level when the server is off due to a vacation or depleted inventory, and when the server is on and working, are derived. Then, we calculate some of the system performance measures. The effect of the server's vacation on the performance measures is investigated analytically. Finally, some numerical results are presented. The simulation study of the model in the context of more general arrival processes and service time distributions is presented.

Keywords Queueing-inventory system · Multiple vacation · Lost sales · Randomized order size · Performance analysis

1 Introduction

A queueing-inventory system (QIS) is a queueing system with attached inventory in which customers arrive one by one and need not only an on-hand item but also some form of time-

✉ Dequan Yue
ydq@ysu.edu.cn

Yuying Zhang
756552686@qq.com

Wuyi Yue
yue@konan-u.ac.jp

1 School of Economics and Management, Yanshan University, Qinhuangdao 066004, China

2 School of Science, Yanshan University, Qinhuangdao 066004, China

3 Faculty of Intelligence and Informatics, Konan University, Kobe 658-8501, Japan

consuming service. For example, items in inventory require time for retrieval, preparation, packing, and loading (see Saffari et al. 2011). Compared to the traditional inventory system, a QIS is more general and realistic. Over the past decades, research on QIS has attracted significant research attention due to its wide-ranging applications in such fields as integrated supply chain management, vehicle maintenance and medical services (see Schwarz et al. 2006; Krishnamoorthy et al. 2016a; Arun 2010).

It seems that the first contribution to QIS research was the work done by Sigman and Simchi-Levi (1992) and Melikov and Molchanov (1992) where the analyses were carried out under the assumptions of an arbitrarily distributed service time in Sigman and Simchi-Levi (1992) and an exponentially distributed service time in Melikov and Molchanov (1992). Sigman and Simchi-Levi (1992) investigated an M/G/1 QIS model, where it was assumed that customers arriving at the system during an out-of-stock period were backlogged. In the literature on inventory systems, these customers are referred to as backorders. Sigman and Simich-Levi proposed a light traffic heuristic approximation procedure to derive performance for their model. Melikov and Molchanov (1992) considered a QIS in a transportation/storage system (TSS), where a user request is lost if the request arrives when the system already contains the maximum number N of user requests. The exact and approximate solution methods were proposed. Subsequently, many research papers on QIS models with backorders were presented. We refer to the survey paper by Krishnamoorthy et al. (2011) for more details on this topic.

Another aspect of QIS research has been on the lost sales model. In this model, it is assumed that customers arriving at the system during an out-of-stock period are lost. Many research papers on QIS with lost sales have been published. A special mention should be paid to the paper by Schwarz et al. (2006) who studied an M/M/1 QIS model under three different inventory management policies including random order size (ROS) policy, (r, Q) inventory policy and (s, S) inventory policy, respectively. The authors derived a product form solution for the stationary joint probability of the queue length and the inventory level by using the probability generating function method. This solution is special because a strong correlation exists between the number of customers joining the system during the lead time and the number of items in the inventory over that period. Krishnamoorthy and Viswanath (2013) subsumed the work in Schwarz et al. (2006) to a (s, S) production inventory model with an M/M/1 service queue where the inventory items were gradually replenished by an internal production process. They obtained the production form solution for the system state distribution in steady state by using a matrix theoretical approach. Baek and Moon (2014) studied a production-inventory system with an M/M/1 service queue and lost sales where the stocks were replenished by both an external order under (r, Q) policy and an international production. They derived the stationary joint distribution of the queue length and the on-hand inventory in product form.

Melikov et al. (2016) considered an M/M/1 QIS model with either a finite or an infinite queue of impatient customers, where ROS policy was considered. The exact and approximate methods to calculate the characteristics of the systems under given lead policies were developed. Melikov et al. (2017) further considered a Markovian QIS model with impatient customers and a variable size of order policy in which the size of the order is dependent on the on-hand inventory level. The exact and approximate methods were developed to calculate the characteristics of the systems under a proposed restocking policy. For other QIS research that includes either (r, Q) inventory policy or (s, S) inventory policy or both, we refer to a spate of research papers including Saffari et al. (2011, 2013), Krenzler and Daduna (2015), Krishnamoorthy et al. (2015, 2016b), Yue et al. (2018), Barron (2019) and several others.

Queueing systems with server vacations have been extensively applied in many fields such as communication systems, manufacturing systems, call centers, and production inventory systems. We refer to Doshi (1986), Takagi (1991), Tian and Zhang (2006) and Ke et al. (2010) for more details on this topic. However, there has been only limited research into QIS that considers server vacations. Viswanath et al. (2008) introduced server vacations into a QIS with a (s, S) inventory policy, where the customers who waited for service may renege after a period of random time. They computed the steady-state probabilities by using the level dependent quasi-birth-and-death (QBD) process theory. Sivakumar (2011) studied an M/M/1 QIS model with multiple server vacations and a (s, S) inventory policy, where the demands that occurred during an out of stock period and/or during a server vacation period entered the orbit of infinite size. They obtained the joint probability distribution of the inventory level and the number of customers in the orbit in the steady-state case. Various performance measures and the long run expected total cost rate were calculated.

Recently, Padmavathi et al. (2016) investigated a finite-source inventory system with postponed demands and a modified vacation policy, where a (s, S) inventory policy was considered. The vacation time and the lead time followed independent PH distributions. The joint distribution of the mode of the server, the server status, the inventory level, and the number of demands in the pool were obtained in the steady-state. Melikov et al. (2017) proposed a model for a servicing system with perishable inventory and a finite queue of impatient claims where a (s, S) inventory policy was considered, and the server could be in one of three states: operational, early and delaying vacations. They developed a method to approximately compute the system's characteristics. Koroliuk et al. (2017, 2018) proposed Markov QIS models with perishable inventory and a (s, S) inventory policy. Koroliuk et al. (2017), it was assumed that the server took vacations if either the inventory level was zero, the queue was empty, or both. Unlike in Koroliuk et al. (2017), it was assumed in Koroliuk et al. (2018) that the server took a vacation only if there were no customers in the system at the moment its operation completed, and the server returned to operating mode only when the number of customers in the system exceeded some thresholds. In these studies, they developed an exact and an approximate method to find the system's characteristics. Jeganathana and Abdul (2020) considered a two-server Markovian inventory system with modified and delayed working vacations, where a (s, Q) inventory policy was considered. The various measures of system performance in the steady state were obtained.

To the best of our knowledge, the two papers by Schwarz et al. (2006) and Melikov et al. (2016) are the only papers that considered the ROS policy in M/M/1 QIS models in which when the inventory was depleted after the service of a customer was completed, a random order size that followed a discrete probability distribution was instantaneously triggered.

In this paper, we consider an M/M/1 QIS with lost sales and ROS policy by taking into account the server's multiple vacations. When the server finishes the service of a customer and finds the inventory is empty, the server leaves for a vacation. If the server finds that the inventory is empty at the end of a vacation, he/she takes another vacation immediately and continues in the same manner until the server finds the inventory is not empty.

The purpose of this paper is to investigate the following research questions: (a) Does the stationary joint distribution of the queue length, the on-hand inventory level, and the status of the server have a simple product form for the marginal distributions? (b) How does the conditional on-hand inventory level when the server is off due to a vacation differ from the conditional on-hand inventory level when the server is on and working? (c) How does the server's vacation influence the on-hand inventory level and the other performance measures of the system? (d) How does the distribution of random order size influence the performance measures of the system?

The main research contributions of this paper are as follows: (a) We develop a new M/M/1 QIS model with ROS policy by taking into account the server vacation. (b) We derive the stationary joint distribution of the queue length, the on-hand inventory level and the status of the server in product form by using a matrix analytical approach. (c) We obtain the conditional distributions and the conditional expectations of the on-hand inventory level when the server is off due to a vacation or depleted inventory, and when it is on and working. (d) We compute explicitly some performance measures and analytically investigate the effect of the parameter of the server's vacation on these performance measures.

The rest of the paper is organized as follows. We describe the system model in Sect. 2. In Sect. 3, we first derive the stability condition of the system by using a QBD process theory. Then, we derive the stationary joint distribution of the queue length, the on-hand inventory level, and the status of the server. We further investigate the conditional distributions of the on-hand inventory level when the server is off due to a vacation or depleted inventory and when it is on and working. Some performance measures are computed and compared with the model shown in Schwarz et al. (2006) for ROS policy in Sect. 4. The effect of the vacation rate on the performance measures is also investigated analytically in this section. Some numerical results are presented in Sect. 5. In Sect. 6, we perform a simulation study for the case of more general distributions of the inter-arrival times and the service times. In Sect. 7, we present some managerial suggestions. Conclusions are given in Sect. 8.

2 Description of the model

In this section, we describe the proposed model in more detail. Figure 1 shows the schematic diagram of the proposed model.

We consider an M/M/1 QIS under continuous review with ROS policy and multiple server vacations. Customers arrive in the system according to a Poisson process with rate λ. Each customer requires one unit of an item and is served by a single server under a First-Come, First-Service (FCFS) discipline. The service time follows an exponential distribution with rate μ.

The item for a customer is counted in the inventory until the end of service for that customer. A served customer departs immediately from the system and the on-hand inventory decreases by one at the moment of service completion. If the server is ready to serve a customer which is at the head of the line and there is no item of inventory, this service starts only at the time instant when the next replenishment arrives at the inventory.

When the server finishes serving of a customer and finds the inventory is empty, the server leaves for a vacation that follows an exponential distribution with parameter θ. If the server finds that the inventory is not empty at the end of a vacation, the server returns from the vacation and serves any customers waiting for service. If the server finds that the inventory is still empty at the end of a vacation, the server takes another vacation immediately and continues in the same manner until the server finds the on-hand inventory is not empty.

Considering ROS policy, once the inventory is depleted after the service of a customer is completed, a random order size D that follows a discrete probability distribution on integers $E = \{1, 2, \ldots, M\}$ is instantaneously triggered. The size D of the replenishment order is k with the probability p_k, where $\sum_{k=1}^{M} p_k = 1$. The corresponding distribution function is denoted by F_p. The mean order size is denoted by $\overline{p}$. Let q_k be the probability that the size of a replenishment order is at least k, i.e., $q_k = \sum_{j=k}^{M} p_j, k = 1, 2, \ldots, M$. The replenishment lead time is exponentially distributed with parameter η.

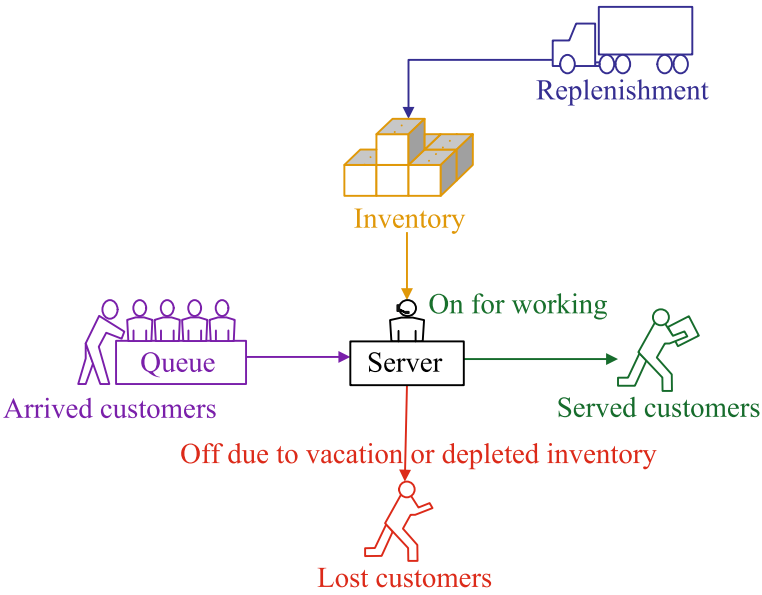

Fig. 1 Schematic diagram of the proposed queueing-inventory model

It is assumed that customers are prevented from entering the system either when the on-hand inventory level is zero or when the server is off due to a vacation. Order size decisions and lead times are independent of the arrival process, the customer service time, and the server's vacation.

3 Steady-state analysis

In this section, we perform the steady-state analysis for the system model described in the previous section.

3.1 Stability condition

Let $\{\boldsymbol{S(t)}, t \geq 0\} = \{(X(t), Y(t), Z(t)), t \geq 0\}$ be the state process of the system, where $X(t)$ denotes the number of customers at time t, $Y(t)$ denotes the inventory level at time t, and $Z(t)$ denotes the status of the server at time t where $Z(t)$ is defined to be either 0 or 1 according to whether the server is off due to a vacation or depleted inventory, or on and working. Then, the process $\{\boldsymbol{S(t)}, t \geq 0\}$ is a QBD process with state space:

$$\Omega = \cup_{n=0}^{\infty}\{n\}$$

where

$$n = \{(n, 0, 0), (n, 1, 0), (n, 1, 1), \ldots, (n, M, 0), (n, M, 1)\}$$

is the collection of states with $X(t) = n$, $n \geq 0$, called the level n. The state-transition diagram of the QIS with server vacations is presented in Fig. 2.

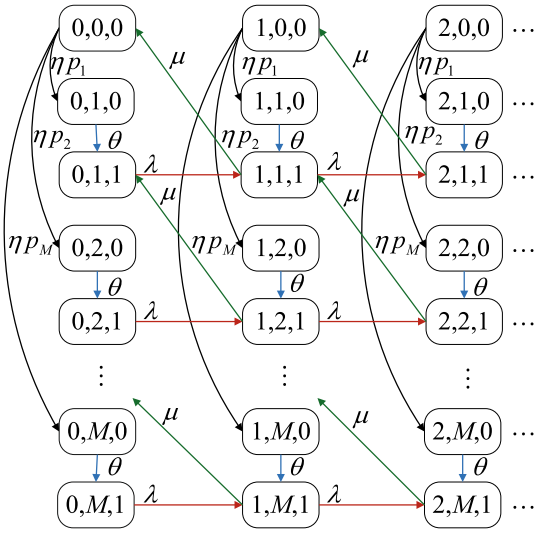

Fig. 2 State-transition diagram of the QIS with server vacations

The infinitesimal generator of the process $\{\boldsymbol{S(t)}, t \geq 0\}$ is as follows:

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{A}_0 & \boldsymbol{C} & & & \\ \boldsymbol{B} & \boldsymbol{A} & \boldsymbol{C} & & \\ & \boldsymbol{B} & \boldsymbol{A} & \boldsymbol{C} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

where $\boldsymbol{A}_0$, $\boldsymbol{B}$, $\boldsymbol{A}$ and $\boldsymbol{C}$ are all square matrices of the order $2M+1$, and they are given as follows:

$$\boldsymbol{A}_0 = \begin{pmatrix} -\eta & \eta p_1 & 0 & \eta p_2 & 0 & \cdots & \eta p_M & 0 \\ 0 & -\theta & \theta & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -\lambda & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & -\theta & \theta & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & -\lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -\theta & \theta \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & -\lambda \end{pmatrix},$$

$$\boldsymbol{B} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \mu & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \mu & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & \mu & 0 & 0 \end{pmatrix},$$

$$\boldsymbol{A} = \boldsymbol{A}_0 - \frac{\mu}{\lambda}\boldsymbol{C}$$

where,

$$\boldsymbol{C} = \text{diag}\{0, 0, \lambda, \ldots, 0, \lambda\}$$

is a diagonal matrix.

Theorem 1 *The process* $\{\boldsymbol{S(t)}, t \geq 0\}$ *with the infinitesimal generator* $\boldsymbol{Q}$ *is positive recurrent if and only if* $\rho = \dfrac{\lambda}{\mu} < 1$.

Proof To derive the stability condition of the process $\{\boldsymbol{S(t)}, t \geq 0\}$, we consider the matrix $\boldsymbol{H} = \boldsymbol{A} + \boldsymbol{B} + \boldsymbol{C}$, which is given by

$$\boldsymbol{H} = \begin{pmatrix} -\eta & \eta p_1 & 0 & \eta p_2 & 0 & \cdots & 0 & \eta p_M & 0 \\ 0 & -\theta & \theta & 0 & 0 & \cdots & 0 & 0 & 0 \\ \mu & 0 & -\mu & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & -\theta & \theta & \cdots & 0 & 0 & 0 \\ 0 & 0 & \mu & 0 & -\mu & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & -\theta & \theta \\ 0 & 0 & 0 & 0 & 0 & \cdots & \mu & 0 & -\mu \end{pmatrix}.$$

Let $\boldsymbol{\pi} = (\pi(0,0), \pi(1,0), \pi(1,1), \ldots, \pi(M,0), \pi(M,1))$ be the steady-state probability vector of the generator $\boldsymbol{H}$. Then, $\boldsymbol{\pi}$ satisfies equations $\boldsymbol{\pi H} = \boldsymbol{0}$ and $\boldsymbol{\pi e} = 1$, where $\boldsymbol{e}$ is a column vector of 1's of appropriate dimension. Solving these equations, we obtain

$$\pi(0,0) = \frac{\mu}{\eta} K_h^{-1}, \tag{1}$$

$$\pi(i,0) = \frac{\mu}{\theta} p_i K_h^{-1}, \ i = 1, 2, \ldots, M, \tag{2}$$

$$\pi(i,1) = q_i K_h^{-1}, \ i = 1, 2, \ldots, M \tag{3}$$

where

$$K_h = \frac{\mu}{\theta} + \frac{\mu}{\eta} + \overline{p}. \tag{4}$$

From Neuts (1981), the process $\{\boldsymbol{S(t)}, t \geq 0\}$ is positive recurrent if and only if

$$\boldsymbol{\pi C e} < \boldsymbol{\pi B e},$$

which is equivalent to

$$\lambda \sum_{i=1}^{M} \pi(i,1) < \mu \sum_{i=1}^{M} \pi(i,1). \tag{5}$$

Using Eqs. (3) and (4), it is easy to verify that

$$\sum_{i=1}^{M} \pi(i,1) = \overline{p} K_h^{-1} > 0.$$

Thus, Eq. (5) implies $\dfrac{\lambda}{\mu} < 1$. □

Remark 1 Theorem 1 shows that the stability condition for the present model is the same as that of the M/M/1 classical queueing system, and this stability condition is independent from the parameters of the server's vacation, the replenishment lead time, and the distribution of the random order size.

3.2 Stationary distribution

For computing the stationary distribution of the process $\{\boldsymbol{S(t)}, t \geq 0\}$, we first consider a QIS with ROS policy and negligible service time. The other assumptions are the same as those given earlier. The corresponding Markov process for this case is defined as $\{\hat{\boldsymbol{S}}(\boldsymbol{t}), t \geq 0\} = \{(Y(t), Z(t)), t \geq 0\}$, where $Y(t)$ and $Z(t)$ are defined as above. The state space of the process $\{\hat{\boldsymbol{S}}(\boldsymbol{t}), t \geq 0\}$ is given as follows:

$$\hat{\Omega} = \{(0, 0), (1, 0), (1, 1), \ldots, (M, 0), (M, 1)\}.$$

The state-transition diagram of the QIS with negligible service time and server vacations is presented in Fig. 3.

This QIS's infinitesimal generator $\hat{\boldsymbol{Q}}$ is given by

$$\hat{\boldsymbol{Q}} = \begin{pmatrix} -\eta & \eta p_1 & 0 & \eta p_2 & 0 & \cdots & 0 & \eta p_M & 0 \\ 0 & -\theta & \theta & 0 & 0 & \cdots & 0 & 0 & 0 \\ \lambda & 0 & -\lambda & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & -\theta & \theta & \cdots & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & -\lambda & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & -\theta & \theta \\ 0 & 0 & 0 & 0 & 0 & \cdots & \lambda & 0 & -\lambda \end{pmatrix}.$$

Let $\hat{\boldsymbol{\pi}} = (\hat{\pi}(0, 0), \hat{\pi}(1, 0), \hat{\pi}(1, 1), \ldots, \hat{\pi}(M, 0), \hat{\pi}(M, 1))$ be the steady-state probability vector of the generator $\{\hat{\boldsymbol{S}}(\boldsymbol{t}), t \geq 0\}$. Then, $\hat{\boldsymbol{\pi}}$ satisfies the set of equations:

$$\begin{cases} \hat{\boldsymbol{\pi}}\,\hat{\boldsymbol{Q}} = \boldsymbol{0} \\ \hat{\boldsymbol{\pi}}\boldsymbol{e} = 1. \end{cases} \tag{6}$$

From matrix $\boldsymbol{H}$ and matrix $\hat{\boldsymbol{Q}}$, we observe that matrix $\hat{\boldsymbol{Q}}$ can be obtained if we change all μ in matrix $\boldsymbol{H}$ by λ. Thus, we can directly get the stationary probability distribution of the

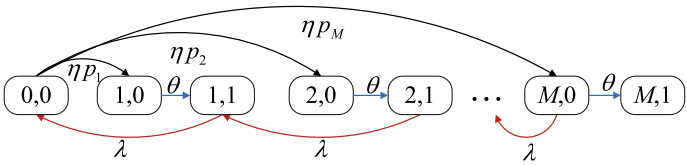

Fig. 3 State-transition diagram of the QIS with negligible service time and server vacations

process $\{\hat{\boldsymbol{S}}(\boldsymbol{t}), t \geq 0\}$ from Eqs. (1)-(4). Therefore, we have

$$\hat{\pi}(0,0) = \frac{\lambda}{\eta} K_v^{-1}, \tag{7}$$

$$\hat{\pi}(i,0) = \frac{\lambda}{\theta} p_i K_v^{-1},\ i = 1, 2, \ldots, M, \tag{8}$$

$$\hat{\pi}(i,1) = q_i K_v^{-1},\ i = 1, 2, \ldots, M \tag{9}$$

where

$$K_v = \frac{\lambda}{\theta} + \frac{\lambda}{\eta} + \overline{p}. \tag{10}$$

Using the steady-state probability vector $\hat{\boldsymbol{\pi}}$ given by Eqs. (7)-(10), we establish the stationary distribution of our system model described in Sect. 2. For this, let $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \ldots)$ be the steady-state probability vector of the process $\boldsymbol{S}(\boldsymbol{t})$, where

$$\boldsymbol{x}_n = (x(n,0,0), x(n,1,0), x(n,1,1), \ldots, x(n,M,0), x(n,M,1)),\ n = 0, 1, \ldots.$$

Then, the steady-state probability vector $\boldsymbol{x}$ satisfies the set of equations:

$$\begin{cases} x\boldsymbol{Q} = \boldsymbol{0} \\ x\boldsymbol{e} = 1. \end{cases} \tag{11}$$

We can obtain the steady-state probability vector $\boldsymbol{x}$ by solving Eq. (11). The solution is given by the following theorem.

Theorem 2 *If $\rho < 1$, the steady-state probability vector $\boldsymbol{x}$ of the process $\{\boldsymbol{S}(\boldsymbol{t}), t \geq 0\}$ is given by*

$$\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots)$$

where

$$\boldsymbol{x}_n = (1-\rho)\,\rho^n \hat{\boldsymbol{\pi}},\ n \geq 0 \tag{12}$$

and the components of the vector $\hat{\boldsymbol{\pi}}$ are given by Eqs. (7)-(10).

Proof The first equation of the set of Eq. (11) can be rewritten as follows:

$$\boldsymbol{x}_0\boldsymbol{A}_0 + \boldsymbol{x}_1\boldsymbol{B} = \boldsymbol{0}, \tag{13}$$

$$\boldsymbol{x}_n\boldsymbol{C} + \boldsymbol{x}_{n+1}\boldsymbol{A} + \boldsymbol{x}_{n+2}\boldsymbol{B} = \boldsymbol{0},\ n \geq 0. \tag{14}$$

Let

$$\boldsymbol{x}_n = \xi\rho^n \hat{\boldsymbol{\pi}},\ n \geq 0 \tag{15}$$

where ξ is a constant. We need to verify that Eqs. (13) and (14) are satisfied by Eq. (15). Substituting Eq. (15) into the left side of Eq. (13), we have

$$\boldsymbol{x}_0\boldsymbol{A}_0 + \boldsymbol{x}_1\boldsymbol{B} = \xi\hat{\boldsymbol{\pi}}\,(\boldsymbol{A}_0 + \rho\boldsymbol{B})\,.$$

Substituting Eq. (15) into the left side of Eq. (14), we have

$$\begin{aligned} \boldsymbol{x}_n\boldsymbol{C} + \boldsymbol{x}_{n+1}\boldsymbol{A} + \boldsymbol{x}_{n+2}\boldsymbol{B} &= \xi\rho^n\hat{\boldsymbol{\pi}}\left[\boldsymbol{C} + \rho\boldsymbol{A} + \rho^2\boldsymbol{B}\right] \\ &= \xi\rho^n\hat{\boldsymbol{\pi}}\left[\boldsymbol{C} + \rho\left(\boldsymbol{A}_0 - \frac{1}{\rho}\boldsymbol{C}\right) + \rho^2\boldsymbol{B}\right] \\ &= \xi\rho^{n+1}\hat{\boldsymbol{\pi}}\,(\boldsymbol{A}_0 + \rho\boldsymbol{B})\,,\ n \geq 0. \end{aligned}$$

From the structure of the matrices $\boldsymbol{A}_0$, $\boldsymbol{B}$ and $\hat{\boldsymbol{Q}}$, it is easy to verify that $\boldsymbol{A}_0 + \rho\boldsymbol{B} = \hat{\boldsymbol{Q}}$. Then, we have

$$\hat{\boldsymbol{\pi}}\,(\boldsymbol{A}_0 + \rho\boldsymbol{B}) = \hat{\boldsymbol{\pi}}\,\hat{\boldsymbol{Q}} = \boldsymbol{0}.$$

Thus, Eqs. (13) and (14) are satisfied by Eq. (15). Applying the normalizing condition $\boldsymbol{x}\boldsymbol{e} = 1$ and noting that $\hat{\boldsymbol{\pi}}\boldsymbol{e} = 1$, we get $\xi = 1 - \rho$. □

Remark 2 Theorem 2 shows that the stationary distribution of the system has a product form of two marginal distributions: One is the stationary distribution of the queue length in the M/M/1 traditional queueing system with the same parameters λ and μ, and the other one is the stationary distribution of the on-hand inventory level of the QIS system with server's multiple vacations and ROS policy when the service time is negligible.

3.3 Conditional distributions of the on-hand inventory level

In this subsection, we first compute the marginal stationary distributions of the queue length, the on-hand inventory level and the status of the server, respectively. Then, we investigate the conditional distributions of the mean on-hand inventory level when the server is off due to a vacation or depleted inventory, and when it is on and working.

Theorem 3 *(a) The marginal stationary distribution of the queue length* $\{X(t), t \geq 0\}$ *is equal to the stationary distribution of the queue length in the classical M/M/1-FCFS system with the same parameter* λ *and* μ.
(b) The marginal stationary distribution of the on-hand inventory level $\{Y(t), t \geq 0\}$ *is given by*

$$P(Y = k) = \begin{cases} \dfrac{\lambda}{\eta} K_v^{-1}, & k = 0 \\ \left(\dfrac{\lambda}{\theta} p_k + q_k\right) K_v^{-1}, & k = 1, 2, \ldots, M. \end{cases} \tag{16}$$

The mean on-hand inventory level is given by

$$\overline{I} = \left(\frac{\lambda}{\theta}\overline{p} + \sum_{i=1}^{M} i q_i\right) K_v^{-1}. \tag{17}$$

(c) The marginal stationary distribution of the status of the server $\{Z(t), t \geq 0\}$ *is given by*

$$P(Z = k) = \begin{cases} \left(\dfrac{\lambda}{\theta} + \dfrac{\lambda}{\eta}\right) K_v^{-1}, & k = 0 \\ \overline{p} K_v^{-1}, & k = 1. \end{cases} \tag{18}$$

Proof The results are directly obtained using Theorem 2. The detail for the proof is omitted.□

Theorem 4 *(a) The conditional distributions of the on-hand inventory level when the server is off due to a vacation or depleted inventory, and when it is on and working are given by*

$$P(Y = k | Z = 0) = \begin{cases} \dfrac{\theta}{\eta + \theta}, & k = 0 \\ \dfrac{\eta}{\eta + \theta} p_k, & k = 1, 2, \ldots, M \end{cases} \tag{19}$$

and

$$P(Y = k|Z = 1) = \frac{q_k}{\overline{p}},\ k = 1, 2, \ldots, M. \tag{20}$$

(b) The conditional mean on-hand inventory levels when the server is off due to a vacation or depleted inventory, and when it is on and working are given by

$$E(Y|Z = 0) = \frac{\eta\overline{p}}{\eta + \theta} \tag{21}$$

and

$$E(Y|Z = 1) = \frac{1}{\overline{p}} \sum_{k=1}^{M} kq_k. \tag{22}$$

Proof The detail for the proof is omitted since it is as simple as that of Theorem 3. □

For Eq. (17), let $\theta \to \infty$, we obtain the mean oh-hand inventory level denoted by $\overline{I}^s$ as follows:

$$\overline{I}^s = \sum_{i=1}^{M} iq_i K_Y^{-1} \tag{23}$$

where

$$K_Y = \frac{\lambda}{\eta} + \overline{p}.$$

This agrees with the corresponding result for the QIS model with ROS policy that was given by Schwarz et al. (2006) (see p. 60, Eq. (5)).

Remark 3 (a) From Eqs. (19) and (21), it is observed that the conditional distribution of the on-hand inventory level when the server is off due to a vacation or depleted inventory and its expectation $E(Y|Z = 0)$ is independent from the arrival rate λ, and that they are not dependent on parameters η and θ individually but only on their proportions η/θ. (b) From Eqs. (20) and (22), it is observed that the conditional distribution of the on-hand inventory level when the server is on and working and its expectation $E(Y|Z = 1)$ is completely independent from parameters λ, η and θ, and that they are only dependent on the distribution of F_p. (c) There is also independence of μ as well throughout for all the performance measures mentioned above.

Remark 4 (a) Equation (21) shows that the conditional mean on-hand inventory level when the server is off due to a vacation or depleted inventory $E(Y|Z = 0)$ is less than the mean order size $\overline{p}$. (b) Using Eq. (23), for Eq. (22), we have

$$E(Y|Z = 1) = \left(1 + \frac{\lambda}{\eta\overline{p}}\right)\overline{I}^s > \overline{I}^s,$$

i.e., the conditional mean on-hand inventory level when the server is on and working is larger than the mean on-hand inventory $\overline{I}^s$ for the QIS model with ROS policy presented in Schwarz et al. (2006).

4 Performance measures and monotonicity

In this section, we derive other performance measures in addition to the mean on-hand inventory level and the conditional mean on-hand inventory levels that have been obtained earlier. Then, we compare the performance measures for this model to those for the QIS model with ROS policy presented in Schwarz et al. (2006). Finally, in order to understand the effect of the vacation rate on the performance measures, we consider the monotonicity of these performance measures in θ.

4.1 Performance measures

In this subsection, we compute some performance measures using the stationary distribution given in Sect. 3.2.

(a) The expected number of inventory replenished per unit of time (reorder rate) is given by

$$\lambda_R = \sum_{k=1}^{M} \eta p_k P(Y = 0) = \lambda K_v^{-1}. \tag{24}$$

(b) The average number of lost sales incurred per unit of time is given by

$$\overline{LS} = \lambda P(Z = 0) = \left(\frac{\lambda^2}{\eta} + \frac{\lambda^2}{\theta}\right) K_v^{-1}. \tag{25}$$

(c) A cycle is defined as the time between the placing of two successive orders, see Schwarz et al. (2006) (p. 60). So, the mean cycle time is λ_R^{-1}. Thus, the mean number of lost sales per cycle is given by

$$\overline{LS}_c = \frac{\overline{LS}}{\lambda_R} = \frac{\lambda}{\eta} + \frac{\lambda}{\theta}. \tag{26}$$

(d) According to Schwarz et al. (2006) (p. 61), β-service level is defined by

$$\beta = \frac{\text{E(satisfied demand per unit of time)}}{\text{E(total demand per unit of time)}}.$$

Thus, the β-service level is given by

$$\beta = \frac{\lambda - \overline{LS}}{\lambda} = 1 - \left(\frac{\lambda}{\eta} + \frac{\lambda}{\theta}\right) K_v^{-1}. \tag{27}$$

(e) The mean arrival rate of customers who are admitted to the system per unit of time is given by

$$\lambda_A = \lambda - \overline{LS} = \lambda\beta. \tag{28}$$

(f) Let $\overline{L}_0$ and $\overline{L}_1$ denote the mean number of customers in the system and the mean number of customers in the queue, respectively. Then, we have

$$\overline{L}_0 = \frac{\lambda}{\mu - \lambda} \tag{29}$$

and

$$\begin{aligned}\overline{L}_1 &= \sum_{n=0}^{\infty}\sum_{i=0}^{M} nx(n,i,0) + \sum_{n=1}^{\infty}\sum_{i=1}^{M}(n-1)x(n,i,1)\\ &= \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu}\overline{p}K_v^{-1}.\end{aligned} \tag{30}$$

(g) From Little's formula, the customer's mean sojourn time $\overline{W}_0$ and the mean waiting time $\overline{W}_1$ are given by

$$\overline{W}_0 = \frac{\overline{L}_0}{\lambda_A} = \frac{K_v}{(\mu-\lambda)\overline{p}} \tag{31}$$

and

$$\overline{W}_1 = \frac{\overline{L}_1}{\lambda_A} = \frac{K_v}{(\mu-\lambda)\overline{p}} - \frac{1}{\mu}. \tag{32}$$

Remark 5 (a) From the above expressions of the performance measures, we observe that some performance measures like $\overline{I}$, $\overline{LS}_c$ and β are not dependent on parameters λ, η and θ individually but only on their proportions λ/η and λ/θ. (b) Concerning the influence of F_p, we observe that all the performance measures derived above, other than $\overline{I}$ and $E(Y|Z=1)$, are only dependent on the first moment $\overline{p}$ of F_p, or completely independent of F_p like $\overline{LS}_c$ and $\overline{L}_0$. (c) μ is also independent throughout for some performance measures related to inventory including $\overline{I}$, λ_R, $\overline{LS}$, β and λ_A. This is because that the stationary distribution of the system has a product form of two marginal distributions (see Theorem 2).

4.2 Comparison with Schwarz et al. (2006)

If we let $\theta \to \infty$ in all the performance measures expressed above, we can obtain the corresponding performance measures of the QIS with ROS policy which have been obtained by Schwarz et al. (2006). We use superscript 's' to denote the corresponding performance measures, e.g., $\overline{I}^s$, λ_R^s, $\overline{LS}^s$, $\overline{LS}_c^s$, β^s, λ_A^s, $\overline{W}_0^s$ and $\overline{W}_1^s$, for the QIS model with ROS policy that was studied by Schwarz et al. (2006).

For the mean on-hand inventory level $\overline{I}$, from Eqs. (17) and (23), we have the following decomposition:

$$\overline{I} = \alpha\overline{I}^s + (1-\alpha)\overline{p} \tag{33}$$

where α is a positive constant and is given by

$$\alpha = \frac{\dfrac{\lambda}{\eta} + \overline{p}}{\dfrac{\lambda}{\theta} + \dfrac{\lambda}{\eta} + \overline{p}} \tag{34}$$

and $\overline{I}^s$ is given by Eq. (23).

Remark 6 This decomposition shows that the mean on-hand inventory level $\overline{I}$ is the weighted average sum of the mean on-hand inventory level $\overline{I}^s$ of the QIS model with ROS policy presented in Schwarz et al. (2006) and the mean order size $\overline{p}$. It is also easy to see that the weight number α increases with θ, and that α approaches to one when $\theta \to \infty$.

Table 1 Relationships of the performance measures to our model and the corresponding model in Schwarz et al. (2006)

Performance measures	Relations
$\overline{I}$	$\overline{I} = \alpha \overline{I}^s + (1-\alpha)\overline{p}$
λ_R	$\lambda_R = \alpha \lambda_R^s$
$\overline{LS}$	$\overline{LS} = \alpha \overline{LS}^s$
$\overline{LS}_c$	$\overline{LS}_c = \overline{LS}_c^s + \frac{\lambda}{\theta}$
β	$\beta = \alpha \beta^s + (1-\alpha)$
λ_A	$\lambda_A = \alpha \lambda_A^s + (1-\alpha)\lambda$
$\overline{W}_0$	$\overline{W}_0 = \overline{W}_0^s + \Delta_1$
$\overline{W}_1$	$\overline{W}_1 = \overline{W}_1^s + \Delta_2$

Similarly, we can derive other relationships that exists between the performance measures for our model and the corresponding performance measures for the model presented in Schwarz et al. (2006). All these relationships are summarized in Table 1, where Δ_1 and Δ_2 are given by

$$\Delta_1 = \frac{\lambda}{\theta \overline{p}(\mu - \lambda)} \tag{35}$$

and

$$\Delta_2 = \frac{\lambda}{\theta \overline{p}(\mu - \lambda)} + \frac{\lambda}{\mu \eta \overline{p}}. \tag{36}$$

4.3 Monotonicity

In the following, we consider the monotonicity of the performance measures on the vacation rate θ by referring to the relationships given in Table 1.

Using the relationships for $\overline{I}$ given in Table 1, we have

$$\frac{\mathrm{d}\overline{I}}{\mathrm{d}\theta} = \frac{\mathrm{d}\alpha}{\mathrm{d}\theta}(\overline{I}^s - \overline{p}).$$

We note that $\frac{\mathrm{d}\alpha}{\mathrm{d}\theta} < 0$. Hence, we find the following conditions for the monotonicity of $\overline{I}$ on θ: (a) If $\overline{I}^s < \overline{p}$ then $\overline{I}$ increases with θ. (b) If $\overline{I}^s > \overline{p}$ then $\overline{I}$ decreases with θ. (c) If $\overline{I}^s = \overline{p}$ then the vacation rate θ does not influence $\overline{I}$.

Noting that $\frac{\mathrm{d}\alpha}{\mathrm{d}\theta} < 0$, it is easy to see from Table 1 that some performance measures like λ_R, $\overline{LS}$, $\overline{LS}_c$, $\overline{W}_0$ and $\overline{W}_1$ decrease with θ.

Using the relationships for β and λ_A given in Table 1, we have

$$\frac{\mathrm{d}\beta}{\mathrm{d}\theta} = \frac{\mathrm{d}\alpha}{\mathrm{d}\theta}(\beta^s - 1) > 0$$

and

$$\frac{\mathrm{d}\lambda_A}{\mathrm{d}\theta} = \frac{\mathrm{d}\alpha}{\mathrm{d}\theta}(\lambda_A^s - \lambda) > 0.$$

Hence, β and λ_A increase with θ.

5 Numerical examples

In this section, we investigate three examples for the distribution of the random order size which have the same mean order size.

(i) *Deterministic distribution (DET).* Let us assume that the order size is fixed and equal to deterministic number $d \in E = \{1, 2, \ldots, M\}$. Hence, we have $p_d = 1$, and $p_k = 0$ for other $k \in E$ and $k \neq d$. The mean order size $\overline{p} = d$. This is the $(0, d)$ inventory policy.

(ii) *Uniform distribution (UNI).* Let the order size be equally distributed in set E. Hence, we have $p_k = 1/M$ for all $k \in E$. We fix $M = 2d - 1$, so that the mean order size $\overline{p} = d$.

(iii) *Modified binomial distribution (MBI).* Let us assume that the order size follows the modified binomial distribution in set E with the following probability distribution:

$$p_k = \begin{cases} q^M + Mpq^{M-1}, & k = 1 \\ C_M^k p^k q^{M-k}, & k = 2, 3, \ldots, M \end{cases}$$

where $p > 0$, $p + q = 1$. It is easy to see the mean order size $\overline{p} = Mp + q^M$. We select an integer M such that it approximately satisfies the equality $Mp + q^M = d$, so that the mean order size will be approximately equal to d.

Following this, we examine the effect of the above three distributions of the random order size on the mean on-hand inventory level $\overline{I}$. We plot the curves for $\overline{I}$ by varying the order size distributions and the parameters λ, η and θ, respectively. We set $d = 6$. For the deterministic distribution, we fix $M = 6$. For the uniform distribution, we fix $M = 11$. For the modified binomial distribution, we fix $M = 15$ and $p = 0.4$. Thus, the mean order sizes for each of the three distributions are equal to 6, or are approximately equal to 6. Figure 4 corresponds to the case of the varying parameter λ and the fixed parameters $\eta = 3$ and $\theta = 5$. Figure 5 corresponds to the case of the varying parameter η and the fixed parameters $\lambda = 10$ and $\theta = 5$. Figure 6 corresponds to the case of the varying parameter θ and the fixed parameters $\lambda = 10$ and $\eta = 3$.

Figure 4 shows that the mean on-hand inventory level $\overline{I}$ decreases with parameter λ for each of the three order size distributions. It is observed that the difference of $\overline{I}$ under the three distributions deceases with an increase in parameter λ. It is observed from Fig. 5 that the mean on-hand inventory level $\overline{I}$ increases with parameter η for each of the three distributions, and the difference of $\overline{I}$ under the three distributions increases with an increase in parameter η. From Fig. 6, we observe that the mean on-hand inventory level $\overline{I}$ decreases with parameter θ for each of the three distributions, and the difference of $\overline{I}$ under the three distributions increases with an increase in parameter θ.

It is observed from Figs. 4, 5 and 6 that the mean on-hand inventory level $\overline{I}$ is minimal for a deterministic distribution, and is maximal for a uniform distribution.

6 Simulation study

From the performance measures obtained in Sects. 3 and 4, we found some invariance properties which can be summarized as follows: (i) some performance measures related to inventory management (including $\overline{I}$, $E(Y|Z = 0)$, $E(Y|Z = 1)$; λ_R, $\overline{LS}$, $\overline{LS}_C$, β and λ_A) are not dependent on the service rate μ (see Remarks 3 and 5); (ii) the mean number of customers

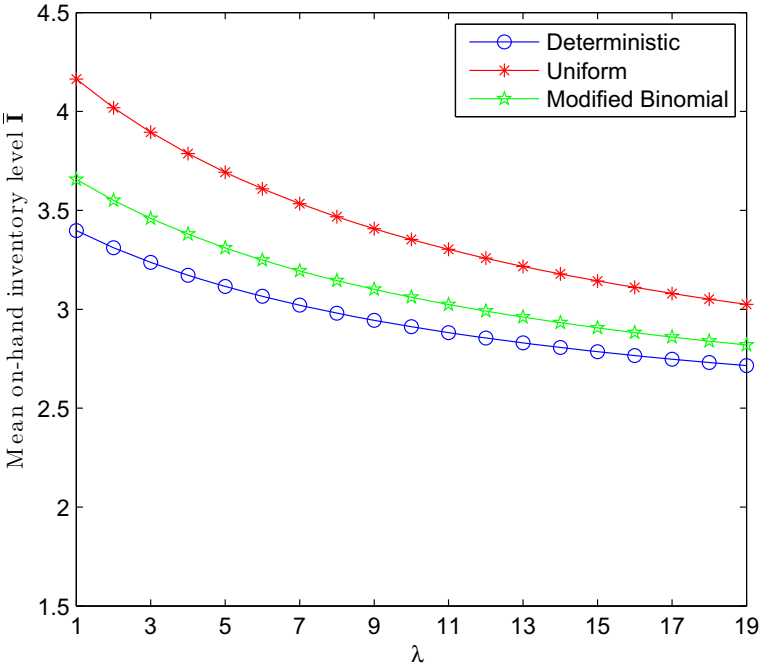

Fig. 4 Effect of parameter λ on the mean on-hand inventory level $\bar{I}$

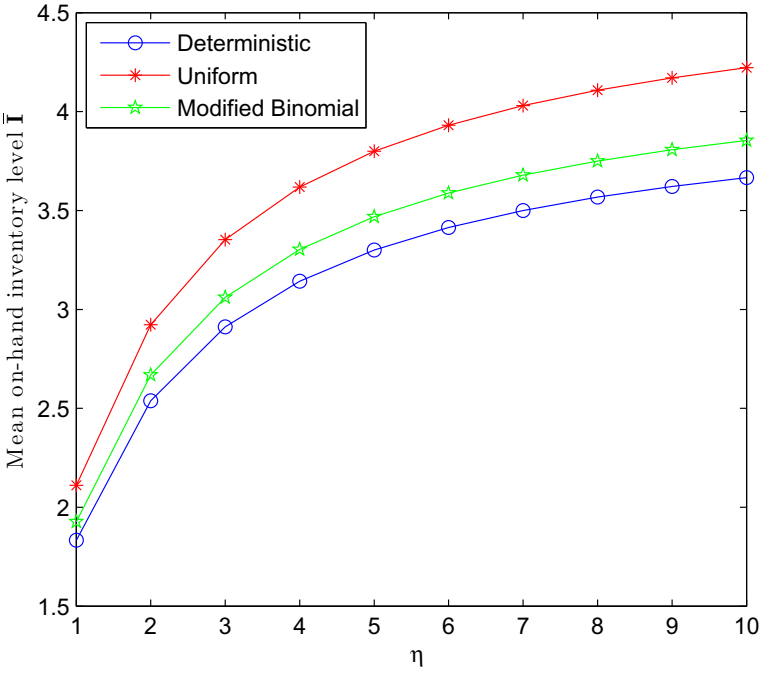

Fig. 5 Effect of parameter η on the mean on-hand inventory level $\bar{I}$

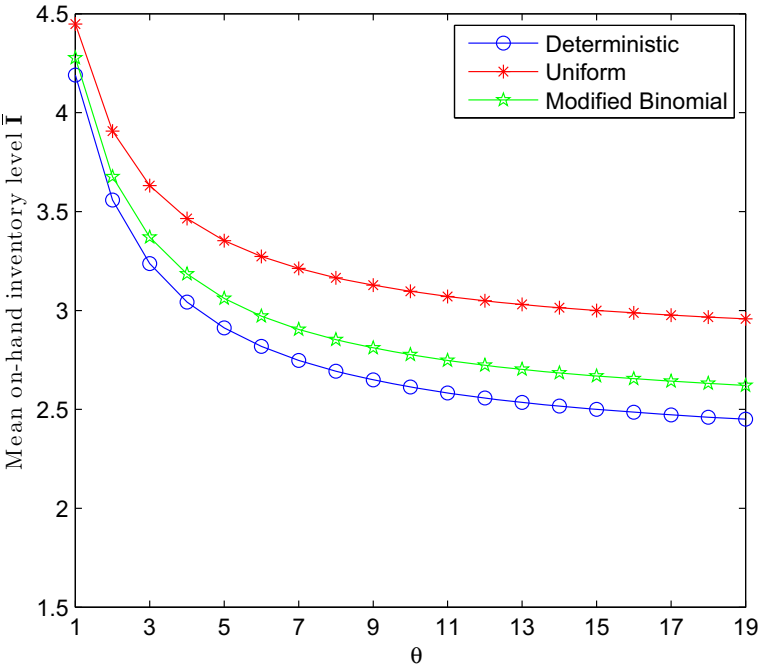

Fig. 6 Effect of parameter θ on the mean on-hand inventory level $\overline{I}$

in the system $\overline{L}_0$ and the conditional mean on-hand inventory level $E(Y|Z = 1)$ are not dependent on the parameters η and θ; (iii) the conditional mean on-hand inventory levels $E(Y|Z = 0)$ and $E(Y|Z = 1)$ are not dependent on the arrival rate λ. These properties have been analytically proven for the case of M/M/1 QIS model described in Sect. 2. However, if the inter-arrival times and the service times are not exponentially distributed and the other assumptions are the same as that in our model described in Sect. 2, can we have these invariance properties?

Since it is not easy to obtain the above performance measures in closed form for the cases of non-exponential distributions of the inter-arrival times and the service times, the simulation method will be used to obtain the above performance measures. Therefore, in this section, we perform simulation experiments with various distribution settings of the inter-arrival times and the service times and examine the effect of the variance of the distributions on some of the system performance measures.

Firstly, we perform a set of experiments to test the accuracy of the simulation. For this purpose, we perform the simulations for the M/M/1 QIS model in Sect. 2 and compare the simulation results with the analytical results for the mean on-hand inventory level $\overline{I}$ obtained in Sect. 3. The relative difference for the performance measures $\overline{I}$ is computed by

$$\epsilon = 100\frac{|\overline{I}_{sim} - \overline{I}_{exa}|}{\overline{I}_{sim}}$$

where $\overline{I}_{exa}$ and $\overline{I}_{sim}$ represent the exact result and the simulation result, respectively.

In this set of experiments, we fix the arrival rate $\lambda = 10$ and the service rate $\mu = 15$. Each of the simulation experiments is characterized by the following factors: (i) the distribution of

Table 2 Relative differences between the exact results and the outcomes of the simulations

	DET	UNI	MBI
Average error	3.10	4.15	3.42
Minimum error	0.53	0.35	0.37
Maximum error	9.74	11.95	10.63

the random order size D; (ii) the lead time parameter η, and (iii) the vacation time parameter θ. We consider the three different distributions of the random order size which are described in Sect. 5. The parameters η and θ are varied over 5 levels: 1, 2, 3, 4, 5. The simulation experiments include a total of $3 \times 5 \times 5 = 75$ scenarios. In our simulation study, we make use of the simulation language Python. We stimulate these scenarios 50,000 unit times. The relative differences between the exact results and the outcomes of the simulation are given in Table 2.

From the summary statistics presented in Table 2, we see that the average errors in estimating $\overline{I}$ for the deterministic distribution, the uniform distribution and the modified binomial distribution are 3.10%, 4.15% and 3.42%, respectively. The maximum error in our set of 75 instances is 11.95%, and the minimal error is 0.35%. For the average error of simulating $\overline{I}$, the deterministic distribution is minimal among the three distributions of the random order size D. Thus, in our following sets of experiments, we fix the order size distribution to be a deterministic distribution with fixed order size $D = 6$.

Now, we perform the following sets of experiments to see if we still have the invariance properties mentioned above when the inter-arrival times and the service times are not exponentially distributed. For this purpose, in the design of simulation study, we consider the following three different probability distributions for the inter-arrival times:

(i) *Exponential distribution (EXA).* The probability density function is

$$f_1(x) = \lambda e^{-\lambda x}, \ x > 0.$$

(ii) *Erlang distribution (ERA).* The probability density function is

$$f_2(x) = \frac{\lambda(\lambda x)^{k-1}}{(k-1)!} e^{-\lambda x}, \ x > 0.$$

(iii) *Hyper-exponential distribution (HEA).* The probability density function is

$$f_3(x) = \sum_{i=1}^{n} p_i \lambda_i e^{-\lambda_i x}, \ x > 0.$$

The parameters of the inter-arrival times are normalized so as to obtain the same arrival rate which is denoted by λ_{ar}. We also use the following four different probability distributions for the service times:

(i) *Exponential distribution (EXS).* The probability density function is

$$g_1(x) = \mu e^{-\mu x}, \ x > 0.$$

(ii) *Erlang distribution (ERS).* The probability density function is

$$g_2(x) = \frac{\mu(\mu x)^{k-1}}{(k-1)!} e^{-\mu x}, \ x > 0.$$

Table 3 The effect of the service rate μ_{sr} on $\overline{I}$ under various scenarios

	EXA	ERA	HEA
μ_{sr}=15			
EXS	3.8062	3.7932	3.7876
ERS	3.8006	3.8120	3.7920
HES	3.8018	3.7912	3.7916
LNS	3.7914	3.7898	3.7884
μ_{sr}=17			
EXS	3.7924	3.8162	3.8084
ERS	3.8002	3.8222	3.7996
HES	3.8242	3.8010	3.7972
LNS	3.7908	3.7974	3.8062
μ_{sr}=19			
EXS	3.8096	3.8188	3.8166
ERS	3.8034	3.8308	3.8192
HES	3.8268	3.8108	3.8176
LNS	3.7936	3.8236	3.8110
μ_{sr}=21			
EXS	3.8170	3.8138	3.8090
ERS	3.8316	3.8072	3.8098
HES	3.8054	3.8196	3.8342
LNS	3.8228	3.8356	3.8026

(iii) *Hyper-exponential distribution (HES).* The probability density function is

$$g_3(x) = \sum_{i=1}^{n} p_i \mu_i e^{-\mu_i x}, \ x > 0.$$

(iv) *Log-normal distribution (LNS).* The probability density function is

$$g_4(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}, \ x > 0.$$

The parameters of the service time are normalized so as to obtain the same service rate which is denoted by μ_{sr}.

In the second set of experiments, we study the effect of the service rate μ_{sr} on some of the performance measures. We mainly look at the following performance measures: $\overline{I}$, $E(Y|Z=0)$, $E(Y|Z=1)$, $\overline{LS}$ and λ_A. We consider the three inter-arrival time distributions and the four service time distributions defined above. The service rate is varied over four levels: $\mu_{sr} = 15, 17, 19, 21$. The parameters η and θ are fixed as: $\eta = 4$, $\theta = 2$. Thus, the simulation experiments include a total of $5 \times 3 \times 4 \times 4 = 240$ (5 performance measures $\times$ 3 inter-arrival time distributions $\times$ 4 service time distributions $\times$ 4 levels of service rates) scenarios. The simulation results for $\overline{I}$, $E(Y|Z=0)$, $E(Y|Z=1)$, λ_R and $\overline{LS}$ are displayed in Tables 3, 4, 5, 6 and 7, respectively.

From these tables, we immediately conclude the following observations:

(i) If the service rate μ_{sr} is fixed to any one of the four levels and the service time distribution is fixed to any one of the four distributions: EXS, ERS, HES and LNS, then the variance

Table 4 The effect of the service rate μ_{sr} on $E(Y|Z=0)$ under various scenarios

	EXA	ERA	HEA
μ_{sr}=15			
EXS	3.9620	3.9664	3.9628
ERS	3.9728	3.9672	3.9704
HES	3.9600	3.9722	3.9404
LNS	3.9692	3.9482	3.9828
μ_{sr}=17			
EXS	3.9528	3.9576	3.9528
ERS	3.9718	3.9764	3.9884
HES	3.9442	3.9752	3.9980
LNS	3.9464	3.9574	3.9454
μ_{sr}=19			
EXS	3.9692	3.9390	3.9404
ERS	3.9404	3.9454	3.9500
HES	3.9524	3.9650	3.9598
LNS	3.9554	3.9602	3.9426
μ_{sr}=21			
EXS	3.9450	3.9556	3.9828
ERS	3.9716	3.9596	3.9676
HES	3.9624	3.9748	3.9558
LNS	3.9524	3.9476	3.9552

Table 5 The effect of the service rate μ_{sr} on $E(Y|Z=1)$ under various scenarios

	EXA	ERA	HEA
μ_{sr}=15			
EXS	3.5036	3.5012	3.5090
ERS	3.4980	3.4998	3.5066
HES	3.4964	3.4912	3.4926
LNS	3.4974	3.4970	3.4990
μ_{sr}=17			
EXS	3.5004	3.5014	3.4920
ERS	3.5030	3.5006	3.5010
HES	3.5022	3.4796	3.4992
LNS	3.4948	3.5036	3.5090
μ_{sr}=19			
EXS	3.5020	3.4932	3.4906
ERS	3.5006	3.4982	3.4922
HES	3.5014	3.4856	3.5058
LNS	3.5042	3.5050	3.5024
μ_{sr}=21			
EXS	3.5032	3.4958	3.5038
ERS	3.5026	3.4974	3.4986
HES	3.4994	3.4924	3.4998
LNS	3.4912	3.4942	3.5012

Table 6 The effect of the service rate μ_{sr} on λ_R under various scenarios

	EXA	ERA	HEA
μ_{sr}=15			
EXS	0.0100	0.0100	0.0100
ERS	0.0100	0.0100	0.0100
HES	0.0100	0.0100	0.0100
LNS	0.0100	0.0100	0.0100
μ_{sr}=17			
EXS	0.0100	0.0100	0.0100
ERS	0.0100	0.0100	0.0100
HES	0.0100	0.0100	0.0100
LNS	0.0100	0.0100	0.0100
μ_{sr}=19			
EXS	0.0100	0.0100	0.0100
ERS	0.0100	0.0100	0.0100
HES	0.0100	0.0100	0.0100
LNS	0.0100	0.0100	0.0100
μ_{sr}=21			
EXS	0.0100	0.0100	0.0100
ERS	0.0100	0.0100	0.0100
HES	0.0100	0.0100	0.0100
LNS	0.0100	0.0100	0.0100

Table 7 The effect of the service rate μ_{sr} on $\overline{LS}$ under various scenarios

	EXA	ERA	HEA
μ_{sr}=15			
EXS	0.0554	0.0562	0.0542
ERS	0.0566	0.0564	0.0534
HES	0.0550	0.0552	0.0556
LNS	0.0556	0.0570	0.0564
μ_{sr}=17			
EXS	0.0586	0.0596	0.0586
ERS	0.0596	0.0594	0.0590
HES	0.0586	0.0600	0.0592
LNS	0.0594	0.0588	0.0582
μ_{sr}=19			
EXS	0.0600	0.0600	0.0600
ERS	0.0600	0.0600	0.0600
HES	0.0598	0.0600	0.0598
LNS	0.0598	0.0600	0.0600
μ_{sr}=21			
EXS	0.0600	0.0602	0.0600
ERS	0.0602	0.0600	0.0602
HES	0.0602	0.0602	0.0600
LNS	0.0602	0.0600	0.0600

Table 8 The effect of the parameters η and θ on $\overline{L}_0$ under various scenarios

	EXA		ERA		HEA	
	$\theta = 2$	$\theta = 12$	$\theta = 2$	$\theta = 12$	$\theta = 2$	$\theta = 12$
$\eta = 4$						
EXS	0.0890	0.0896	0.0890	0.0890	0.0894	0.0882
ERS	0.0894	0.0890	0.0896	0.0890	0.0880	0.0884
HES	0.0888	0.0890	0.0900	0.0894	0.0884	0.0892
LNS	0.0886	0.0890	0.0888	0.0894	0.0886	0.0892
$\eta = 14$						
EXS	0.0882	0.0888	0.0894	0.0896	0.0896	0.0878
ERS	0.0888	0.0886	0.0894	0.0898	0.0888	0.0888
HES	0.0890	0.0894	0.0894	0.0898	0.0892	0.0892
LNS	0.0890	0.0882	0.0900	0.0896	0.0880	0.0886

of the different inter-arrival time distributions barely affects the performance measures $\overline{I}$ and $\overline{LS}$, and it especially does not affect λ_R.

(ii) If the service rate μ_{sr} is fixed to any one of the four levels and the inter-arrival distribution is fixed to any one of the three distributions: EXA, ERA and HEA, then the variance of the different service time distributions barely affects the performance measures $\overline{I}$ and $\overline{LS}$, and it especially does not affect λ_R.

(iii) If the inter-arrival time distribution is fixed to any one of the three distributions: EXA, ERA and HEA, and the service time distribution is fixed to any one of the four distributions: EXS, ERS, HES and LNS, then the variance of the different service rates barely affects the performance measures $\overline{I}$ and $\overline{LS}$, and it especially does not affect λ_R.

Since the other performances $\overline{LS}_C$, β and λ_A can be derived by means of $\overline{LS}$ and λ_A (see Eqs. (27)-(29)), we conclude from these observations that if the inter-arrival times and the service times are not exponentially distributed and the other assumptions are the same as that in our model described in Sect. 2, the performance measures related to inventory management mentioned above are not dependent on the service rate μ_{sr}.

In the third set of experiments, we study the effect of the parameters η and θ on the mean number of customers in the system $\overline{L}_0$ and the conditional mean on-hand inventory level $E(Y|Z = 1)$. We consider the three inter-arrival time distributions with the common arrival rate $\lambda_{ar} = 10$ and the four service time distributions with the common service rate $\mu_{sr} = 15$ defined above. The parameters η and θ are varied over two levels: $\eta = 4, 14$, $\theta = 2, 12$. Thus, the simulation experiments include a total of $2 \times 3 \times 4 \times 2 \times 2 = 96$ (2 performance measures $\times$ 3 inter-arrival time distributions $\times$ 4 service time distributions $\times$ 2 levels of η $\times$ 2 levels of θ) scenarios. The simulation results for $\overline{L}_0$ and $E(Y|Z = 1)$ are displayed in Tables 8 and 9, respectively.

From Tables 8 and 9, we observe that there is no significant dependence of the parameter η and θ on the performance measures $\overline{L}_0$ and $E(Y|Z = 0)$ under various scenarios.

In the last set of experiments, we study the effect of the arrival rate λ_{ar} on the conditional mean on-hand inventory levels $E(Y|Z = 0)$ and $E(Y|Z = 1)$. We consider the three inter-arrival time distributions and the four service time distributions with the common service rate $\mu_{sr} = 15$ defined above. The arrival rate λ_{ar} is varied over 3 levels: $\lambda_{ar} = 6, 8, 10$. The parameters η and θ are fixed as: $\eta = 4$, $\theta = 2$. Thus, the simulation experiments include a total of $2 \times 3 \times 4 \times 3 = 96$ (2 performance measures $\times$ 3 inter-arrival time distributions $\times$ 4

Table 9 The effect of the parameter η and θ on $E(Y|Z=1)$ under various scenarios

	EXA		ERA		HEA	
	$\theta=2$	$\theta=12$	$\theta=2$	$\theta=12$	$\theta=2$	$\theta=12$
$\eta=4$						
EXS	3.5014	3.4934	3.4988	3.4930	3.5126	3.5100
ERS	3.4936	3.4994	3.4958	3.4966	3.5046	3.5010
HES	3.4946	3.5094	3.5008	3.4984	3.5088	3.4974
LNS	3.5028	3.4952	3.4980	3.4922	3.5034	3.5000
$\eta=14$						
EXS	3.4994	3.3862	3.5020	3.3332	3.5010	3.4068
ERS	3.4922	3.3818	3.5018	3.3048	3.5010	3.4074
HES	3.5184	3.3890	3.4952	3.3392	3.4980	3.3890
LNS	3.5006	3.3784	3.4960	3.2794	3.4946	3.4053

Table 10 The effect of the arrival rate λ_{ar} on $E(Y|Z=0)$ under various scenarios

	EXA	ERA	HEA
λ_{ar}=6			
EXS	3.9398	3.9608	3.9524
ERS	3.9530	3.9328	3.9330
HES	3.9886	3.9778	3.9684
LNS	3.9548	3.9434	3.9630
λ_{ar}=8			
EXS	3.9506	3.9504	3.9468
ERS	3.9668	3.9614	3.9596
HES	3.9842	3.9550	3.9710
LNS	3.9414	3.9672	3.9588
λ_{ar}=10			
EXS	3.9684	3.9798	3.9618
ERS	3.9678	3.9472	3.9710
HES	3.9484	3.9518	3.9576
LNS	3.9508	3.9742	3.9446

service time distributions $\times$ 3 levels of the arrival rate λ_{ar}) scenarios. The simulation results for $E(Y|Z=0)$ and $E(Y|Z=1)$ are displayed in Tables 10 and 11, respectively.

From Tables 10 and 11, it is observed that there is no significant dependence of the arrival rate λ_{ar} on the performance measures $E(Y|Z=0)$ and $E(Y|Z=1)$ under various scenarios.

From all the observations above, we get an answer for the question proposed at the beginning of this section, i.e., if the inter-arrival times and the service times are not exponentially distributed and the other assumptions are the same as that in our model described in Sect. 2, we still have the invariance properties mentioned above.

Table 11 The effect of the arrival rate λ_{ar} on $E(Y|Z=1)$ under various scenarios

	EXA	ERA	HEA
λ_{ar}=6			
EXS	3.4866	3.4654	3.4744
ERS	3.4662	3.4550	3.4646
HES	3.4670	3.4586	3.4696
LNS	3.4634	3.4512	3.4870
λ_{ar}=8			
EXS	3.4964	3.5016	3.4874
ERS	3.4926	3.4962	3.4932
HES	3.5032	3.5056	3.4954
LNS	3.4992	3.4980	3.4980
λ_{ar}=10			
EXS	3.4938	3.4882	3.5134
ERS	3.5008	3.4994	3.5020
HES	3.4992	3.5024	3.4984
LNS	3.5028	3.4934	3.5038

7 Managerial suggestions

From the performance measures obtained in Sects. 3 and 4, we have found that all the performance measures except the mean on-hand inventory level $\overline{I}$ and the conditional mean on-hand inventory level $E(Y|Z=1)$ are only dependent on the mean order size $\overline{p}$ of the random order size D, or completely independent of the distribution F_p of D like the mean number of lost sales per cycle $\overline{LS}_c$ and the mean number of customers in the system $\overline{L}_0$ (see Remark 5). The analysis of numerical examples in Sect. 6 show that the mean on-hand inventory level $\overline{I}$ is minimal for the deterministic order size distribution. Therefore, we recommend to managers in queueing-inventory systems that they pay more attention on the mean order size than the distribution of the random order size, and that the deterministic order size might be an optimal replenishment policy from the point of view of the mean on-hand inventory level.

In practice, it is common in service systems to allow the server to have a vacation when the server is idle. From economic point of view, allowing a server to take a vacation can reduce the expenses incurred when the server is idle. Also, where the server is a human being, continuous work creates physical stress and mental pressure that reduce the server's working efficiency. As pointed out by Jeganathana and Abdul (2020), "A vacation period helps persons avoid stress factors and restore their energy and confidence to work efficiently". However, from the comparison of our model with the QIS model with no server's vacation and monotonicity of the performance measures on the vacation rate (see Sects. 4.2 and 4.3), we found that server's vacation in QIS has a significant effect on inventory management and the satisfaction of the customers. Therefore, we recommend that managers in QIS must consider the effect of a server's vacation on the inventory management and the quality of service. Neglecting a server's vacation in QIS will lead to poorer outcomes in inventory decision making.

From the analytical and simulation study, we have found that all the performance measures except the mean number of customers in the system $\overline{L}_0$, the mean number of customers in the queue $\overline{L}_1$, the customer's mean sojourn time $\overline{W}_0$ and the mean waiting time $\overline{W}_1$ are not

dependent on the service rate. Therefore, we recommend that a company should not blindly improve the service capacity since the improvement of the service capacity does not guarantee a decrease in the rate of customer losses. However, the improvement of the service capacity will decrease the waiting time of customers and thus increase customers' satisfaction. This will be helpful for improving the company's reputation. However, it will increase the service operating costs. Therefore, managers should consider the inventory management and the service process integrally.

8 Conclusion

In this paper, we studied the queueing-inventory system with random order size policy and lost sales, where a multiple vacation policy for the server was considered when the on-hand inventory was depleted. It was found that the stability condition was independent of the vacation rate, the parameter of lead time, and the distribution of the random order size. We obtained the stationary distribution of the system as a product of the marginal distributions. We observed that the conditional distribution of the on-hand inventory level when the server is off due to a vacation or depleted inventory and its expectation $E(Y|Z=0)$ are independent of the arrival rate λ. We also observed that the conditional distribution of the on-hand inventory level when the server is on and working and its expectation $E(Y|Z=1)$ are completely independent with parameters λ, η and θ. Some monotonicity for the performance measures on the vacation rate was obtained. Numerical examples show that the mean on-hand inventory level for the deterministic distribution of the order size was the minimal among the three distributions of the order size. The simulation study shows that some invariance properties still hold if the inter-arrival times and the service times are not exponentially distributed and the other assumptions are the same as that in our model. Our model can be further extended to a more general case with Phase-type distributions of service times, lead times and vacation times. However, the computational complexity increases significantly with the state space size.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (No. 71971189), and the Key Research Project of Science and Technology, University of Hebei Province, China (No. ZD2018042), and is supported in part by MEXT, Japan.

References

Arun, C. P. (2010). Queueing and inventory theory in clinical practice: Application to clinical toxicology. *Annals of the New York Academy of Sciences*, *919*(1), 284–287.

Baek, J. W., & Moon, S. K. (2014). The M/M/1 queue with a production-inventory system and lost sales. *Applied Mathematics and Computation*, *233*(1), 534–544.

Barron, Y. (2019). A state-dependent perishability (s, S) inventory model with random batch demands. *Annals of Operations Research*, *280*, 65–98.

Doshi, B. T. (1986). Queueing systems with vacations: A survey. *Queueing Systems*, *1*, 29–66.

Jeganathana, K., & Abdul, R. M. (2020). Two parallel heterogeneous servers Markovian inventory system with modified and delayed working vacations. *Mathematics and Computers in Simulation*, *172*, 273–304.

Ke, J. C., Wu, C. H., & Zhang, Z. G. (2010). Recent developments in vacation queuing models: A short survey. *International Journal of Operation Research*, *7*, 3–8.

Koroliuk, V. S., Melikov, A. Z., Ponomarenko, L. A., & Rustamov, A. M. (2017). Asymptotic analysis of the system with server vacation and perishable inventory. *Cybernetics and Systems Analysis*, *53*(4), 543–553.

Koroliuk, V. S., Melikov, A. Z., Ponomarenko, L. A., & Rustamov, A. M. (2018). Model of perishable queueing-inventory system with server vacations. *Cybernetics and Systems Analysis*, *54*(1), 31–44.

Krenzler, R., & Daduna, H. (2015). Loss systems in a random environment steady-state analysis. *Queueing Systems*, *80*(1), 127–153.
Krishnamoorthy, A., Lakshmy, B., & Manikandam, R. (2011). A survey on inventory models with positive service time. *Opsearch*, *48*(2), 153–169.
Krishnamoorthy, A., Manikandan, R., & Lakshmy, B. (2015). A revisit to queueing-inventory system with positive service time. *Annals of Operations Research*, *233*, 221–236.
Krishnamoorthy, A., Shajin, D., & Lakshmy, B. (2016a). Product form solution for some queueing-inventory supply chain problem. *Opsearch*, *53*(1), 85–102.
Krishnamoorthy, A., Shajin, D., & Lakshmy, B. (2016b). On a queueing-inventory with reservation, cancellation, common life time and retrial. *Annals of Operations Research*, *247*, 365–389.
Krishnamoorthy, A., & Viswanath, N. C. (2013). Stochastic decomposition in production inventory with service time. *European Journal of Operational Research*, *228*, 358–366.
Melikov, A. A., & Molchanov, A. A. (1992). Stock optimization in transport/storage. *Cybernetics and Systems Analysis*, *28*(3), 484–487.
Melikov, A. Z., Ponomarenko, L. A., & Bagirova, S. A. (2016). Models of queueing-inventory systems with randomized lead policy. *Journal of Automation and Information Sciences*, *48*(9), 23–35.
Melikov, A. Z., Ponomarenko, L. A., & Bagirova, S. A. (2017). Markov models of queueing-inventory systems with variable order size. *Cybernetics and Systems Analysis*, *53*(3), 373–386.
Melikov, A. Z., Rustamov, A. M., & Ponomarenko, L. A. (2017). Approximate analysis of a queueing-inventory system with early and delayed server vacations. *Automation and Remote Control*, *78*(11), 1991–2003.
Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: An algorithmic approach*. Baltimore: John Hopkins Press.
Padmavathi, I., Lawrence, A. S., & Sivakumar, B. (2016). A finite-source inventory system with postponed demands and modified M vacation policy. *Opsearch*, *53*(1), 41–62.
Saffari, M., Asmussen, S., & Haji, R. (2013). The M/M/1 queue with inventory, lost sale, and general lead times. *Queueing Systems*, *75*(1), 65–77.
Saffari, M., Haji, R., & Hassanzadeh, F. (2011). A queueing system with inventory and mixed exponentially distributed lead times. *International Journal of Advanced Manufacturing Technology*, *53*(9), 1231–1237.
Schwarz, M., Sauer, C., Daduna, H., Kulik, R., & Szekli, R. (2006). M/M/1 queueing systems with inventory. *Queueing Systems*, *54*(1), 55–78.
Sigman, K., & Simchi-Levi, D. (1992). Light traffic heuristic for an M/G/1 queue with limited inventory. *Annals of Operations Research*, *40*(1), 371–380.
Sivakumar, B. (2011). An inventory system with retrial demands and multiple server vacation. *Quality Technology and Quantitative Management*, *8*(2), 125–146.
Takagi, H. (1991). *Queueing analysis—A foundation of performance evaluation* (Vol. 1). Amsterdam: Elsevier.
Tian, N., & Zhang, Z. G. (2006). *Vacation queuing models: Theory and applications*. New York: Springer.
Viswanath, C. N., Deepak, T. G., Krishnamoorthy, A., & Krishkumar, B. (2008). On (s, S) inventory policy with service time, vacation to server and correlated lead time. *Quality Technology and Quantitative Management*, *5*(2), 129–144.
Yue, D., Zhao, G., & Qin, Y. (2018). An M/M/1 queueing-inventory system with geometric batch demands and lost sales. *Journal of System Science and Complexity*, *4*, 1–18.