MANAGEMENT SCIENCE
Vol. 56, No. 1, January 2010, pp. 110–124
ISSN 0025-1909 | EISSN 1526-5501 | 10 | 5601 | 0110

informs

DOI 10.1287/mnsc.1090.1099

Global Dual Sourcing: Tailored Base-Surge Allocation to Near- and Offshore Production

Gad Allon, Jan A. Van Mieghem
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208
{g-allon@kellogg.northwestern.edu, vanmieghem@kellogg.northwestern.edu}

When designing a sourcing strategy in practice, a key task is to determine the average order rates placed to each source because that affects cost and supplier management. We consider a firm that has access to a responsive nearshore source (e.g., Mexico) and a low-cost offshore source (e.g., China). The firm must determine an inventory sourcing policy to satisfy random demand over time. Unfortunately, the optimal policy is too complex to allow a direct answer to our key question. Therefore, we analyze a tailored base-surge (TBS) sourcing policy that is simple, used in practice, and captures the classic trade-off between cost and responsiveness. The TBS policy combines push and pull controls by replenishing at a constant rate from the offshore source and producing at the nearshore plant only when inventory is below a target. The constant base allocation allows the offshore facility to focus on cost efficiency, whereas the nearshore facility's quick response capability is utilized only dynamically to guarantee high service. The research goals are to (i) determine the allocation of random demand into base and surge capacity, (ii) estimate corresponding working capital requirements, and (iii) identify and value the key drivers of dual sourcing. We present performance bounds on the optimal cost and prove that economic optimization brings the system into heavy traffic. We analyze the sourcing policy that is asymptotically optimal for high-volume systems and present a simple "square-root" formula that is insightful to answer our questions and sufficiently accurate for practice, as is demonstrated with a validation study.

Key words: inventory production; stochastic models; applications; probability; stochastic model applications
History: Received September 25, 2008; accepted August 21, 2009, by Paul H. Zipkin, operations and supply chain management. Published online in *Articles in Advance* November 6, 2009.

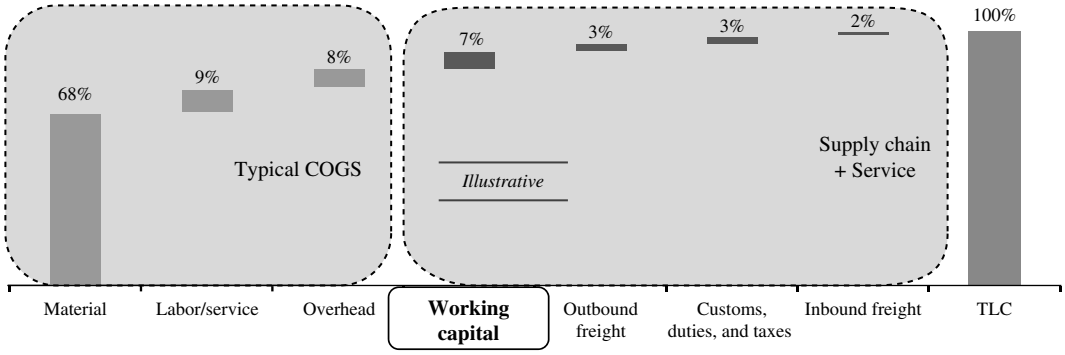
1. Introduction and Summary

A $10 billion high-tech U.S. manufacturer of wireless transmission components was at a crossroads regarding its global network.[1] The company had two assembly plants, one in China and another in Mexico. Although the Chinese facility enjoyed lower costs, ocean transportation made its order lead times 5 to 10 times as long as those from Mexico. With highly uncertain product demand—coefficients of variations of monthly demand for some products were as high as 1.25—sole sourcing was unattractive: Mexico was too expensive and China too unresponsive. The firm had to decide how it could best utilize these two sources by properly allocating product demand to them. In practice, specifying supply allocations is a key task of any sourcing strategy—be it global or domestic—because it affects costs and supplier management. Although also relevant to domestic sourcing, the policy studied in this paper is most naturally applied and interpreted in a global setting during a single-season planning horizon when supply and demand volatility dominate currency exchange risk considerations. In this paper, we will refer to the average order rates as strategic allocation.

The manufacturer retained a management consultant company for advice. Their analysis focused on computing the total landed cost as a function of the allocation to China. The total landed cost represents the end-to-end cost to transform inputs at the source to outputs at the destination (Van Mieghem 2008, p. 208). It captures not only the traditional cost of goods sold (material, labor, and overhead, shown in Figure 1), but also accounts for supply chain costs such as transportation, customs, duties, and taxes, as well as required working capital carrying costs. We will refer to all but working capital cost components as the "sourcing cost." Computing the sourcing cost is tedious yet straightforward. In contrast, working capital greatly depends on lead times (which determine pipeline inventory), volatility and service levels (which determine safety stock). Whereas working capital is easily estimated for single sourcing using readily available standard inventory formulae, there are no such formulae for dual sourcing because the required inventory not only depends on the allocation to both sources but also on the replenishment policy. Therefore, as part of their analysis, the management

[1] The sourcing strategy that motivated this paper is further described in Mini-Case 6 in Van Mieghem (2008).

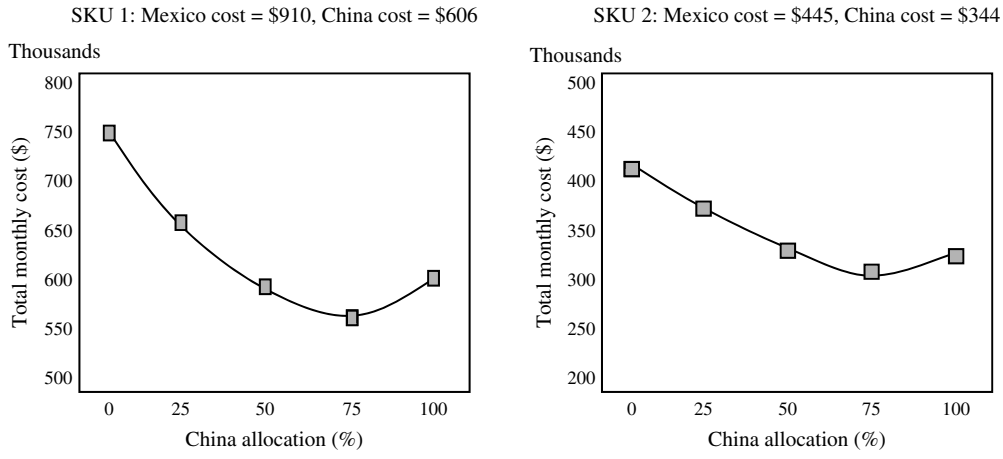
Figure 1 Total Landed Cost Is the Cost to Transform Inputs at the Source to Outputs at the Destination



consultants resorted to an extensive simulation study of the total landed cost.

The simulation study captured a variety of product parameters as well as five distinct China allocations (0%, 25%, 50%, 75%, and 100%) using the firm's replenishment policy, which we will refer to as a tailored base-surge (TBS) sourcing policy. This policy combines push and pull controls by replenishing at a constant rate from China (push), yet ordering from Mexico only when inventory is below a target (pull). Policies that assume a constant order rate are also known as "standing order policies" and have been used for decades (Rosenshine and Obee 1976, Janssen and De Kok 1999). The presumption is that the low-cost source cannot rapidly change volumes because of frictions such as long lead times or an inflexible level production process, which is essential to achieve this cost advantage. The benefits of this policy are that it is simple to administer and it eliminates the need to explicitly account for the long lead time. In addition, the policy aligns the order dynamics with each source's competencies: The slow source replenishes "base" demand, and the fast source reacts to the remaining "surge" demand. As such, the TBS policy captures the classic trade-off between cost and responsiveness: The constant base allocation allows China to operate under level production, and thereby focus on cost efficiency, whereas Mexico's quick response is utilized only dynamically to guarantee high service. The consultants' simulation (Figure 2 shows two representative results) indicated that the total cost was convex and, for the majority of parameter values, minimal when around 75% (i.e., more than 50% but less than 100% given that only five allocations were investigated) was sourced from China. The objective of this paper is to present an analytic model and formulae to predict the optimal allocation, understand its drivers, and tailor the sourcing strategy to the demand and supply characteristics. In our numerical study, we validate the robustness of the "three-quarter" allocation rule of thumb as a good starting point during strategic planning. The studied model applies to the dual-sourcing setting where (1) the lower-cost supplier has a sufficiently long lead time (making a standing order a reasonable alternative to dynamically changing orders), (2) the more expensive supplier has a short transportation lead time (relative to his order fulfillment/production lags),

Figure 2 Consultants' Simulated Total Landed Cost Was Minimal When Allocating More Than 50% But Less Than 100% to China



and (3) there is a single-season planning period during which demand is reasonably stationary. A global offshore/nearshore dual-sourcing setting is a natural example, provided demand and supply risks dominate currency exchange risks.

We consider a model of a single-stage inventory system that replenishes from two supply sources using a TBS policy. The demand and supply processes can be general, correlated stationary stochastic processes. Even our simple TBS policy is not amenable to exact analysis. There are two options to proceed: (1) solve the exact problem numerically or via simulation, or (2) solve an approximate problem analytically. Given that we seek simple formulae to determine the allocation and its key determinants, we develop a Brownian analytic model that is asymptotically optimal for high sourcing volumes. Analytic optimization of the Brownian model provides us with an analytic prescription for the sourcing allocations, the base-stock level, and its corresponding cost.

Our main results can be summarized as follows.

(1) We present performance bounds on the optimal cost and prove that economic optimization brings the system into the so-called heavy-traffic regime. We provide an analytic characterization of the asymptotically optimal TBS dual-sourcing policy, including its strategic allocation, base-stock level, and expected cost, as well as an analytic expression for the corresponding "overshoot" process. In addition, we present a simple square-root formula to predict the near-optimal allocation and cost.

(2) The analytic characterizations, including the simple square-root formula, capture and quantify the classic trade-off between cost and responsiveness. They highlight the key drivers of the dual sourcing allocations: (i) the monetary ratio of the China cost advantage to the unit holding cost; (ii) average demand rate; (iii) the volatility of demand and China supply; and (iv) demand–supply correlations as well as serial time correlations. Our results not only confirm intuition but also provide new insight and permit easy quantification of the allocation and corresponding cost. For example, an increase in the monetary ratio (either because of a larger China cost advantage or a smaller holding cost) results, as expected, in a larger China allocation. Our formulae predict that this relationship is nonlinear and follows a square root. Similarly, an increase in demand volatility decreases the China allocation. Intuitively, this reduces the base demand while increasing the surge demand. Our formula quantifies what constitutes "base demand," thereby providing the scientific underpinnings of the principle of strategic alignment when applied to dual sourcing. We also quantify and investigate the value of dual sourcing over single sourcing.

(3) A numerical study shows that our analytic characterization and the simple square-root formula provide sufficiently accurate prescriptions relative to simulation-based optimization of the TBS policy as well as more complex policies. This study initially assumes parameter values traditionally used in the literature but then continues with applying the model to real data from the motivating example. During the latter, we discuss how to calibrate model parameters in practice and validate the robustness of the "three-quarter" allocation rule of thumb. This suggests that our results are readily applicable.

The remainder of this paper is structured as follows. The next section provides a review of the relevant literature and is followed by a discussion of the model. Section 4 specifies inventory dynamics under the TBS policy, presents bounds on the optimal cost, and proves that optimization brings the system into heavy traffic. Section 5 analyzes the Brownian model and the asymptotic performance of the TBS policy. Section 6 identifies key drivers and the value of dual sourcing. Section 7 discusses the impact of demand–supply correlations as well as serial time correlations. Section 8 reports the numerical validation study. Section 9 provides a conclusion and discussion of limitations. All proofs are relegated to the online appendix (provided in the e-companion).[2]

2. Literature Review

The dual-sourcing literature dates back to Barankin (1961), who studied a single-period model with emergency orders. The literature distinguishes between single- and dual-index policies, depending on whether one or two inventory positions are tracked, and, somewhat independently, between single- and dual-base-stock policies, depending on the number of order-up-to levels used by the policy.[3] Our TBS policy is a single-index, single-base-stock policy. The dual-sourcing literature can also be divided into discrete and continuous review models.

Discrete review models include Fukuda (1964) who studies a dynamic inventory model with stochastic demand in which the deterministic lead times of both sources differ by exactly one period. He shows that single-index, dual-base-stock policies are optimal

[2] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

[3] A single-index policy uses one state variable, usually the total inventory position I. A single-base-stock policy uses one parameter and usually brings the inventory position after ordering as close to the base stock level s as possible (Porteus 2002, p. 67). A single-index, dual-base-stock policy has two parameters $s < s_0$. As with a TBS policy, the fast source shuts off when $I > s$. In addition, the slow source also shuts off when $I > s_0$. A dual-index policy tracks two state variables, e.g., inventory position of the emergency supplier and the total inventory position.

under mild conditions. Whittemore and Saunders (1977) extend Fukuda's (1964) model to allow for arbitrary (yet still deterministic) lead times. They show that when lead times differ by more than one period, the optimal policy is no longer a simple function of one or two inventory positions, but depends on the entire ordering history. The model in Rosenshine and Obee (1976) assumes a regular lead time, but immediate emergency replenishment. Their standing order policy, which was evaluated numerically, assumes a constant order rate from the regular source, a feature shared by our TBS policy. Tagaras and Vlachos (2001) allow emergency replenishment within the regular review period. Veeraraghavan and Scheller-Wolf (2008) introduce a dual-index policy for capacitated dual-source models that can be computed using a simple simulation-based optimization procedure. They show that such a dual-index policy is nearly optimal when compared to state-dependent policies found via multidimensional dynamic programming. Scheller-Wolf et al. (2006) establish computationally that a single-index policy can be highly effective, and even outperform a dual-index policy. Sheopuri et al. (2007) generalize the dual-index policy by considering two classes of policies that have an order-up-to structure for the emergency supplier. The authors show that the "lost sales inventory problem" is a special case of the dual-sourcing problem. They use this property to suggest near-optimal policies within this class that often improve on the already excellent performance of dual-index policies.[4] One of their policy classes uses a single order-up-to level throughout and then allocates, in each period, demand to each supplier. The idea of determining the allocation is similar in spirit to our approach. However, to address our research question of strategic allocation, we first determine the average allocation throughout. This average allocation then determines a single order-up-to level that specifies dynamically when to source from the fast supplier. In addition, the papers above consider deterministic lead times; in contrast, one of the goals of our model is to explore the relationship between the optimal strategic allocation to each source and the volatility of the supply sources.

Continuous review models include Moinzadeh and Nahmias (1988), who consider two sources with deterministic lead times and fixed order costs. They extend the (Q, R) policy to two different lot sizes and two different reorder levels, and optimize over these four parameters. Assuming negligible fixed order costs, Moinzadeh and Schmidt (1991) consider a more sophisticated dual-base-stock policy in which real-time supply information on the age of all outstanding orders and the inventory level is used. Song and Zipkin (2009) extend Moinzadeh and Schmidt (1991) by considering a system with multiple supply sources under stochastic demand and lead times. The authors develop performance evaluation tools for a family of policies that utilize real-time supply information and under which the supply system becomes a network of queues with a routing mechanism called an overflow bypass. Bradley (2004), which is the closest to our model and inspired our analysis, considers a production–inventory problem where the inventory can be replenished from in-house production or through a subcontractor. The author constructs a Brownian approximation of the optimal control problem, assuming that the manufacturer uses a single-index, dual-base-stock policy. By using only a single base stock, our replenishment policy is simpler and provides greater tractability. This allows us to specify and investigate the optimal allocations explicitly. Zipkin (2000) highlights the connection between inventory and queuing theory, and argues that "queuing theory remains our richest source of models for supply processes" (p. 13).

The aforementioned papers focus on determining or optimizing the control parameters, or on evaluating the performance of dual-sourcing policies. Although we also derive the optimal base-stock level of the TBS policy, our focus is on determining the optimal allocation of demand to either source. The latter is also the focus of the literature on "order splitting," which studies inventory models with deterministic demand. Lau and Zhao's (1994) paper belongs to this stream, and studies a system with stochastic lead times and explores the impact of splitting rules on inventory costs and stockout risks.

In addition to dual-source inventory models, our work is also related to the literature on inventory models with returns. Under a TBS policy, the net demand experienced by the firm after subtracting the base-demand replenishment can be negative. Inventory models with returns, as studied by Fleischmann et al. (2002) and DeCroix et al. (2005), are characterized by the same feature. Fleischmann et al. (2002) studied a Markovian model with fixed cost. The behavior of the inventory cost as a function of the return ratio is closely related to the behavior of the total cost in our model as a function of the allocation to China. DeCroix et al. (2005) studied a more general serial system with returns and show that an echelon base-stock policy is optimal.

Our model explores the cost-responsiveness tradeoff when allocating supply to a responsive yet expensive source, and a low-cost but remote source within

[4] Dual-index policies and their generalizations require the stationary distribution of an "overshoot" process, which typically is obtained through simulation. We provide an analytic expression of an overshoot distribution that may be useful in the computation of the former policies.

an *existing* network. It does not explore financial hedging of currency exchange rate risk or configuring global networks. For such models, we refer the reader to Ding et al. (2007) and Lu and Van Mieghem (2009) and references therein.

3. Model

Consider a continuous-time model of a single-stage inventory system with two supply sources. The cumulative demand up to time t is a stationary stochastic process $D(t)$; demand in excess of available inventory is backlogged. Initially, D, is modeled as a counting (renewal) process whose independent and identically distributed (i.i.d.) interarrival times have mean $1/\lambda$ and coefficient of variation v_D; later we will generalize to allow for correlated interarrival times. Similarly, to model production variability as well as congestion and disruption, the actual supply from either source is stochastic around its mean rate. To be precise, let $S_i(t)$ denote the cumulative quantity received from source i if it were continuously supplying during $[0, t]$. To align with our motivating example, we will use $i \in \{M, C\}$ for Mexico and China as concrete placeholders for the nearshore and offshore sources, respectively. Initially, we assume that $S_i(t)$ is a renewal process whose associated i.i.d. service times have mean $1/\mu_i$ and coefficient of variation v_i; later we will generalize to allow for correlated intersupply as well as for cross-correlations between intersupply and interdemand times. We shall refer to μ_i as the capacity of source i, which is a decision variable and incurs a capacity cost of $c_i^K \mu_i$ per unit of time.

The variable order cost from source i is c_i^V per unit ordered. As discussed in the Introduction, this sourcing cost includes all components of the total landed cost with the exception of working capital (i.e., inventory) cost. Source M is responsive but expensive, whereas source C is cheap but slow. M is more expensive both on a variable-cost basis ($c_M^V > c_C^V$) as well on a full-cost basis: $c_M = c_M^K + c_M^V > c_C = c_C^K + c_C^V$.

Let the control $T_i(t)$ denote the actual cumulative amount of time that source i is supplying during $[0, t]$ so that $S_i(T_i(t))$ is the actual supply from source i during $[0, t]$. Let $I(t)$ denote the net-inventory process, i.e., the amount of inventory on hand minus the amount on backorder at time t. We then have the following dynamics:

$$I(t) = I(0) + S_C(T_C(t)) + S_M(T_M(t)) - D(t).$$

Let $I(\infty)$ denote the steady-state net-inventory process for a given control policy T. On-hand inventory I^+ incurs the familiar per-unit holding cost h per unit of time.[5] Stockouts are backlogged, and backorders I^- incur a per-unit backlogging penalty cost b per unit of time. In the usual way, the average inventory (or demand–supply mismatch) cost rate under this policy is $G = \mathbb{E}g(I(\infty))$, where $g(x) = hx^+ + bx^- = hx + (b + h)x^-$. Let ζ denote the critical fractile $b/(b + h)$ and $\bar{\zeta} = 1 - \zeta$.

The research question is to determine the capacity vector μ and the allocation policy T that minimizes total cost C, the sum of capacity, inventory, and sourcing costs. We seek simple characterizations of how the sourcing volume λ should be allocated to the two sources. In other words, we want to characterize the "base demand" that should be allocated to China, and when tailored dual sourcing outperforms single sourcing.

Addressing these questions involves determining the optimal dynamic order policy, which is generally complex and not amenable to exact analysis. Therefore, in what follows, we first restrict attention to a particular allocation policy (the TBS policy) for which we provide some general results. To further quantify its performance, we then provide an analytic characterization using a Brownian model of the TBS policy that is asymptotically correct for high volume ($\lambda \to \infty$). In a third step, we present a simple square-root formula that is a lower bound of the predicted optimal allocation in the Brownian model. Finally, our numerical study validates the accuracy of our approximation.

4. The Tailored Base-Surge Policy

The simplest tailored allocation policy orders a constant rate from the offshore source and orders only occasionally when needed from the nearshore source. Specifically, China supplies at a constant rate μ_C; clearly, $0 \le \mu_C < \lambda$ to prevent unlimited inventory buildup. In contrast, the policy orders from Mexico only when the net inventory falls below a target level s. During that time, supply from Mexico is received at rate μ_M. Obviously, $\mu_C + \mu_M > \lambda$ to keep up with demand.

As stated in the Introduction, a TBS policy is used in practice because it is simple to administer and it allows the efficient source to operate under level production. It also is amenable to analysis and, hence, simple to tailor to particular demand–supply characteristics. The underlying assumption of the TBS policy is that the offshore source is not capable of implementing feedback control because of various

[5] Here, h is the average unit holding cost rate. Under the policy analyzed in this paper, its opportunity cost component is a weighted average $r((1-\rho_C)c_M + \rho_C c_C)$, where ρ_C denotes the fraction sourced from China and r is the cost of capital. We shall see that ρ_C^* is close to 1, so that the opportunity holding cost $\simeq rc_C$, and we assume h is constant. Incorporating the dependence of h on the allocation does not impact our main asymptotic results, but it does significantly complicate exposition.

frictions such as long transportation times or inflexible production.

Flow balance dictates that the long-run average supply from Mexico is $\lambda - \mu_C$, and the long-run average sourcing cost rate is $c_M^V\lambda - \mu_C\Delta c^V$, where $\Delta c^V = c_M^V - c_C^V > 0$. The difference $\mu_M - (\lambda - \mu_C) > 0$ between the Mexican capacity and its long-run average supply rate is called the Mexican safety capacity, which is positive. In contrast, the constant order obviates the need for China safety capacity, and its average supply rate equals its capacity. Observe that, by design, the replenishment lead time from the slow source does not impact the TBS policy. However, we can easily account for any pipeline inventory holding costs.[6]

Under the TBS policy, continuous supply from China implies that the control $T_C(t) = t$ so that the model simplifies to a single-source inventory model with remaining demand $D(t) - S_C(t)$, which can be negative. This is mathematically equivalent to an inventory system with returns, and it is well established that a base-stock policy is optimal. In the typical base-stock dynamics, once inventory falls below s (after a potential transient initial regime), the inventory position stays at or below s and is a demand-replacement policy. This is not the case under a TBS policy because the slow source supply may occasionally exceed the actual demand, resulting in excess inventory excursions above s. A similar "overshoot" phenomenon is observed in the dual-index policies of Veeraraghavan and Scheller-Wolf (2008) and the generalizations by Sheopuri et al. (2007). This overshoot is a key disadvantage that is not present in Bradley's (2004) dual-base-stock policy.

We will adopt a continuous-review base-stock policy that requests supply at rate μ_M from the fast source whenever the net inventory falls below s. Let $Z = I - s$ denote the "excess inventory process," which is the inventory above the base stock. Under a TBS policy, the excess inventory dynamics simplify to

$$Z(t) = Z(0) + S_M(T_M(t)) + S_C(t) - D(t),$$

where

$$T_M(t) = \int_0^t 1\{Z(u) < 0\}\,du.$$

Essentially, Z is a random walk stemming from the conventional order-up inventory dynamics with a superimposed $GI/G/1$ queue capturing the occasional excess inventory excursions. For a given capacity vector μ, let F_μ denote the stationary distribution of Z (we will show that such limiting distribution does exist).

[6] Let L_M and L_C denote the average transportation times from Mexico and China, respectively, so that the associated in-transit holding costs is $L_M(\lambda - \mu_C)h + L_C\mu_C h$. The allocation only affects the terms in μ_C and that effect can be captured by inflating c_M^V by $h\Delta L$, where $\Delta L = L_C - L_M > 0$.

The benefit of analyzing the excess inventory process Z is that it is independent of the actual value of the base stock s.

The average steady-state total cost rate under a TBS policy with base stock s and capacity vector μ is the sum of inventory, sourcing, and capacity costs:

$$\begin{aligned} C(\mu_C, \mu_M, s) &= \mathbb{E}g(Z(\infty)+s) + c_C^V\mu_C + c_M^V\mu_M\mathbb{P}(Z(\infty) < 0) \\ &\quad + c_C^K\mu_C + c_M^K\mu_M \\ &= G(\mu, s) + c_M^V\lambda + (c_C^K - \Delta c^V)\mu_C + c_M^K\mu_M, \end{aligned}$$

given that a stationary solution requires stability so that $\mu_M\mathbb{P}(Z(\infty) < 0) = \lambda - \mu_C$. The inventory cost $G = \mathbb{E}g(Z(\infty)+s) = h\mathbb{E}(Z(\infty)+s) + (b+h)\mathbb{E}(Z(\infty)+s)^-$ and integration by parts of the last term yields

$$G(\mu, s) = hs + h\int_{-\infty}^{+\infty} x\,dF_\mu(x) + (b+h)\int_{-\infty}^{-s} F_\mu(x)\,dx.$$

Proposition 1. *The inventory cost $G(\mu, s)$ is convex in s for any μ, and the optimal base stock s^* is a fractile of the steady-state excess inventory distribution: If F_μ is continuous, then $F_\mu(-s^*) = \bar{\xi}$.*

This type of newsvendor solution has appeared in previous analyses of inventory shortfall as discussed by Bradley and Glynn (2002). To optimize the total cost, it "only" remains to specify the stationary distribution F_μ. Given that our system involves $GI/G/1$ queue dynamics, its stationary distribution cannot be solved analytically in general. We can, however, obtain a useful upper bound on the optimal cost as follows. Observe that the optimal dual-sourcing cost dominates the minimal cost under single sourcing from Mexico with $s = 0$: $\min C(\mu_C, \mu_M, s) \le \min C(0, \mu_M, 0)$. Under such single sourcing, $Z(\infty) \le 0$, and the backlog $-Z$ is a $GI/G/1$ queue so that

$$\begin{aligned} C(0, \mu_M, 0) &= -b\mathbb{E}Z(\infty) + \mu_M c_M^K + c_M^V\lambda \\ &\le b\frac{\lambda}{\mu_M - \lambda}\frac{v_M^2 + v_D^2}{2} + \mu_M c_M^K + c_M^V\lambda, \end{aligned}$$

using Kingman's bound. The right-hand side is convex in μ_M and reaches a minimum at $\mu_M = \lambda + \sqrt{(b/c_M^K)((v_M^2 + v_D^2)/2)\lambda}$, which yields an exact upper bound: $\min C(0, \mu_M, 0) \le \bar{C}^\lambda$, where

$$\bar{C}^\lambda = (c_M^V + c_M^K)\lambda + \sqrt{2bc_M^K(v_M^2 + v_D^2)\lambda}.$$

The upper bound also bounds the inventory and the capacity cost. This directly shows how the optimal inventory and capacities depend on the volume λ, which is key to our analysis. To emphasize this dependence, we will add a superscript λ to the notation. For example, C^λ denotes the total cost given volume λ, and $(\mu_C^{\lambda*}, \mu_M^{\lambda*}, s^{\lambda*})$ denotes an optimal solution.

Proposition 2. *The optimal cost* $C^\lambda(\mu_C^{\lambda*}, \mu_M^{\lambda*}, s^{\lambda*})$ *is bounded,*

$$(c_C^V + c_C^K)\lambda + \sqrt{2c_C^K h(v_C^2 + v_D^2)\ln\bar{\zeta}^{-1}}\sqrt{\lambda} + o(\sqrt{\lambda}) \le C^\lambda(\mu_C^{\lambda*}, \mu_M^{\lambda*}, s^{\lambda*}) \le \bar{C}^\lambda,$$

and there exist nonnegative scalars $\hat{\mu}_M$, $\hat{\mu}_C$, *and* $\hat{s}$ *such that the optimal solution satisfy*

$$\mu_C^{\lambda*} = \lambda - \hat{\mu}_C\sqrt{\lambda} + o(\sqrt{\lambda}), \tag{1}$$

$$\mu_M^{\lambda*} = \hat{\mu}_M\sqrt{\lambda} + o(\sqrt{\lambda}), \tag{2}$$

$$s^{\lambda*} = \hat{s}\sqrt{\lambda} + o(\sqrt{\lambda}). \tag{3}$$

Proposition 2 has two important implications. First, the optimal allocation sources the majority from the cheap source, but a small amount of Mexican capacity is necessary. Second, economic optimization naturally brings the system into a parameter regime called "heavy traffic." Loosely speaking, this means that the China resource is heavily utilized. Indeed, expression (1) implies that the optimal China utilization $\mu_C^{\lambda*}/\lambda \simeq 1 - \hat{\mu}_C/\sqrt{\lambda}$ tends to 100% as $\lambda \to \infty$. In addition, the optimal Mexican capacity, although small, is just sufficient to stabilize the inventory process. The theoretical significance of the proposition is that heavy traffic is not assumed, but the proved result of capacity optimization. From a practical perspective, the proposition guarantees that the system converges to a tractable Brownian limiting system as $\lambda \to \infty$.

Proposition 3. *The scaled excess inventory* $Z^\lambda/\sqrt{\lambda}$ *converges almost surely uniformly in compact sets*[7] *to a dual-drift Brownian motion* $\hat{Z}$:

$$\frac{Z^\lambda(t)}{\sqrt{\lambda}} \to \hat{Z}(t) = \hat{Z}(0) + \hat{\mu}_M\hat{T}(t) - \hat{\mu}_C t + \sigma\hat{B}(t) \quad a.s.,$$

where $\sigma^2 = v_C^2 + v_D^2$, $\hat{B}$ *is a standard Brownian motion, and* $\hat{T}(t) = \int_0^t 1\{\hat{Z}(u) < 0\}\,du$. *Furthermore, the steady-state limit* $\hat{Z}(\infty)$ *exists if* $\hat{\mu}_M > \hat{\mu}_C$.

The limiting scaled inventory $\hat{Z}$ is a diffusion process with negative drift $-\hat{\mu}_C$ if $\hat{Z} \ge 0$, and positive drift $\hat{\mu}_M - \hat{\mu}_C$ elsewhere. As we will show, the limiting system *is* amenable to analytic optimization and allows us to prescribe a solution for a system with volume λ as follows. Denote the total cost of the Brownian limiting system by

$$\hat{C}(\hat{\mu}_C, \hat{\mu}_M, \hat{s}) = \mathbb{E}g(\hat{Z}(\infty) + \hat{s}) + c_M^K\hat{\mu}_M + (c_M^V - c_C^K - c_C^V)\hat{\mu}_C, \tag{4}$$

and let $(\hat{\mu}_C^*, \hat{\mu}_M^*, \hat{s}^*)$ denote a minimizer of $\hat{C}$. We can now state our prescribed solution.

[7] This means that, with probability 1, for every compact set $A \subset \mathbb{R}_+$, $\lim_{\lambda\to\infty}\sup_{t\in A}\|Z^\lambda(t)/\sqrt{\lambda} - \hat{Z}(t)\| = 0$.

Proposition 4. *The prescription* $(\lambda - \hat{\mu}_C^*\sqrt{\lambda}, \hat{\mu}_M^*\sqrt{\lambda}, \hat{s}^*\sqrt{\lambda})$ *is asymptotically optimal*:

$$\lim_{\lambda\to\infty}\frac{C^\lambda(\lambda - \hat{\mu}_C^*\sqrt{\lambda}, \hat{\mu}_M^*\sqrt{\lambda}, \hat{s}^*\sqrt{\lambda}) - (c_C^V + c_C^K)\lambda}{\sqrt{\lambda}} = \min\hat{C}(\hat{\mu}_C, \hat{\mu}_M, \hat{s}).$$

The solution that we prescribe for a system with volume λ is based on the optimal solution of the Brownian limiting model, and thus is guaranteed to perform well for large volumes. In the remainder, we analyze the asymptotic cost $\hat{C}$, characterize its optimal solution $(\hat{\mu}_C^*, \hat{\mu}_M^*, \hat{s}^*)$, and validate the prescription for various volume levels.

5. Asymptotic Analysis of the TBS Policy

5.1. Steady-State Distribution $\hat{F}_{\hat{\mu}}$

The steady-state distribution of $\hat{Z}(\infty)$ follows directly from Browne and Whitt (1995):

Proposition 5. *The steady-state limit* $\hat{Z}(\infty)$ *has distribution function*

$$\hat{F}_{\hat{\mu}}(x) = \begin{cases} \dfrac{\hat{\mu}_C}{\hat{\mu}_M}\exp\left(\dfrac{2(\hat{\mu}_M - \hat{\mu}_C)}{\sigma^2}x\right) & x < 0, \\ 1 - \dfrac{(\hat{\mu}_M - \hat{\mu}_C)}{\hat{\mu}_M}\exp\left(-\dfrac{2\hat{\mu}_C}{\sigma^2}x\right) & x \ge 0, \end{cases} \tag{5}$$

which decreases as $-\hat{\mu}_C$ *or* $\hat{\mu}_M$ *increases.*

For Markovian systems (demand follows a Poisson process with rate λ, service times from source i are i.i.d. exponentially distributed with rate μ_i, and all service and demand times are independent), we can calculate the exact steady-state distribution of Z^λ using the detailed balance equations

$$\lim_{t\to\infty}\mathbb{P}(Z^\lambda(t) \le x) = \begin{cases} \dfrac{\pi_0}{1-\alpha}\alpha^{-\lfloor x\rfloor} & \text{for } x < 0, \\ 1 - \dfrac{\pi_0}{1-\beta}\beta^{\lfloor x\rfloor+1} & \text{for } x \ge 0, \end{cases}$$

where

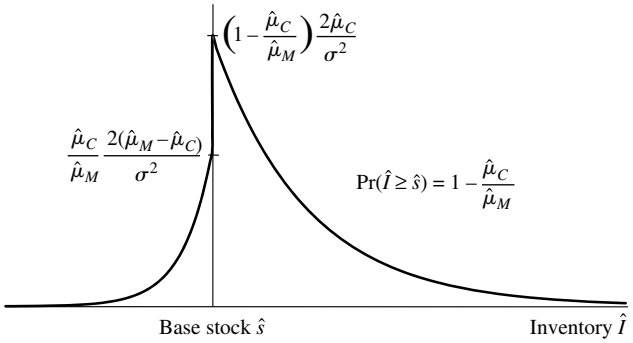
$$\alpha = \frac{\lambda}{\mu_M + \mu_C}, \quad \beta = \frac{\mu_C}{\lambda},$$
$$\pi_0 = \frac{(\lambda - \mu_C)(\mu_M + \mu_C - \lambda)}{\lambda\mu_M}.$$

Using the optimal capacities (1) and (2) and the fact that $\lim_{\lambda\to\infty}(1 + x/\lambda)^\lambda = e^x$, it is easy to verify that indeed

$$\lim_{\lambda\to\infty}\lim_{t\to\infty}\mathbb{P}\left(\frac{Z^\lambda(t)}{\sqrt{\lambda}} \le x\right) = \hat{F}_{\hat{\mu}}(x) = \lim_{t\to\infty}\lim_{\lambda\to\infty}\mathbb{P}\left(\frac{Z^\lambda(t)}{\sqrt{\lambda}} \le x\right).$$

The scaled excess inventory has a biexponential density, as shown in Figure 3. As expected, higher

Figure 3 **Stationary Scaled Inventory $\hat{I} = \hat{Z}(\infty) + \hat{s}$ Has a Biexponential Distribution**



supply rates μ (hence, smaller $\hat{\mu}_C$ or larger $\hat{\mu}_M$) stochastically increase inventory, and thus the average stationary inventory:

$$\mathbb{E}\hat{Z}(\infty) = -\frac{\sigma^2}{2(\hat{\mu}_M - \hat{\mu}_C)}\frac{\hat{\mu}_C}{\hat{\mu}_M} + \frac{\sigma^2}{2\hat{\mu}_C}\frac{\hat{\mu}_M - \hat{\mu}_C}{\hat{\mu}_M}.$$

Notice that the proposition provides an analytic characterization of the overshoot process that may be useful in the computation of dual-index policies and their generalizations. It certainly allows us to compute and optimize the holding cost of the TBS policy.

5.2. Optimal Scaled Base Stock $\hat{s}^*$ and Associated Inventory Cost $G(\hat{\mu}, \hat{s}^*)$

The explicit characterization (5) allows the specification of the optimal base stock $\hat{s}^* = -\hat{F}_{\hat{\mu}}^{-1}(\bar{\zeta})$ and of the associated inventory cost for fixed capacities. Given the specific biexponential structure of $\hat{F}$, we distinguish two operating regimes: $\hat{s}^* \geq 0$ versus $\hat{s}^* < 0$. Following Bradley (2004), we say that the control policy is preventive if the fast Mexican source is engaged while inventory is on hand ($\hat{s}^* \geq 0$), and reactive when Mexico supplies only backorders ($\hat{s}^* < 0$).

Proposition 6. *Consider fixed capacities $\hat{\mu}_M$ and $\hat{\mu}_C$. If $\bar{\zeta} = h/(h+b) \leq \hat{\mu}_C/\hat{\mu}_M$, then the optimal base stock $\hat{s}^*$ is positive so that the fast source supplies to stock ("preventive mode") with*

$$\hat{s}^* = -\frac{\sigma^2}{2(\hat{\mu}_M - \hat{\mu}_C)}\ln\frac{\hat{\mu}_M}{\hat{\mu}_C}\bar{\zeta} \geq 0 \quad \textit{and}$$

$$G(\hat{\mu}, \hat{s}^*) = h\hat{s}^* + h\frac{\sigma^2}{2\hat{\mu}_C} \geq 0.$$

Otherwise, $\hat{s}^* < 0$, and the fast source engages only to cover backlog ("reactive mode") with

$$\hat{s}^* = \frac{\sigma^2}{2\hat{\mu}_C}\ln\frac{\hat{\mu}_M}{\hat{\mu}_M - \hat{\mu}_C}\zeta < 0 \quad \textit{and}$$

$$G(\hat{\mu}, \hat{s}^*) = -b\hat{s}^* + b\frac{\sigma^2}{2(\hat{\mu}_M - \hat{\mu}_C)} \geq 0.$$

As expected, the optimal base stock $\hat{s}^*$ is decreasing in $\hat{\mu}_M$ and $-\hat{\mu}_C$. In addition, for a constant allocation, the absolute value of the optimal base stock and the total cost are increasing in volatility σ. Operating in preventive (reactive) mode is similar to operating the Mexico source in a make-to-stock (make-to-order) fashion. The optimal regime is preventive when relative holding costs h/b and the contingent supply μ_M are small; otherwise, it is better to move to a make-to-order model in which we operate in reactive mode and use the ample capacity of the fast source to cover backlogs.

5.3. Optimal Scaled Capacities $\hat{\mu}^*$ and Cost $\hat{C}^*$

We can now characterize the optimal capacities that minimize the total asymptotic cost $\hat{C}$:

Proposition 7. *The inventory cost $G(\hat{\mu}, \hat{s}^*)$, and thus the total cost, is strictly convex over the set $0 \leq \hat{\mu}_C \leq \hat{\mu}_M$. The optimal cost is*

$$\hat{C}^* = \hat{C}(\hat{\mu}_C^*, \hat{\mu}_M^*, \hat{s}^*) = 2c_M^K\hat{\mu}_M^* + 2(c_M^V - c_C)\hat{\mu}_C^* \quad (6)$$

where the optimal capacities $(\hat{\mu}_C^*, \hat{\mu}_M^*)$ are the unique solutions to

$$\hat{\mu}_C^* = \sqrt{\frac{h\sigma^2}{2\Delta c} + \alpha^2} - \alpha, \quad \textit{where } \alpha = \frac{1}{\hat{\mu}_M^*}\frac{h\sigma^2}{4\Delta c}, \quad (7)$$

$$\frac{2}{h\sigma^2}c_M^K = \frac{1}{\hat{\mu}_M^*(\hat{\mu}_M^* - \hat{\mu}_C^*)} - \frac{1}{(\hat{\mu}_M^* - \hat{\mu}_C^*)^2}\ln\frac{\hat{\mu}_M^*}{\hat{\mu}_C^*}\bar{\zeta} \quad (8)$$

if $\hat{s}^* > 0$, and otherwise,

$$\hat{\mu}_M^* - \hat{\mu}_C^* = \sqrt{\frac{b\sigma^2}{2c_M^K} + \beta^2} - \beta, \quad \textit{where } \beta = \frac{1}{\hat{\mu}_M^*}\frac{b\sigma^2}{4c_M^K}, \quad (9)$$

$$\frac{2}{b\sigma^2}(c_C - c_M^V) = \frac{1}{\hat{\mu}_C^{*2}}\ln\frac{\hat{\mu}_M^*}{\hat{\mu}_M^* - \hat{\mu}_C^*}\zeta + \frac{2\hat{\mu}_C^* - \hat{\mu}_M^*}{\hat{\mu}_C^*(\hat{\mu}_M^* - \hat{\mu}_C^*)^2}. \quad (10)$$

At the optimal control variables, the inventory cost $G(\hat{\mu}^*, \hat{s}^*)$ equals the sourcing and capacity costs so that the optimal cost $\hat{C}^*$ equals twice the latter. This property is similar to the familiar economic order quantity model and useful to verify whether the controls are close to optimal in numerical work. Indeed, the transcendental first-order equations (7) and (9) are easily solved numerically, yet also suggest the following simple square-root expressions to be used as a starting point for allocations in the preventive and reactive modes:

$$\hat{\mu}_p = \sigma\sqrt{\frac{h}{2\Delta c}} \quad \text{and} \quad \hat{\mu}_r = \sigma\sqrt{\frac{b}{2c_M^K}}. \quad (11)$$

PROPOSITION 8. *The simple square-root formulae provide upper bounds on allocation,*

$$\hat{\mu}_p - \frac{1}{\hat{\mu}_M^*}\frac{h\sigma^2}{4\Delta c} \le \hat{\mu}_C^* \le \hat{\mu}_p \quad \text{if } \hat{s}^* > 0 \text{ (preventive)}, \tag{12}$$

$$\hat{\mu}_r - \frac{1}{\hat{\mu}_M^*}\frac{b\sigma^2}{4c_K^1} \le \hat{\mu}_M^* - \hat{\mu}_C^* \le \hat{\mu}_r \quad \text{otherwise (reactive)}, \tag{13}$$

and on the optimal cost,

$$\hat{C}^* \ge 2\Delta c\hat{\mu}_C^* \ge \sigma\sqrt{2h\Delta c}. \tag{14}$$

Our analysis thus culminates in these simple square-root formulae which provide a useful estimate or starting point for the optimal China allocation and cost in the more likely preventive mode. Indeed, when dual sourcing a given volume λ, our analysis gives the following prescriptions:

$$\text{China allocation} \quad \mu_C^{\lambda*} \simeq \lambda - \hat{\mu}_C^*\sqrt{\lambda} \ge \lambda - \sigma\sqrt{\frac{h\lambda}{2\Delta c}}, \tag{15}$$

$$\text{Total cost} \quad C^{\lambda^*} \simeq c_C\lambda + \hat{C}^*\sqrt{\lambda} \ge c_C\lambda + \sigma\sqrt{2h\lambda\Delta c}, \tag{16}$$

which are accurate up to $o(\sqrt{\lambda})$ and directly identify key drivers that we discuss next.

6. Drivers and Value of Dual Sourcing

6.1. Key Drivers of Strategic Allocation

The square-root formula (15) directly provides the following key drivers, insights, and quantification on strategic allocation. First, the key monetary trade-off in determining the China allocation is $\Delta c/h$, which can be expressed as follows. Recall that the unit holding cost $h = (\text{cost of capital } r + \text{physical holding cost } p)c_C$, so that the key trade-off simplifies to

$$\frac{\Delta c}{h} = \frac{\Delta c_{\text{source}} - h\Delta L}{h} = \frac{\Delta c/c_C}{r+p} - \Delta L$$
$$= \frac{\textit{relative cost advantage}}{\textit{cost of capital} + \textit{physical holding cost}} - \textit{transportation time difference}.$$

Note that this equation is in time units and that it captures the combined impact of monetary cost concerns as well as responsiveness. This is exactly the trade-off at the essence of this model. It shows that the China allocation is high when (i) China has a high relative cost advantage (as expected), (ii) the cost of capital and the physical holding cost are low (meaning small opportunity costs as well as low risk of obsolescence), and (iii) there is a relatively small transportation time difference between China and Mexico. Not only does this confirm intuition, the equation also quantifies the factors and their interaction.

Second, the total China allocation as a fraction of average demand is

$$\frac{\mu_C^{\lambda*}}{\lambda} \simeq 1 - \frac{\hat{\mu}_C^*}{\sqrt{\lambda}} \ge 1 - \sigma\sqrt{\frac{h}{2\lambda\Delta c}}.$$

It strongly depends on product volume, and thus on its stage in the product life cycle: As the volume grows, the China allocation should increase. Later, during the decline phase, that allocation should decrease, thereby reflecting a shift in the relative importance from cost to responsiveness.

Third, the allocation depends mostly on the China supply volatility and the demand volatility. Our approximation depends equally on both, but is rather insensitive to the Mexico supply volatility.[8] As expected, as China becomes a less reliable source, its allocation is reduced. Interestingly, as the demand volatility increases, the allocation to China is reduced as well. Both effects reflect the fact that China is the less flexible source.

The combined impact of these three key drivers is summarized through the ratio

$$\frac{h((v_C^2 + v_D^2)/2)}{2\lambda\Delta c} = \frac{\textit{holding cost of safety stock}}{\textit{sourcing cost savings}},$$

which captures the natural trade-off in dual sourcing and quantifies it: as the ratio increases, the China allocation reduces.

6.2. Cost and Value of Dual Sourcing

The square-root formula (16) provides similar insights on the cost of dual sourcing. Given that the majority is sourced from China, the first-order component in the total cost is simply the China cost rate $c_C\lambda$. With proper sourcing management, the cost of safety capacity and safety stock is of second order, $O(\sqrt{\lambda})$. That cost increases linearly in volatility and sublinearly with the unit holding cost h and the total cost differential Δc.

The square-root formula (16) can also be used to quantify the value of dual sourcing, which is the reduction in total cost relative to single sourcing from Mexico. The latter is a special case of dual sourcing, and its cost is

$$C_{1M}^{\lambda^*} = (c_M^V + c_M^K)\lambda + \sqrt{2c_M^K h\sigma_1^2 \ln\bar{\xi}^{-1}}\sqrt{\lambda} + o(\sqrt{\lambda})$$
$$= c_M\lambda + 2c_M^K\hat{\mu}_{1M}\sqrt{\lambda} + o(\sqrt{\lambda}),$$

[8] The Brownian analysis in the online appendix shows that v_M is $O(\lambda^{1/4})$, and thus a third-order effect.

where $\hat{\mu}_{1M}$ can be shown to be $\sqrt{(h\sigma_1^2/2c_M^K)\ln\bar{\zeta}^{-1}}$ (see proof of Proposition 2 in the online appendix) so that

$$\begin{aligned} V_{\text{dual}} &= C_{1M}^{\lambda^*} - C^{\lambda^*} \\ &= \lambda\Delta c + 2(c_M^K(\hat{\mu}_{1M}^* - \hat{\mu}_M^*) - (c_M^V - c_C)\hat{\mu}_C^*)\sqrt{\lambda} + o(\sqrt{\lambda}) \\ &\geq \lambda\Delta c - 2\hat{\mu}_p\sqrt{\lambda}\Delta c + o(\sqrt{\lambda}) \\ &= \lambda\Delta c - \sigma\sqrt{2h\lambda\Delta c} + o(\sqrt{\lambda}), \end{aligned}$$

where we used the fact that $\hat{\mu}_C^* \leq \hat{\mu}_p$ and $\hat{\mu}_C^* \leq \hat{\mu}_M^*$. The lower bound on value is tight if c_M^K is small (so that $\hat{\mu}_{1M}^* \simeq \hat{\mu}_M^*$). The relative value of dual sourcing,

$$\frac{V_{\text{dual}}}{C_{\text{single}}} \gtrsim \frac{\Delta c}{c_M} - \sigma\sqrt{2\frac{h}{\lambda c_M}\frac{\Delta c}{c_M}},$$

is bounded from above by the relative sourcing cost savings (which are deterministic), but is reduced by increased working capital requirements (which reflect the cost of variability). The latter reflect the corrupting influence of variability, which increases when (1) demand or supply volatility, (2) holding costs relative to total (capacity + variable) Mexico cost, or (3) the relative China advantage increase, or when (4) average demand is low. Any of those four factors can lead to the lower bound on the relative value of dual sourcing becoming negative, suggesting that the value of dual sourcing would be small.

The TBS policy assumes that feedback control on China is not feasible; which precludes the comparison of dual sourcing with single sourcing from China. If one allows feedback control, this comparison follows the same lines as our comparison with single sourcing from Mexico.

7. Serial and Cross-Correlated Demand and Supply

So far we have confined the analysis to settings in which the demand and supply processes are tractable, independent renewal processes. The strength of the Brownian approximation, however, is not only analytic tractability but also generality: It can handle complex correlated processes, provided the variance terms are adjusted appropriately. Bradley and Glynn (2002) derived the asymptotic time-average variance of a general stationary demand and supply process $\{(D(t), S(t)): t \geq 0\}$ with interarrival times $\{(U_i, V_i): i \in \mathbb{N}\}$. We use that result to discuss two applications that highlight specific correlation structures observed in practice.

First, assume that demand has autocorrelation function $\text{corr}(U_1, U_{k+1}) = \theta^k$ with $|\theta| < 1$, and the supply processes are independent renewal processes. Then, $\text{covar}(U_1, U_{k+1}) = \theta^k \text{var}\, U_1$ and

$$\begin{aligned} \sigma^2 &= \lambda^2 \text{var}\, U_1(1 + 2\sum_{k=1}^{\infty}\theta^k) + \mu_C^2 \text{var}\, V_1^C \\ &= v_D^2\left(\frac{1+\theta}{1-\theta}\right) + v_C^2. \end{aligned}$$

Relative to our earlier setting of independent renewal processes, demand that is serially correlated over time has the effect of adjusting v_D^2. Positive time correlations increase volatility, and thus reduce the China allocation and the value of dual sourcing. In contrast, negative time correlations are mean reversing, and increase the China allocation and value of dual sourcing.

Second, assume the demand and China supply are correlated renewal processes with correlation coefficient ϕ, and the Mexico supply process is an independent renewal process. Then

$$\begin{aligned} \sigma^2 &= \lambda^2 \text{var}\, U_1 + \mu_C^2 \text{var}\, V_1^C - 2\lambda\mu_C \text{covar}(U_1, V_1^C) \\ &= v_D^2 + v_C^2\left(1 - 2\phi\frac{v_D}{v_C}\right). \end{aligned}$$

Relative to independent renewal processes, cross-correlated demand and the China supply have the effect of adjusting the China supply volatility v_C^2. Positive cross-correlations $\phi > 0$ could represent a situation where economic cycles impact both demand and China supply productivity. This would decrease China volatility, and thus increase the China allocation and the value of dual sourcing. In contrast, negative cross-correlations may arise because of congested transportation and import/customs processes, which would decrease the value of dual sourcing. To our knowledge, there is no empirical evidence as to which effect dominates.

Last, although the China and Mexico supply processes could be correlated, this would not impact our model given that Mexico volatility is a third-order effect. It is interesting that supply correlation—which has been advocated as an important reason to diversify the supply base—has little impact on sourcing allocation, and hence on the value of dual sourcing in our model.

8. Numerical Validation Study

We conduct a numerical study to illustrate and validate some of the key results discussed above. The goal of this validation study is to answer four questions: (i) How well does the TBS policy perform relative to the dual-base-stock policy, both in terms of cost minimization and allocation prediction? (We use the dual-base-stock policy as a proxy of the optimal policy

given that Bradley (2005) showed that a dual-base-stock policy is optimal when all interarrival and intersupply times are exponentially distributed.) (ii) How well does the Brownian prescription perform relative to the simulation-based optimal allocation? (iii) How well does the square-root allocation perform relative to the exact Brownian allocation? (iv) How well does our Brownian prescription perform if Mexico has a lead time?

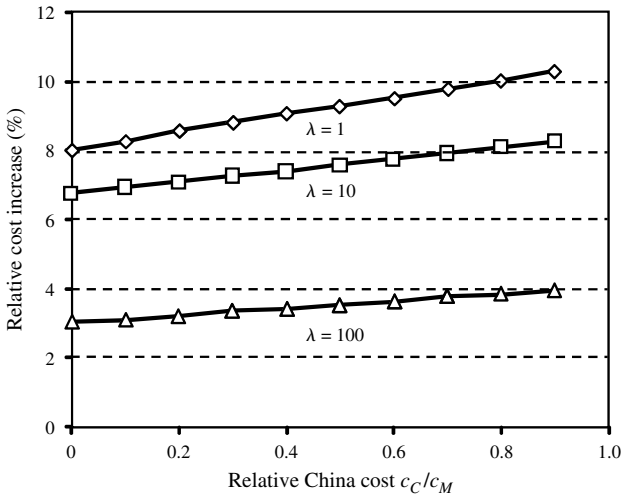
We shall address these questions using parameters similar to those studied in Bradley (2004) and Moinzadeh and Nahmias (1988): interdemand times are independent and identically normally distributed with coefficient of variation $v_D = 1$; intersupply times from Mexico are independent and identically normally distributed with $v_M = 1$; and intersupply times from China are independent and identically normally distributed with $v_C = 0.5$. (When simulating, negative sampled interarrival times were truncated to 0.[9]) The holding and backlogging costs are $h = \$1$ per period of time per unit and $b = \$50$ per period of time per unit. We set $c_M^K/c_M = 0.25$, and results were rather insensitive to changes in this fraction. Results shall be displayed typically as a function of the relative China cost $0 < c_C/c_M < 1$ and of the demand rate λ.

8.1. Comparing the TBS With Dual-Base-Stock Policies

The first step in the validation study investigates how well the optimized TBS policy performs relative to the more complex dual-base-stock policy. For various values of China's relative cost advantage $\Delta c/c_M$ and volumes λ, we simulated the total cost under a TBS policy with various base-stock levels. A numerical search then found the cost-minimizing capacities and base-stock level, and the corresponding optimal cost under TBS. For the dual-base-stock policy, we first kept the TBS-cost-minimizing capacities. Then, we simulated the total cost for a grid of possible base-stock pairs and obtained both optimal base-stock levels using a numerical search over this grid. An extensive numerical study in which we also optimized over the capacities under a dual-base-stock policy found no significant improvement, suggesting that the dual-base-stock performance is relatively insensitive to the capacities. This also suggests that the optimal China supply rate for our single-base-stock policy remains nearly optimal for the dual-base-stock policy, echoing the finding in Scheller-Wolf et al. (2006) that compares single-index with dual-index policies.

Figure 4 depicts the relative cost penalty of using the (optimized) TBS policy compared to using the (optimized) dual-base-stock policy for $\lambda = 1, 10$, and 100. (The sample error was less than 2% of the either cost.) As the China cost increases, the benefit of controlling the China supply (shutting it off and preventing large excess inventory) increases. Yet this benefit decreases as the volume increases, reflecting the fact that the inventory cost is of second order compared to the sourcing cost. This suggests that the use of the optimized TBS policy in practice (instead of a more complex policy) can be justified for larger volumes or when China's cost advantage is large.

Figure 4 Relative Cost Increase of TBS Over Dual-Base-Stock Policy



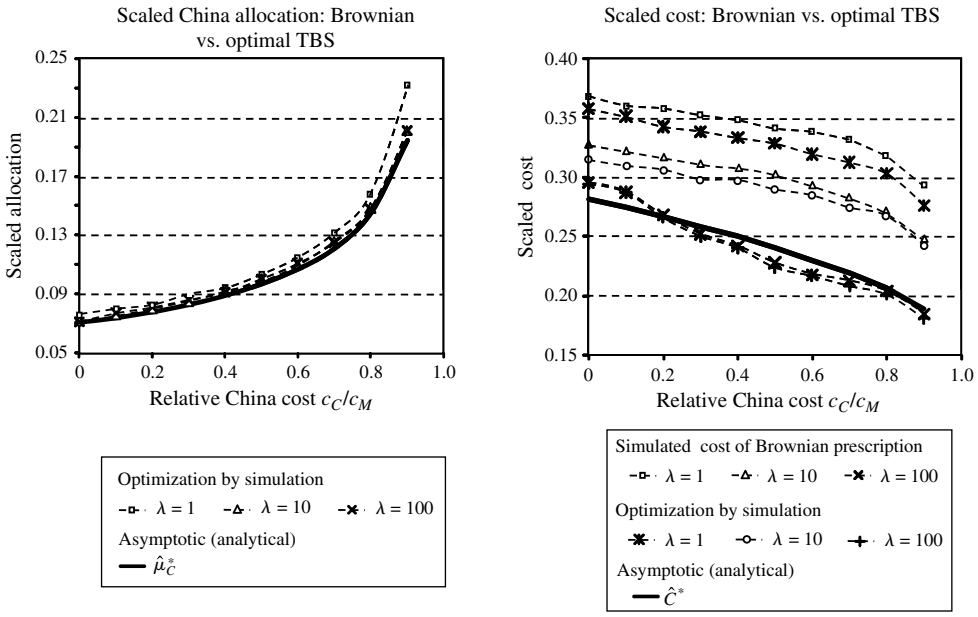
8.2. Comparing the Brownian Prescription to Simulation-Based Optimization

The second question addressed by our validation study is as follows: Under the TBS policy, how close is the Brownian scaled allocation $\hat{\mu}_C^*$ to the optimal scaled allocation $(\mu_C^{\lambda*} - \lambda)/\sqrt{\lambda}$, where $\mu_C^{\lambda*}$ is obtained by simulation-based optimization?

The left panel of Figure 5 shows how the optimal scaled China allocations for $\lambda = 1$, 10, and 100 converge to the allocation $\hat{\mu}_C^*$ that minimizes the asymptotic Brownian cost. It is remarkable that even for $\lambda = 1$, the relative error is less than 16% and below 8% as long as the China cost advantage exceeds 10%. For $\lambda = 10$ and above, the relative error was not statistically significant. Note, however, that the relative error on the *total* allocation prescription $\lambda - \hat{\mu}_C^*\sqrt{\lambda}$ depends on the volume λ and is much smaller, especially as λ increases. The same comment applies to the relative cost difference, which is shown in the right panel of Figure 5. That panel shows the scaled cost of the optimal control $(\mu_C^{\lambda*}, \mu_M^{\lambda*}, s^{\lambda*})$ obtained by optimization via simulation against the simulated scaled cost of the Brownian prescription $(\lambda - \hat{\mu}_C^*\sqrt{\lambda},\ \hat{\mu}_M^*\sqrt{\lambda},\ \hat{s}^*\sqrt{\lambda})$, both for $\lambda = 1$, 10, and 100. As proved, both converge to the asymptotic cost $\hat{C}^*$ (which is of the $\sqrt{\Delta c}$ form).

[9] We computed the coefficient of variation of the truncated sample and found it was very close to that of the nontruncated distribution.

Figure 5 Comparing the Brownian Allocation to the Allocation Optimized via Simulation



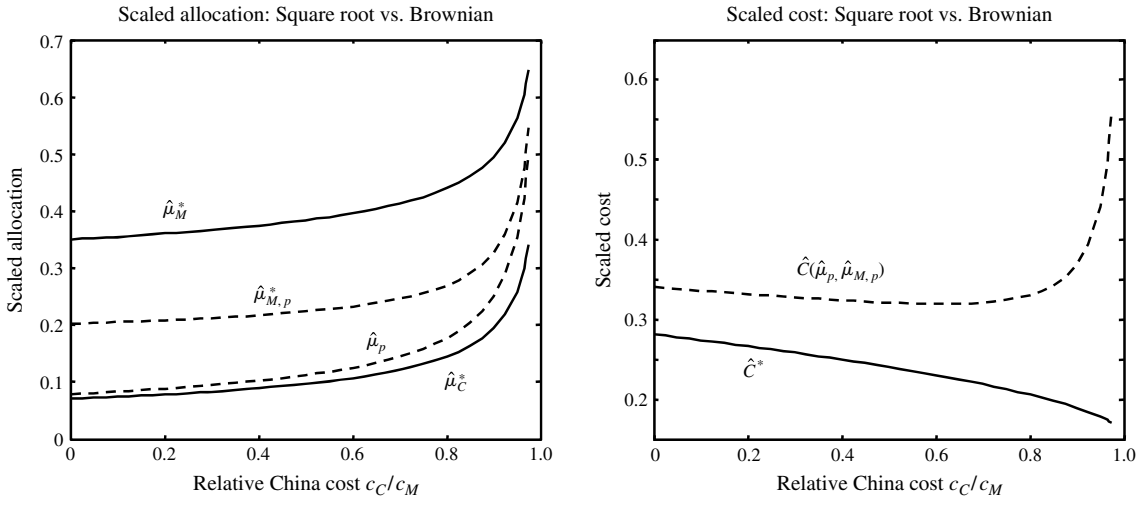
Yet even for $\lambda = 1$, the relative error in *scaled* cost between the prescription and the optimal control was less than 7%. The main implication is that the Brownian prescription is a good and useful approximation of the optimal strategic China allocation, even for small volumes.

8.3. Comparing the Square-Root Allocation to the Brownian Allocation

The left panel in Figure 6 shows the optimal capacities $\hat{\mu}_C^*$ and $\hat{\mu}_M^*$ in the Brownian model, as well as the square-root approximation $\hat{\mu}_p$. To evaluate cost differences, we also solved first-order condition (8) for the optimal Mexican capacity given $\hat{\mu}_C = \hat{\mu}_p$ and denote it by $\hat{\mu}_{M,p}$. The right panel shows the optimal cost $\hat{C}^*$ and the cost $\hat{C}(\hat{\mu}_p, \hat{\mu}_{M,p})$ when using the square-root formulae.

One can observe that the scaled square-root allocation $\hat{\mu}_p$ is a reasonable approximation of the exact scaled Brownian allocation $\hat{\mu}_C^*$, but the error increases as the China cost increases. Indeed, in our numerical study, the allocation difference is about 10% and below 28% as long as the China cost advantage exceeds 10%. Keep in mind that the relative error on the *total* allocation prescription $\lambda - \hat{\mu}_C^*\sqrt{\lambda}$ depends on the volume λ and will be much smaller, especially as λ increases. The same comment applies to the relative cost difference. The main implication is that the simple square-root formulae provide a reasonable starting point for the strategic China allocation.

Figure 6 Comparing the Square-Root Allocation $\hat{\mu}_p$ to the Brownian Allocation $\hat{\mu}_C^*$

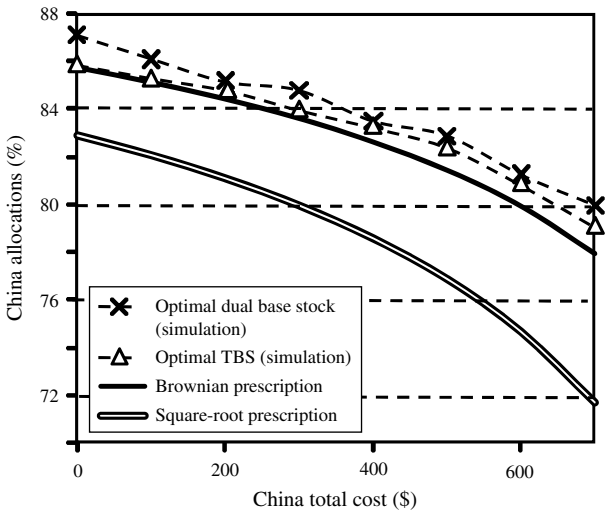


8.3.1. Practice-Based Validation Study. The numerical study reported so far assumed parameter values that have traditionally been used in the literature to assess the quality of our analytic approximations and of the TBS policy. Now, in an attempt to test the robustness of these results, we next calibrate the parameters using the actual data observed in practice in our motivating example. In addition to choosing the unit of time and unit of money, the key calibration involves the coefficients of variation using real data.

The monthly demand experienced by the firm varied between 5,000 and 67,000 units. Using a time unit of one month, we set $\lambda = 5{,}000$. The sourcing cost c_C varied between a few hundred and a couple thousand dollars. Using \$1 as the monetary unit, we set $c_C = 1{,}000$. The annual holding cost h equals (interest rate + physical holding cost) $\times\ c_C$ and was estimated at about $0.6c_C = \$600/\text{year}$ to reflect the short product life cycle, or $h = 50$ (per month). We kept the ratio b/h equal to 50 as before. Next, the monthly demands in the actual observed data exhibited a coefficient of variation between 0.05 to 1.25. These values were converted into the coefficient of variation of interarrival times using renewal theory: $v_D = \lambda \times$ (the coefficient of variation of demand rate) $= 5{,}000 \times (0.05\text{to } 1.25) = 15$ to 88. The fact that the interarrival coefficients of variations are significantly higher than the monthly coefficient of variation is because the latter exhibits strong aggregation effects. In addition, actual order patterns are staggered or batched. For example, an order for 1,000 units results in one high interarrival time followed by 999 interarrival times of 0. To simulate these arrival processes, we sampled interarrival times that were independent and identically gamma distributed.

We used these parameters to obtain both the allocation and the corresponding total cost using (1) simulation-based optimization, (2) the Brownian prescription, and (3) the square-root approximation, all assuming a TBS policy, and (4) simulation-based optimization assuming a dual-base-stock policy. As shown in Figure 7, the Brownian prescription was very close to the optimal allocations under both TBS (the relative error was smaller than 1.4% and smaller than 0.8% if China cost < 500) and dual base stock (the relative error was smaller than 2.5% and smaller than 1.0% if China cost < 500). As proved, the square-root prescription $\lambda - \hat{\mu}_p\sqrt{\lambda}$ is a lower bound on the China allocation: its error is between 3.4% and 9.3% relative to optimal TBS, and between 4.8% and 10.4% relative to dual base stock.

Figure 7 Comparing the Optimal Allocation Using Parameters Consistent with Practice



8.4. How Well Does the Brownian Prescription Perform If Mexico Has a Transportation Lead Time?

Our analysis has assumed that the Mexico supply incurs an endogenous delay because of orders queuing up for stochastic production. To compare this with traditional inventory models, we simulated the optimal allocation using the practice-based parameters when Mexico supply incurs an additional transportation or information lead time. Figure 8 shows that the optimal China allocation decreases as the Mexico lead time increases. An informal explanation is that the relevant metric of volatility is the volatility of the lead time demand, which increases, and our square-root formula thus predicts a decrease in the China allocation. The main conclusion, however, is that the Brownian prescription derived in this paper is quite robust under practical settings where the nearshore transportation lead time is about a week or less.

Figure 8 Brownian Prescription vs. the Optimal Allocation If Mexico Has a Transportation Lead Time

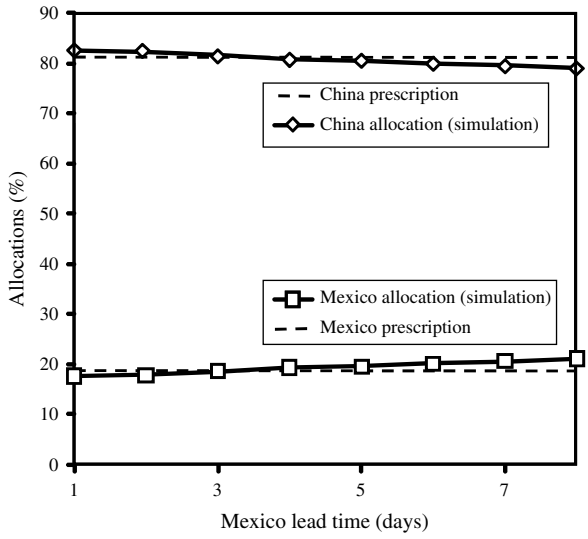
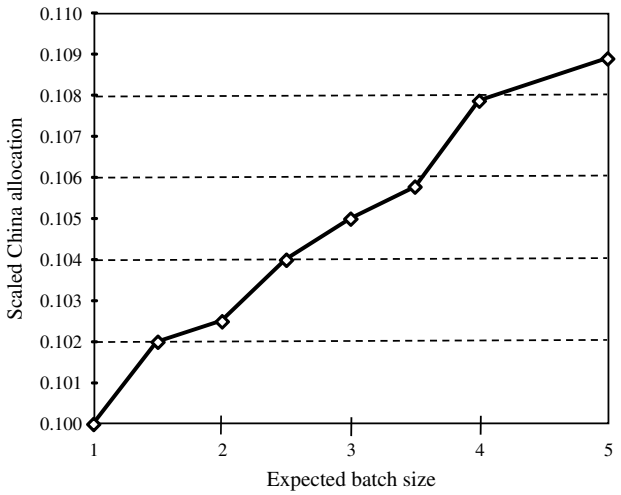


Figure 9 Scaled China Allocation as a Function of the Batch Size



8.5. How Well Does the Brownian Prescription Perform If Batch Orders Are Possible?

Our analysis has assumed that orders always occur one unit at a time. We next explore the robustness of our results when this assumption is violated. We report a numerical validation study in which we allowed for batch arrivals, using the parameters as reported at the beginning of §8 with China cost = $0.5 \times$ Mexico cost. For the same arrival point process with arrival rate $\lambda = 100$ as used in §8, we increased the expected number of units demanded (which we refer to as the batch size) at each arrival epoch. We used a geometric distribution to generate the actual batch size at each epoch. As Figure 9 shows, the scaled optimal China allocation increases in the batch size (holding total expected demand per period fixed). However, the increase is very small and leads to an error of less than 1% in the scaled allocation. Note that this means that the error in the total allocation will be much smaller, especially as λ increases.

9. Summary and Discussion

The dual-sourcing literature has traditionally focused on determining sophisticated dynamic policies that approach optimal performance. The main research objective of this paper, however, is more strategic in nature: to determine the near-optimal average sourcing allocation. We were able to answer this question by assuming a simpler policy that is used in practice. This tailored base-surge policy echoes a fundamental tenet in strategy: it aligns the ordering patterns with the core competencies of the suppliers. The constant base allocation allows China to focus on cost efficiency, whereas Mexico's quick response is utilized only dynamically to guarantee high service. Our model is a first attempt to provide some theory and quantification of this intuition of tailoring the sourcing strategy.

The model provides the following insights and quantification on strategic allocation. First, we present an analytic characterization of the TBS dual-sourcing policy that culminates in a simple square-root formula. This formula specifies the near-optimal strategic allocation that separates stochastic demand into "base" and "surge." Second, we determine the target inventory level and the corresponding cost under this near-optimal allocation. Our formulas allow an estimation of working capital requirements under dual sourcing, which have been lacking in the literature. Third, we identify and value the key drivers of dual sourcing. The square-root formula suggests a classification into first- and second-order drivers, and highlights the key role of supply and demand volatilities in dual sourcing. Our mode of analysis allows us to go beyond the typical assumptions of independence and also discuss the impact of serial time correlations as well as intra-demand/supply correlations.

A numerical study demonstrates that the results are robust and validates practice. We demonstrate robustness by showing that the TBS policy is near optimal in terms of total cost minimization, that the Brownian model provides reasonably accurate predictions of allocation and cost, and that the square-root formula results in a simple and useful estimate of strategic China allocation. The numerical study also validates the consultants' recommendation of allocating roughly three-quarters to the slow source as a starting point. With more specific data, the three-quarter allocation can and should be further tailored to the specific demand and supply characteristics using our results.

As with every model, ours has limitations. We do not explicitly model scale economies (such as those arising from fixed costs) in ordering, production, or capacity costs. Our results, however, show the presence of scale economies (our expressions are nonlinear in the demand rate λ) due to statistical economies of scale. Our policy assumes that feedback control on China is not feasible, which precludes the comparison of dual sourcing with single sourcing from China. If one allows feedback control, this comparison follows the same lines as our comparison with single sourcing from Mexico. Finally, we have focused on a single product and a single market setting under centralized control. Future work should extend to multiproduct, multimarket settings under decentralized control.

10. Electronic Companion and Teaching Game

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/. We have developed an in-class game to port our academic insights to the

classroom and practice; see Allon and Van Mieghem (2009) for a description.

Acknowledgments

The authors are grateful to Cort Jacoby, Ruchir Nanda, and Brian Bodendein from Deloitte Consulting, and to Achal Bassamboo, Marty Lariviere, Hyo duk Shin, Jing-Sheng Song, and the anonymous reviewers for detailed suggestions that improved the content of this paper.

References

Allon, G. , J. A. Van Mieghem. 2009. The Mexico-China sourcing game: Teaching global dual sourcing. Working paper, Kellogg School of Management, Northwestern University, Evanston, IL.

Barankin, E. W. 1961. A delivery-lag inventory model with an emergency provision. *Naval Res. Logist. Quart.* **8**(3) 285–311.

Bradley, J. R. 2004. A Brownian approximation of a production-inventory system with a manufacturer that subcontracts. *Oper. Res.* **52**(5) 765–784.

Bradley, J. R. 2005. Optimal control of a dual service rate M/M/1 production-inventory model. *Eur. J. Oper. Res.* **161** 812–837.

Bradley, J. R., P. W. Glynn. 2002. Managing capacity and inventory jointly in manufacturing systems. *Management Sci.* **48**(2) 273–288.

Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. J. H. Dshalalow, ed. *Advances in Queueing: Theory, Methods, and Open Problems*. CRC Press, Boca Raton, FL, 463–480.

DeCroix, G., J. S. Song, P. Zipkin. 2005. A series system with returns: Stationary analysis. *Oper. Res.* **53**(2) 350–362.

Ding, Q., L. Dong, P. Kouvelis. 2007. On the integration of production and financial hedging decisions in global markets. *Oper. Res.* **55**(3) 470–489.

Fleischmann, M., R. Kuik, R. Dekker. 2002. Controlling inventories with stochastic item returns: A basic model. *Eur. J. Oper. Res.* **138** 63–75.

Fukuda, Y. 1964. Optimal policies for the inventory problem with negotiable leadtime. *Management Sci.* **10**(4) 690–708.

Janssen, F., T. De Kok. 1999. A two-supplier inventory model. *Internat. J. Production Econom.* **59** 395–403.

Lau, H. S., L. G. Zhao. 1994. Dual sourcing cost-optimization with unrestricted lead-time distributions and order-split proportions. *IIE Trans.* **26**(5) 66–75.

Lu, X. L., J. A. Van Mieghem. 2009. Multimarket facility network design with offshoring applications. *Manufacturing Service Oper. Management* **11**(1) 90–108.

Moinzadeh, K., S. Nahmias. 1988. A continuous review model for an inventory system with two supply modes. *Management Sci.* **34**(6) 761–773.

Moinzadeh, K., C. P. Schmidt. 1991. An (*S*-1, *S*) inventory system with emergency orders. *Oper. Res.* **39**(2) 308–321.

Porteus, E. L. 2002. *Foundations of Stochastic Inventory Theory*. Stanford University Press, Stanford, CA.

Rosenshine, M., D. Obee. 1976. Analysis of a standing order inventory system with emergency orders. *Oper. Res.* **24**(6) 1143–1155.

Scheller-Wolf, A., S. K. Veeraraghavan, G.-J. van Houtum. 2006. Inventory policies with expedited ordering: Single index policies. Working paper, Wharton at the University of Pennsylvania, Philadelphia.

Sheopuri, A., G. Janakiraman, S. Seshadri. 2007. New policies for the stochastic inventory control problem with two supply sources. Working paper, Stern School of Business, New York University, New York.

Song, J. S., P. Zipkin. 2009. Inventories with multiple supply sources and network of queues with overflow bypasses. *Management Sci.* **55**(3) 362–372.

Tagaras, G., D. Vlachos. 2001. A periodic review inventory system with emergency replenishments. *Management Sci.* **47**(3) 415–429.

Van Mieghem, J. A. 2008. *Operations Strategy: Principles and Practice*. Dynamic Ideas, Belmont, MA.

Veeraraghavan, S. K., A. Scheller-Wolf. 2008. Now or later: A simple policy for effective dual sourcing in capacitated systems. *Oper. Res.* **56**(4) 850–864.

Whittemore, A. S., S. C. Saunders. 1977. Optimal inventory under stochastic demand with two supply options. *SIAM J. Appl. Math.* **32**(2) 293–305.

Zipkin, P. H. 2000. *Foundations of Inventory Management*. McGraw-Hill, New York.