Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Supply Chain Planning for Random Demand Surges: Reactive Capacity and Safety Stock

Lu Huang, Jing-Sheng Song, Jordan Tong

To cite this article:
Lu Huang, Jing-Sheng Song, Jordan Tong (2016) Supply Chain Planning for Random Demand Surges: Reactive Capacity and Safety Stock. Manufacturing & Service Operations Management 18(4):509-524. http://dx.doi.org/10.1287/msom.2016.0583

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 18, No. 4, Fall 2016, pp. 509–524
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2016.0583

Supply Chain Planning for Random Demand Surges: Reactive Capacity and Safety Stock

Lu Huang
Google Inc., Mountain View, California 94043, huanglu@google.com

Jing-Sheng Song
Fuqua School of Business, Duke University, Durham, North Carolina 27708, jssong@duke.edu

Jordan Tong
Wisconsin School of Business, University of Wisconsin–Madison, Madison, Wisconsin 53706, jordan.tong@wisc.edu

Globalization, innovation, social media, and exposure to natural and man-made disasters have increased organizations' need to cope with demand surges: random, significant increases in demand in an otherwise relatively stable demand environment. To build supply chain capabilities, organizations face a choice between two fundamentally different sourcing strategies—reactive capacity and safety stock. We develop a framework to guide the joint sourcing strategy that minimizes the long-run average cost under a target service level. A salient feature of our modeling framework is a novel demand model that captures important continuous-time, non-Markovian characteristics of demand surge trajectories. In addition to the total magnitude as typically modeled in the literature, we define several other metrics of surges, including duration, intensity, compactness, peak position, volatility, and frequency. The resulting optimization problem is challenging because it requires evaluating, for any sample path, whether surge demand can be satisfied at every point in time—a refined feature that traditional models do not have. To identify the optimal strategy, we first characterize the optimal production and deployment policy for any given strategy and then transform the original problem into two equivalent but more tractable problems. Finally, through stochastic comparison techniques, we show how the magnitude and predictability of surge demand characteristics (mentioned above) and the cost profiles of each strategy impact the optimal joint strategy.

Keywords: supply chain risk management; inventory planning; demand surges; Markov modulated demand; reactive capacity; safety stock; stochastic comparison
History: Received: June 5, 2015; accepted: March 7, 2016. Published online in *Articles in Advance* August 24, 2016.

1. Introduction

A central challenge in supply chain management is to develop system capabilities to match supply with demand in a cost-efficient manner. While there exist several demand models in the operations and supply chain literature that include random components, this paper considers a new and particularly challenging type of demand uncertainty—random demand surges. These are unpredictable yet substantial increases in demand in an otherwise relatively stable environment. Consider the following three illustrative examples:

- *Severe weather and disasters*. Severe weather (e.g., a hurricane or tsunami) and man-made disasters (e.g., an oil spill or fire) generate unpredictable and large spikes in demand for both private firms and humanitarian organizations. Moreover, when disasters disrupt suppliers of critical industrial components, alternative suppliers of those components typically experience sudden and substantial demand surges. The combination of the rapid speed of globalization and the increased frequency of natural disasters has exposed more firms to these types of demand surges.
- *Structural demand*. Beyond normal demand derived from a firm's own sales processes, firms occasionally secure sudden and large demands through external circumstances (e.g., new business partners, mergers, acquisitions). These structural events represent important and unique operational planning challenges due to their unpredictability and magnitude. For example, in one of the authors' work with a high-tech Fortune 500 company, preparing for surges due to structural events is a far more pressing problem than managing normal demand fluctuations. Similarly, an executive at a large firm in the semiconductor industry reported to us that one of their most difficult challenges is dealing with the unpredictability and infrequency of winning large contracts in developing

countries that put enormous additional pressure on their supply chain.

• *New product introductions and "viral" products.* Finally, new product introductions typically create surges in demand for both retailers and suppliers. Especially for supply chains in the fashion and high-tech industries, the demand process can be well conceptualized by a series of unpredictable surges associated with the release and/or sudden popularity of new products. One recent example is Apple's release of the iPhone 6 and iPhone 6 Plus in 2014. Two months after the product launch, it was reported that only 58% of iPhone 6/6 Plus models were in stock at retail stores nationwide (Yarow 2014). Furthermore, social media has created the possibility for products to go "viral," causing sudden and dramatic increases in demand (e.g., Shontell 2010). For example, in December of 2012, a story about a small sweatshirt company went "viral" on Twitter and Facebook, causing the small company, American Giant, to sell out in less than 36 hours (Manjoo 2013).

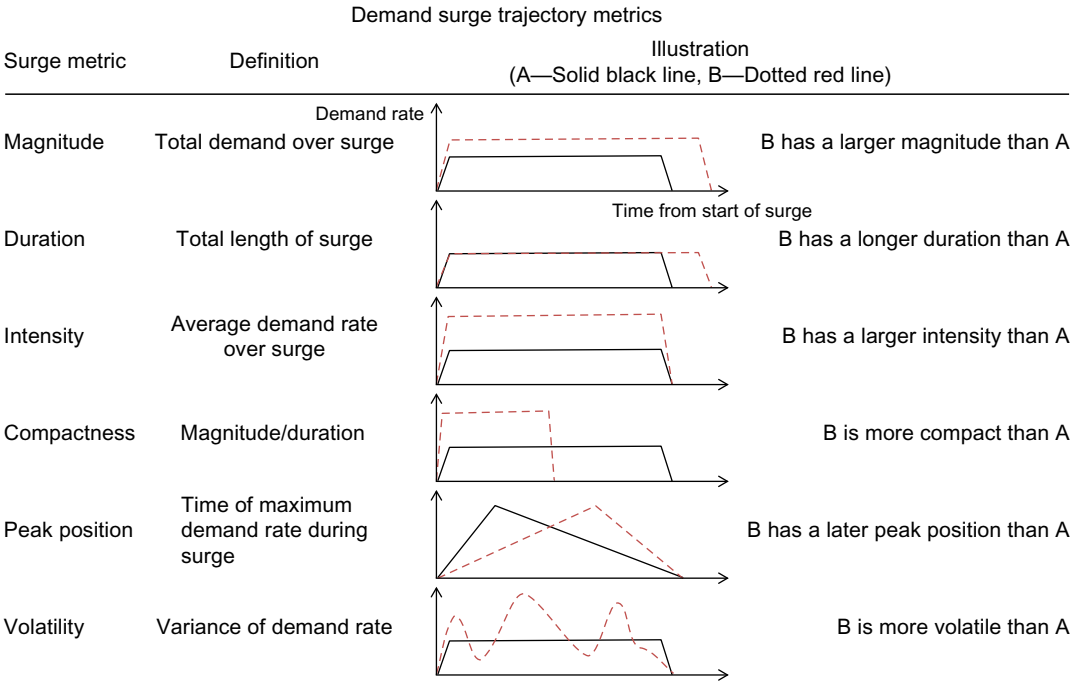
A demand process with surges differs drastically from a process with ordinary, stationary fluctuations. There are several new dimensions of uncertainty to consider with a random demand surge model: What is the likelihood of a surge demand occurring? What is the expected duration and intensity of a surge? What kind of trajectory will the surge demand follow? How difficult is it to predict surge demand characteristics? Our underlying demand model (see Section 3) captures these types of uncertainties. It is partly inspired by the state-space model of the environment (Song and Zipkin 1993). However, a key difference is that our model provides an explicit characterization of the entire evolution of a demand surge. A demand surge realization is not a single magnitude, but an entire trajectory. These trajectories can be described by certain metrics, which we develop. Figure 1 defines and illustrates some of these metrics. In addition to the total magnitude, there are several other metrics by which we can characterize surges, including duration, intensity, compactness, peak position, and volatility. Furthermore, these metrics portray *realizations* of demand surges. Therefore, we can characterize demand surges not only by the expected size of each metric (first-order effect), but also by how confident one is concerning each metric (second-order effect).

This paper examines a fundamental supply chain planning question that the particular nature of surge demands presents: How should an organization design its supply chain to position itself to efficiently cope with random demand surges? There are two fundamental ways a firm can prepare for surges: build reactive capacity or hold safety stock. Reactive capacity is the ability to ramp up production (above normal levels) when needed. To build reactive capacity, a firm can, for example, hold additional equipment and standby labor to assemble products quickly if needed. Or, they can contract with an outside firm to give priority to their needs. On the other hand, safety stock is inventory that is held at all times and can be deployed when a surge occurs. To keep safety stock for demand surges, a firm can set aside extra inventory in their own warehouses, or they can outsource safety stock by contracting with another vendor to hold the inventory for them. Of course, one can also employ a combination of reactive capacity and safety stock. For example, the high-tech retailer we worked for, mentioned earlier, not only holds extra machines in its service centers, but also compensates its vendors to hold additional components so that the vendors can quickly assemble the components into machines if needed. Similarly, to prepare for disasters, humanitarian and governmental organizations both maintain safety stocks of key relief items at their own warehouses and contract with manufacturers for standby access to their production capacity should a disaster strike (e.g., see National Association of State Procurement Officials 2013).

The main objective of this paper is to better understand the optimal combination of reactive capacity and safety stock and how it depends on the type of surge demands the supply chain faces. To accomplish this goal, we construct an analytical model. There are two policies to determine: a planning policy (i.e., how much reactive capacity and safety stock to maintain) and a deployment policy (i.e., how to use the reactive capacity and safety stock to satisfy surge demands). We are able to characterize the optimal deployment policy and then obtain the optimal planning policy. The model is somewhat stylized, yet it includes essential features of real systems. We capture reactive capacity via a fixed maximum production rate with a linear reservation cost. Safety stock is immediately available and carries a linear holding cost. In this way, the model captures two fundamental differences: on one hand, safety stock is immediately available while inventory from reactive capacity requires some time for production; on the other hand, reactive capacity can generate product for a long time while only a limited amount of safety stock is available. We assume, plausibly, that a surge demand must be met immediately or it is lost forever. The objective is to minimize total cost, subject to a minimum service-level constraint.

Our main contributions are as follows. First, the new demand model captures demand surges in a more nuanced manner than traditional models in the operations management literature. By specifying surges through a set (countable or uncountable) of sample paths, our model is remarkably flexible and can capture heretofore unstudied aspects of demand

Figure 1 **(Color online) Surge Trajectory Metrics and Illustrations**



uncertainty. These features are important because they allow a more precise and explicit comparison between the advantages of safety stock and reactive capacity; the optimal way to prepare for some demand trajectories may be significantly different from that for others, even if demands are identical in terms of simple measures such as magnitude.

Second, we are able to characterize the optimal planning policy analytically. To do this, we first find the optimal deployment policy for any fixed reactive capacity and safety stock level and characterize its cost. We are able to prove that the service-level constraint (which in its original form requires calculating the probability that demand is satisfied at *every* point in time during the surge demand) is equivalent to another constraint that only requires evaluating the probability of satisfying demand at *one* point in time, which we refer to as the critical time. This simplification is key to making a seemingly difficult problem tractable. Finally, we show that the optimal policy type is determined by two thresholds that define four regions—do nothing, reactive capacity only, safety stock only, or both—and we demonstrate how to evaluate the best policy in each region.

Third, we conduct extensive sensitivity analyses to better understand how different surge demand/product characteristics (and the ability to accurately predict them) impact the optimal planning policy. Specifically, we leverage several forms of stochastic comparison to illuminate how surge demand metrics affect the optimal absolute and relative levels of reactive capacity and safety stock. Table 1 in Section 5 provides a high-level summary of the results. Some results are intuitively appealing. For example, we find that one should favor reactive capacity as surge demands last longer or their durations become more uncertain. In contrast, one should invest more heavily in safety stock for short and intense demand surges. Other insights provide structural intuition. For example, we find that the two strategies remain proportionally attractive as the intensity of surge demand grows, so long as a particular type of stochastic equality holds. Finally, some insights are counterintuitive. For example, we find that if an organization seeks to achieve a higher service level, although they should always increase reactive capacity, they may actually want to decrease safety stock. And, if safety stock is less expensive than reactive capacity, it is optimal to use safety stock only when surge demands are very frequent or very infrequent, but not necessarily in between. This set of results offers clear "strategic guidelines" that managers can understand and use.

The remainder of this paper proceeds as follows. Section 2 reviews related literature. Section 3 presents the demand model and formulates the optimization problem. Section 4 derives the optimal policy. Section 5 contains sensitivity analyses and stochastic comparisons. Section 6 discusses extensions and makes some concluding remarks. All proofs are in the online appendix (available as supplemental material at http://dx.doi.org/10.1287/msom.2016.0583).

2. Literature Review

This paper is related to those that study multiple supply modes to hedge against uncertainty. Several papers consider capacity reservation or backup agreements between a buyer and its supplier in the presence of supply risk or demand uncertainty. Such agreements can be designed to provide flexibility in industries where demand is highly volatile and difficult to forecast. For example, Eppen and Iyer (1997) consider a backup agreement between a catalog company and its manufacturers and show that the backup option can increase the committed order quantities. Henig et al. (1997) study a backup contract between an automobile company and its supplier. Capacity reservations can also be designed to mitigate supply and disruption risks for the buyer. Literature in this stream includes Tomlin (2006), Chopra et al. (2007), Hou et al. (2010), and Saghafian and Oyen (2012). Our work differs from previous work in the way we model surge demands and in our focus on the strategic differences between the two supply modes: reactive capacity is slower but more flexible than safety stock.

Our work is also related to studies on quick response and accurate response initiatives in fashion supply chains. In this context, enabled by advanced technologies, manufacturers of fashion products can provide a second production run before the selling season starts, albeit at a higher cost than the first run. Because the second production run takes place at a time closer to the beginning of the selling season, when more market information is available, it serves as reactive capacity. The key issue faced by a buyer is how to allocate his order quantity for the entire season to different production runs; see, for example, Fisher and Raman (1996) and Donohue (2000). These decisions are similar in spirit to our deployment policies. Whereas this line of work assumes fixed timing of the second production run with updated (one-time) demand forecast, we assume the timing of demand surge (and hence the timing of deploying the reactive capacity) is random and the surge demand has a dynamically evolving trajectory.

Procurement strategy under supply uncertainty and disruption has recently received a great deal of attention. Based on the type of supply uncertainty addressed, most of this work can be differentiated into three groups: random yield (Henig and Gerchak 1990, Anupindi and Akella 1993, Federgruen and Yang 2009), random capacity (Ciarallo et al. 1994, Erdem et al. 1999), and random disruption (Parlar and Perry 1996, Tomlin 2006, Babich et al. 2007). Our work is related to these papers in the sense that all consider procurement strategy under extreme uncertainty. However, whereas this literature deals with extreme uncertainty on the supply side, we address extreme uncertainty on the demand side.

Our state-dependent demand process is related to the Markov-modulated demand process widely adopted in the inventory literature, in which an underlying Markov chain that describes the demand environment drives the demand distribution (e.g., Iglehardt and Karlin 1962, Song and Zipkin 1993, Sethi and Cheng 1997, Abhyankar and Graves 2001, Chen and Song 2001, Muharremoglu and Tsitsiklis 2008). Relevant to our work, Aviv and Federgruen (1997) and Kapuściński and Tayur (1998) investigate optimal inventory policies for cyclic demand when the supply process has finite capacity. Furthermore, Bhat and Krishnamurthy (2014) consider the optimal control of a make-to-stock system with exponential processing time and a Markov-modulated Poisson demand process. This stream of research shows that the optimal inventory policy should be state dependent. In our model, the underlying demand environment is a renewal process, switching between normal and surge states. While the transition from the normal state to the surge state is Markovian, the transition in the other direction is not. Furthermore, we use a sample-path approach to model the non-Markovian demand evolution in the surge state. Nonetheless, we too follow a state-dependent policy, but focus on the policy in the surge state (by normalizing the policy parameters to zero for the normal state). In addition, we consider a dual sourcing problem—we decide *both* inventory and capacity reservation levels. In this regard, our work is somewhat related to Song and Zipkin (2009), who study a multisourcing inventory problem with a Poisson demand. While they analyze how to deploy the alternative sources in a Markovian world, we consider both high-level planning and lower-level deployment decisions in a non-Markovian environment.

We note that planning for surge demands is of particular interest for disaster preparedness. There has been a growing interest in studying some of the unique supply chain problems that disasters create and how to apply or extend our existing theories to these challenges (e.g., Van Wassenhove 2006, Ergun et al. 2012). Disaster supply chains are special in several ways, including the supply process, the demand process, financial resources, the participating players, and the information technology available (Ergun et al. 2012). Although our model is not specific to disaster preparedness, our new way of representing surge demands does address some aspects of disasters not covered in standard models.

Finally, it is worth mentioning that, in a broader sense, our demand model is related to regime switching models in economics and finance (for a review, see Piger 2007). The main premise of these models is that, over time, model parameters may change significantly (reflecting a different "regime"). Regime-shifting models have been a popular and effective

technique for capturing dynamic macroeconomic behavior for which constant parameter time series models are inadequate. In a similar spirit, our demand surge model is an attempt to capture more realistic demand patterns that stationary demand models cannot.

3. Model

We begin by describing the surge demand process for a single product. Then, we formulate the joint safety stock and reactive capacity planning problem. Throughout this paper, $y^+ = \max\{y, 0\}$, $x \wedge y = \min\{x, y\}$, and $\mathbf{1}_{\{\cdot\}}$ is the indicator function. In general, random variables are denoted by capital letters.

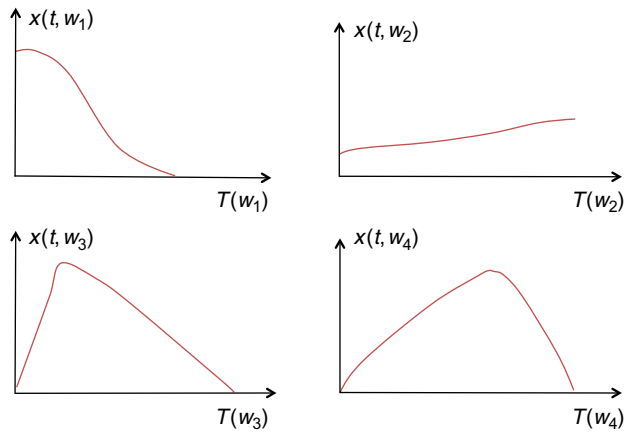
3.1. Demand Process

Consider an environment with a normal and a surge state. We discuss the extension to more normal and surge states in Section 6. The environment begins in a normal state, during which there is a constant demand rate. We assume the normal demand rate is zero without loss of generality. The system stays in the normal state for a random period following an exponential distribution with rate λ. Once it enters the surge state, the environment stays there for a random time T. After T, the environment returns to the normal state. We denote by $\tau = E[T]$ the expected duration of the surge state and by $\nu = 1/\lambda$ the expected time in the normal state.

A salient feature of our modeling framework is a novel way to describe surge demands. We assume the demand in the surge state is a continuous-time stochastic process $\{X(t),\ t \geq 0\}$. Instead of assuming parametric distributions of T and $\{X(t), t \geq 0\}$ (such as a lognormal distribution for T and a Brownian motion for X), we take a sample-path approach. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space for all possible demand sample paths (or scenarios) in the surge state. Each ω corresponds to a unique forecasted demand sample path. More specifically, each $\omega \in \Omega$ specifies surge duration $T(\omega)$ and the corresponding demand trajectory $\{X(t, \omega), 0 \leq t \leq T(\omega)\}$. Consequently, it also specifies a total surge demand magnitude, denoted $D(\omega) = \int_0^{T(\omega)} X(s, \omega)\, ds$. The measure $\mathbb{P}(\omega)$ gives the probabilities of these ω's. The sample space Ω can be either countable or uncountable.

Figure 2 provides an illustration of four sample demand trajectories. These sample paths vary in total magnitudes. They also vary in other dimensions. Some surges start with a high intensity but do not last long, as in $x(t, \omega_1)$, while others last longer with gradually increasing intensity, as in $x(t, \omega_2)$. Surge trajectories may peak early, as in $x(t, \omega_3)$ or later, as in $x(t, \omega_4)$. A very simple example of our model of random demand surges is to consider a random draw from one of these sample paths. Of course, the set of possible demand trajectories may be much larger than four—it need not even be countable.

Figure 2 (Color online) Depiction of Possible Surge Demand Paths



This sample-path definition of surges is consistent with the observation that demand forecasting usually provides a variety of possible scenarios along with a probability of each scenario (Sheffi 2005). It is also complementary to the growing trend of "big data" and increasingly common forecasting methods through statistical tools, such as those that rely on the bootstrapping method. Indeed, many companies and government agencies maintain databases containing long histories of trajectories of surge demands and the incidents that triggered them. They also have the analytical capability to generate probabilistic predictions for future demand surges based on this data (Sheffi 2005).

Although we define random demand surges in a sample-path manner, the model can also accommodate surge demand patterns that are defined parametrically. For example:

Example 1 (Rectangular Surge Demand). A rectangular surge demand has sample paths that stay flat over the surge duration. Let $\boldsymbol{\omega} = (l, s) \geq 0$, where l and s represent the random duration and intensity of surge demand, respectively. Then, $X(t, \boldsymbol{\omega}) = s$, $0 \leq t \leq T(\boldsymbol{\omega}) = l$.

Example 2 (Triangular Surge Demand). A triangular surge demand is characterized by its duration l, peak time β, and demand intensity s at the peak time βl. Let $\boldsymbol{\omega} = (l, \beta, s)$, where l, β, and s are positive random variables with $\beta \leq 1$. Then, $T(\boldsymbol{\omega}) = l$, and

$$X(t, \boldsymbol{\omega}) = \begin{cases} \dfrac{t}{\beta l} s, & 0 \leq t \leq \beta l, \\ \dfrac{l - t}{(1-\beta) l} s, & \beta l < t \leq l. \end{cases}$$

These simple demand evolution patterns can also be used as constituent elements for more complex

demand processes. Nevertheless, the crucial benefit of this surge demand model is not that it can accommodate several of these parametric types of demand surges. Rather, to foreshadow our analytical results in Section 5, the benefit is that the sample-path definition of surge demands facilitates stochastic comparisons via sample-path arguments.

For tractability, we assume that any realization of the demand trajectory $\{X(t,\omega), 0 \le t \le T(\omega)\}$ is continuous and unimodal with respect to t. This assumption allows us to solve the optimal deployment policy. However, it is also reasonable to assume that even if a surge demand may contain some local small oscillations, it usually maintains the unimodal property globally. Finally, we note that in the rest of this paper, when ω is given, we often suppress ω for notational simplicity. For example, we write T instead of $T(\omega)$ and denote $x(t) = X(t,\omega), 0 \le t \le T(\omega)$, when the context is clear.

3.2. Problem Formulation

We consider two operational levers: *safety stock* and *reactive capacity*. A safety stock of size m guarantees the immediate availability of up to m units in inventory when a surge state occurs. Reactive capacity of level μ guarantees the maximum available production rate μ to produce the product in demand while in the surge state. Let (m,μ) denote the joint planning policy. Once in the surge state, the manager must make a production decision for how to use the reserved capacity at any time t. We call the inventory produced from this production the *reactive stock*. Finally, the manager must decide how to use the safety stock and the reactive stock to meet surge demand. These decisions comprise the production policy from reactive capacity and the deployment policy.

Because the production and deployment problem depends on the planning policy, we proceed by formulating the problem backward: first formulating the production and deployment problem for a fixed (m,μ) before formulating the planning problem.

3.2.1. Stage 2: Production and Deployment Problem. Under any given joint planning policy, at each time point t in the surge state, the decision maker must decide a production policy $p_\mu(t)$ for how to produce from the reserved capacity, and a deployment amount $y_m(t)$ from the remaining safety stock and $y_\mu(t)$ from the remaining reactive stock. Let $(y_m(t), y_\mu(t) \mid 0 \le t \le T(\omega))$ denote a *deployment policy*. To determine the optimal deployment policy, we specify the economic factors associated with different operational levers. Safety stock incurs a holding cost h per unit/time. Capacity reservation costs r per unit/time. We also assume linear purchasing costs: it costs c to purchase units from safety stock and w to purchase units produced with reactive capacity. We focus our analysis on the case $c < w$ because the marginal production cost for a supplier is typically greater when he needs to increase production with short notice. Nevertheless, a similar analysis can be repeated for $c \ge w$, and most of our results still hold, although we omit those details here.

Given a planning policy (m,μ), let $C(m,\mu)$ denote the optimal expected deployment costs. It can be expressed as follows:

$$C(m,\mu) = E_\omega\left[\min_{p_\mu(t),\, y_m(t),\, y_\mu(t) \in S(t,\omega),\, 0 \le t \le T(\omega)} \left\{ c\int_0^{T(\omega)} y_m(t)\,dt + w\int_0^{T(\omega)} p_\mu(t)\,dt \right\}\right],$$

$$S(t,\omega) = \operatorname*{argmax}_{\tilde{p}_\mu(t),\, \tilde{y}_m(t),\, \tilde{y}_\mu(t) \in \mathbb{R}^+} \{\tilde{y}_m(t) + \tilde{y}_\mu(t), 0 \le t \le T(\omega)\} \quad (1)$$

$$\text{s.t.} \quad \tilde{y}_m(t) + \tilde{y}_\mu(t) \le X(t,\omega), \quad (2)$$

$$\tilde{y}_m(t) = 0, \quad \text{if } \int_0^t \tilde{y}_m(s)\,ds = m, \quad (3)$$

$$\tilde{y}_\mu(t) \le \tilde{p}_\mu(t), \quad \text{if } \int_0^t \tilde{y}_\mu(s)\,ds = \int_0^t \tilde{p}_\mu(s)\,ds, \quad (4)$$

$$\tilde{p}_\mu(t) \le \mu, \quad 0 \le t \le T(\omega).$$

Once a surge occurs, the planner aims to satisfy as much demand as possible using the available levels of reactive and safety stock, which is captured by the service-level maximization constraint (1). Here, we have defined the service level as the probability that all surge demand is immediately satisfied. Therefore, similar to the in-stock rate, a 95% service level is interpreted as there being a 0.95 probability that all surge demand is met upon arrival. Given this service level, the planner seeks to minimize costs. Constraint (2) guarantees that the deployed stock does not exceed the demand. Constraints (3) and (4) ensure that the amounts deployed from safety stock and reactive capacity are feasible under the safety stock level m and capacity level μ. Denote the optimal deployment policy by $(p^*_\mu(t), y^*_m(t), y^*_\mu(t))$.

3.2.2. Stage 1: Planning Problem. Now, given the costs from the production and deployment problem, the expected total cost rate function of the joint planning policy can be written as

$$G(m,\mu) = \frac{1}{\nu+\tau}([hm + r\mu]\nu + C(m,\mu)). \quad (5)$$

It is the sum of two types of costs. The first term is an expected preparedness costs, which is the combined safety stock holding and the reactive capacity reservation costs incurred during the normal state. The second term is the expected optimal deployment costs incurred during the surge state, which is obtained

from Stage 2 above. Here, we have assumed that neither inventory holding nor capacity reservation costs are incurred during the surge. Nevertheless, these assumptions are not critical, and it is straightforward to alter the analysis to account for the preparedness costs both during the normal and surge state without affecting the main results.

The *joint planning problem* is to minimize the long-run average system cost, under the surge demand service-level constraint

$$A: \quad \min_{m,\,\mu\in\mathbb{R}^+} \; G(m,\mu) \tag{6}$$

$$\text{s.t.} \quad P\{\omega\colon X(t,\omega) = y_m^*(t) + y_\mu^*(t), \; 0 \le t \le T(\omega)\} \ge \alpha. \tag{7}$$

The inequality (7) is the service-level constraint with a service-level lower bound α. Note that when making planning decisions, one does not know the total demand surge magnitude nor the kind of trajectory it will take (i.e., ω is unknown).

Another way to formulate this problem is to maximize surge demand coverage under a budget constraint of the long-run system cost $\bar{G}$. We will later show that these two formulations are equivalent.

4. Optimization

The joint planning problem is difficult to solve in its original form A; one must optimize over both the planning decisions (m,μ) and the deployment decisions $(y_m(t), y_\mu(t) \mid 0 \le t \le T(\omega))$ and determining the objective service level requires evaluating an inequality at every time point t. Furthermore, the optimal deployment decision is constrained by the underlying production policy $p_\mu(t)$ that shapes the reactive stock. We first address these challenges before presenting the optimal production and deployment policy.

4.1. Optimal Deployment Policy

We begin by solving the problem backward and characterize the optimal production and deployment policy. Observe that to determine the optimal production policy, we must specify what information about the surge demand is available to the manager in real time. Let $\mathcal{F}_t(\omega)$ be the manager's demand information about the total surge demand magnitude $D(\omega) = \int_0^{T(\omega)} x(s,\omega)\,ds$ at time t. Here $\mathcal{F}_t(\omega)$ is a σ-algebra satisfying $\mathcal{F}_s(\omega) \sqsubset \mathcal{F}_t(\omega)$ for $s < t$, implying more information is available as time passes. Therefore, $E_\omega[D(\omega) \mid \mathcal{F}_t(\omega)]$ is the manager's estimated total surge demand magnitude based on the information at time t. We make the following assumption:

Assumption 1. *At every time point t in the surge state, the manager's estimate of total surge demand magnitude $E[D \mid \mathcal{F}_t]$ is accurate enough to predict whether it is greater or less than the threshold $m + \mu t$, i.e., $E[D \mid \mathcal{F}_t] \le m + \mu t$ if and only if $D \le m + \mu t$.*

To be clear, Assumption 1 does not imply that one knows the total demand surge magnitude when making the planning decisions (m,μ). The demand surge is unknown when making planning decisions. Rather, Assumption 1 implies that upon entering a demand surge the manager can assess the surge well enough to determine whether its total magnitude is greater than or less than the total safety stock m. Then, as the surge continues and more information becomes available, the manager updates his or her estimate and the updated estimate is accurate enough to determine whether the actual surge magnitude is greater or less than the total safety and reactive stock if producing at full capacity $m + \mu t$. This assumption is sufficient to ensure the following proposed production policy is optimal:

Lemma 1. *For a given planning policy (m,μ), the optimal production policy produces at maximum capacity unless the estimated surge demand $E[D \mid \mathcal{F}_t]$ has been fulfilled or a stockout occurs. Production stops if the estimated surge demand $E[D \mid \mathcal{F}_t]$ is fulfilled, but continues at the minimum of μ and the demand rate if a stockout occurs. Mathematically,*

$$p_\mu^*(t) = \begin{cases} 0 & \text{if } m + \mu t \ge E[D \mid \mathcal{F}_t], \\ \mu & \text{if } \int_0^t x(s)\,ds < m + \mu t < E[D \mid \mathcal{F}_t], \\ \min\{\mu, x(t)\} & \text{otherwise.} \end{cases} \tag{8}$$

The optimal deployment policy $(y_m^*(t), y_\mu^*(t))$ is a greedy policy, deploying the stockpiled inventory first. Mathematically,

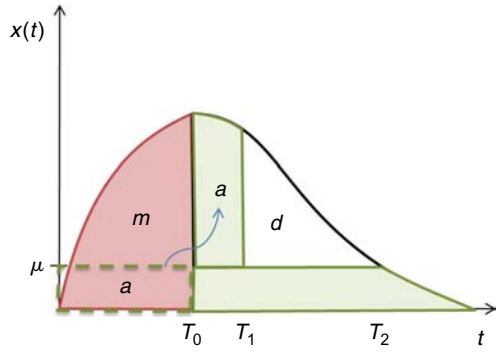
$$(y_m^*(t), y_\mu^*(t)) = \begin{cases} (x(t), 0), & \text{if } m > \int_0^t y_m^*(s)\,ds, \\ (0, x(t)), & \text{if } m = \int_0^t y_m^*(s)\,ds \text{ and} \\ & \quad \int_0^t p_\mu^*(s)\,ds > \int_0^t y_\mu^*(s)\,ds, \\ (0, p_\mu^*(t)), & \text{otherwise.} \end{cases} \tag{9}$$

Under the optimal production and deployment policy, the service level can be simplified to

$$P\left\{\omega\colon \int_0^t X(s,\omega)\,ds \le m + \mu t,\; 0 \le t \le T(\omega)\right\}. \tag{10}$$

For a depiction of the optimal production and deployment policy, refer to Figure 3. Once one has entered the surge state at $t = 0$, the optimal deployment policy is to deploy the stockpiled inventory m first because it is less expensive. Therefore, up to

Figure 3 **(Color online) Illustration of the Optimal Production and Deployment Policy**



time T_0, where $m = \int_0^{T_0} x(s)\,ds$, the optimal deployment rate from safety stock $y_m^*(t)$ is the demand rate $x(t)$. In the meantime, the manager needs to decide if he should be producing inventory from his reactive capacity. He produces at the maximum rate as long as his updated forecast for the total surge demand magnitude is greater than a threshold $E[D \mid \mathcal{F}_t] > m + \mu t$. In the figure, the total surge demand magnitude is indeed larger than this threshold, so the optimal production policy produces at full capacity such that at time T_0 there is μT_0 amount of reactive stock is available (depicted by area "a"). The manager uses this inventory to meet the shortfall of safety stock until it is also depleted (at time T_1). It then continues to produce at the minimum of the reactive capacity rate and demand rate. At time T_2 in the figure, one cannot produce more than can be consumed immediately to make up for the shortfall area "d" (demand is lost).

Assumption 1 allows us to satisfy demand as much as possible without overproduction due to demand overestimate. More importantly, it allows us to characterize $C(m, \mu)$ without dependency on the manufacturer's demand updating process $\mathcal{F}_t$, and thus helps us to focus on the impact of surge demand trajectories on the optimal planning policy, which is the main focus of this paper. Moreover, it is arguably consistent with demand forecast updating that occurs in practice. For example, once a disaster strikes, a new contract is won, or a new product is launched, the associated organization can often use early data to make an estimate of its total magnitude accurately enough to determine whether it is greater than or less than the threshold specified in Lemma 1.

4.2. Problem Transformation

4.2.1. Service-Level Transformation. Next, we simplify the service-level formulation. It is difficult to optimize over service level (10) in problem A because one needs to check whether it holds at every moment in time over the random surge duration T. The following definition helps us simplify this constraint to make the problem tractable. For any ω, we say time point T' is a critical time, if the fact that inventory position at this point is positive suggests that it is positive at all other points. Mathematically, T' is the time point such that if $\int_0^{T'} x(s)\,ds \le m + \mu T'$, then $\int_0^t x(s)\,ds \le m + \mu t$, $0 \le t \le T$.

Recall that the surge demand is unimodal, so its trajectory must follow one of the following patterns: (i) stationary, (ii) increasing, (iii) decreasing, or (iv) first increasing and then decreasing. This observation, combined with the nature of the optimal deployment policy above, leads to the following result.

Definition 1 (Critical Time). For any given ω, there exists a unique critical time T', which is a function of the reactive capacity μ. In particular, if $x(t)$ is stationary or increasing, then $T'(\mu) = T$. If $x(t)$ is first increasing then decreasing or is only decreasing, let $t_p = \arg\max\{x(t)\colon 0 \le t \le T\}$, and then

$$T'(\mu) = \begin{cases} T, & \text{if } x(T) \ge \mu, \\ \arg\{x(t) = \mu, t_p \le t \le T\}, & \text{if } x(T) < \mu < x(t_p), \\ t_p, & \text{if } x(t_p) \le \mu. \end{cases}$$

Figure 4 provides an illustration of the critical time. The notion of the critical time enables us to simplify the service-level characterization as follows.

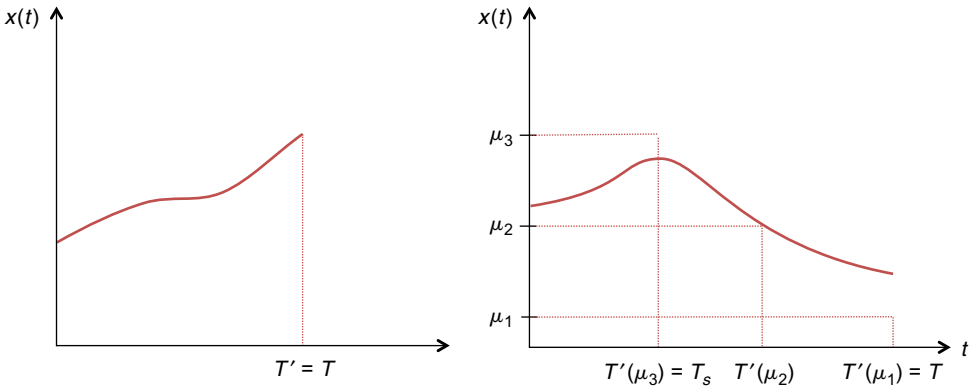
Lemma 2. *The service level* (10) *is equivalent to* $P\{\omega\colon \int_0^{T'(\mu,\omega)} X(s,\omega)\,ds \le m + \mu T'(\mu,\omega)\}$.

This equivalent expression of the service level is significantly easier to evaluate than the original one, because for each ω, we only need to check one inequality instead of checking the inequality for every moment in time. For instance, for rectangular surge demands, $T'(\mu,\omega) = l$. Thus, according to Lemma 2, we can rewrite the service level as $P\{\omega = (l,s)\colon sl \le m + \mu l\}$; namely, the service level can be determined by comparing the cumulative demand with the cumulative supply at the end of the surge. Similarly, for triangular surge demands, the critical time is $T'(\mu,\omega) = (\beta l + (1-\beta)l((s-\mu)/s)\mathbf{1}_{\{\mu \le s\}})$. Thus, according to Lemma 2, we can rewrite the service level as $P\{\omega = (l,s)\colon sl/2 - (\mu^2 l/(2s))(1-\beta) \le m + \mu l\}$ if $\mu \le s$ and 1 otherwise. We will return to these special cases in our analysis of the impact of surge demand characteristics in Section 5.

4.2.2. Cost Function Transformation. By leveraging the optimal deployment policy and the definition of the critical time, $C(m,\mu)$ can be simplified to

$$C(m,\mu) = E\big[w((D-m)^+ - (B'(\mu)-m)^+) + c[D \wedge m]\big], \quad (11)$$

Figure 4 **(Color online) Depiction of Critical Time Determination for the Increasing and Stationary Case (Left) and the Unimodal or Decreasing Case (Right)**



where $D(t) = \int_0^t X(s,\omega)\,ds, 0 \le t \le T(\omega)$, and $B'(\mu) = D(T'(\mu)) - \mu T'(\mu)$. Consequently, problem A can be simplified to

$$A': \quad \min_{m,\mu\in\mathbb{R}^+} G(m,\mu)$$

$$\text{s.t. } P\left\{\omega: \int_0^{T'(\mu,\omega)} X(s,\omega)\,ds \le m + \mu T'(\mu,\omega)\right\} \ge \alpha. \quad (12)$$

It can be shown that at optimality, the constraint (12) is tight, from which we obtain an analytical relationship between m and μ. Thus, A' can be reduced to a univariate optimization problem. However, it is still intractable because the convexity of the objective function is not guaranteed. To overcome this difficulty, we introduce a unit penalty cost b for each unsatisfied demand $[D(T') - \mu T' - m]^+$. In this way, we can transform the cost $C(m,\mu)$ and obtain the associated transformed total average cost:

$$G^p(m,\mu) = \frac{1}{\nu+\tau}((hm + r\mu)\nu + C^p(m,\mu)), \quad (13)$$

$$C^p(m,\mu) = cE[D] + (w-c)E[D-m]^+ + (b-w)E[B'(\mu) - m]^+. \quad (14)$$

The corresponding transformed optimization problem is

$$B: \quad \min_{m,\mu\ge 0} G^p(m,\mu), \quad (15)$$

which we refer to as the *penalty cost formulation*.

4.3. Optimal Planning Policy

We proceed by first solving the penalty cost formulation in problem B. Later, we prove that solving problem B is equivalent to the original service-level constraint formulation A. The function $G^p(m,\mu)$ is well behaved. Specifically:

Lemma 3. *$G^p(m,\mu)$ is jointly convex in (m,μ). The optimal solution (m^*,μ^*) to the problem B exists and can be obtained by solving the Kuhn–Tucker first-order conditions.*

It is intuitive that for any fixed capacity reservation cost r, the optimal stockpile decision should follow some kind of threshold policy; that is, keep stockpile if and only if the holding cost is not too high. Of course, the threshold also ought to depend on other parameters. Similarly, for any fixed holding cost h, we expect the optimal capacity reservation policy to be of a threshold type; that is, reserve capacity if and only if doing so is not too expensive, compared with other options. Below we show that this intuition is correct—the optimal joint planning policy is one of the following four types depending on the parameter range: (i) no action, (ii) employ safety stock only, (iii) employ reactive capacity reservation only, and (iv) employ a joint strategy. (The optimal joint planning policy may not be unique, but only under some trivial cases, which we ignore.) On the other hand, for the joint optimization, the thresholds are characterized by switching curves $h^*(r)$ and $r^*(h)$ in Proposition 1 below. To describe these switching curves, we first make the following definitions.

When $h < (b-c)/\nu$, define $m_1(h)$ as the solution to

$$P\{D > m_1\} = \frac{h\nu}{b-c}. \quad (16)$$

When $r < (b-w)\tau/\nu$, define $\mu_1(r)$ as the solution to

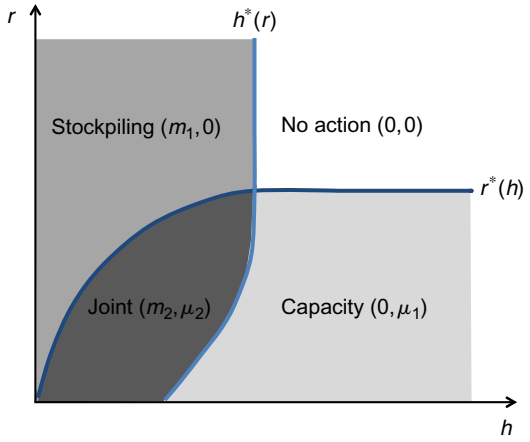
$$E[T'(\mu_1) \mid B'(\mu_1) > 0]P\{B'(\mu_1) > 0\} = \frac{r\nu}{b-w}, \quad (17)$$

and define (m_2,μ_2) as the solution to

$$\begin{aligned} &(b-w)P\{B'(\mu_2) > m_2\} + (w-c)P\{D > m_2\} = h\nu,\\ &E[T'(\mu_2) \mid B'(\mu_2) > m_2]P\{B'(\mu_2) > m_2\} = \frac{r\nu}{b-w}. \end{aligned} \quad (18)$$

The following proposition provides a full characterization of the optimal planning policy.

Figure 5 **(Color online) Optimal Planning Policy Regions**



Proposition 1. *Let*

$$h^*(r)=\begin{cases}\dfrac{w-c}{v}+\dfrac{b-w}{\nu}P\{B'(\mu_1(r))>0\}, & r<\dfrac{(b-w)\tau}{\nu},\\ \dfrac{b-c}{\nu}, & \textit{otherwise};\end{cases}$$

$$r^*(h)=\begin{cases}\dfrac{b-w}{v}E[T\,|\,D>m_1(h)]P\{D>m_1(h)\}, & h<\dfrac{b-c}{\nu},\\ \dfrac{(b-w)\tau}{\nu}, & \textit{otherwise}.\end{cases}$$

(i) *If $r\ge r^*(h)$ and $h\ge h^*(r)$, then no action is optimal, i.e., $(m^*,\mu^*)=(0,0)$.*

(ii) *If $r\ge r^*(h)$ and $h<h^*(r)$, then it is optimal to deploy safety stock only, i.e., $(m^*,\mu^*)=(m_1,0)$.*

(iii) *If $r<r^*(h)$ and $h\ge h^*(r)$, then it is optimal to deploy reactive capacity only, i.e., $(m^*,\mu^*)=(0,\mu_1)$.*

(iv) *If $r<r^*(h)$ and $h<h^*(r)$, then it is optimal to deploy a joint strategy, i.e., $(m^*,\mu^*)=(m_2,\mu_2)$.*

Moreover, $h^*(r)$ is increasing in r, and $r^*(h)$ is increasing in h.

Thus, the optimal planning policy is characterized by two switching curves, as depicted in Figure 5. As shown in the figure, it is optimal to employ a particular operational lever if and only if the unit cost of that lever is below a certain threshold, which depends on the unit cost of the other lever. For example, it is optimal to stockpile inventory if and only if the unit inventory holding cost h is below a certain threshold $h^*(r)$, as shown in the dark gray area. This threshold $h^*(r)$ is increasing in unit capacity reservation cost r; namely, when reactive capacity has lower unit reservation costs, inventory stockpiling must also have lower unit inventory holding costs to be deployed. Moreover, no action is taken when both strategies are too expensive ($h\ge(b-c)/\nu$ and $r\ge(b-w)\tau/\nu$), and a joint strategy is optimal when both strategies are sufficiently inexpensive ($h<h^*(r)$ and $r<r^*(h)$).

The optimal policy can be determined by the following steps: (1) Solve $m_1(h)$ and $\mu_1(r)$ according to (16) and (17). (2) Based on those values, calculate the switching curves $r^*(h)$ and $h^*(r)$. (3) With these two switching curves, compare r with $r^*(h)$ and h with $h^*(r)$ to determine which action is optimal and calculate the corresponding optimal policy (m^*,μ^*).

The computation of the optimal policy involves numerically solving nonlinear Equations (16)–(18). For simple surge patterns, such as the rectangular or triangular surge demands, the nonlinear equations are easy to obtain. As a result, solving the optimal policy (m^*,μ^*) is equivalent to finding the unique zero point. Moreover, the nonlinear equations are monotone. Therefore, well established algorithms such as binary search can be applied here. For more complex surge demand patterns, the computation is more complex, requiring the generation of numerical approximations for the nonlinear equations. On the other hand, because we can always devise a convex combination of the simple surge demand patterns to approximate any complex surge demand pattern, we can use the solution of the former as a near optimal policy for the latter.

4.3.1. Equivalence of Problem Formulations. Finally, we verify that solving the optimal policy for the penalty cost problem B is equivalent to solving other formulations of the problem.

Proposition 2. *There exists a one-to-one increasing function $b(\alpha)$ such that the optimal solution (m^*,μ^*) to B under penalty cost $b(\alpha)$ achieves the same optimal solution under the service-level constraint α in problem A. Similarly, there exists a one-to-one increasing function $b(\bar{G})$ such that the optimal solution (m^*,μ^*) to B under penalty cost $b(\alpha)$ achieves the same optimal solution under the budget $\bar{G}$ constraint formulation.*

In other words, the three formulations of the problem—minimizing cost given a target service level, minimizing total cost given a shortage penalty cost, and maximizing service given a fixed budget—are equivalent.

5. Impact of Product and Demand Characteristics

We now conduct sensitivity analyses to better understand how the optimal planning policy depends on the product's cost structure and surge demand characteristics. We are interested in the effect of these factors on the absolute and relative levels of the optimal reactive capacity and safety stock. To study the relative levels, we define

$$\gamma_m=\frac{m^*}{m^*+E[D-B'(\mu^*)]},\qquad \gamma_\mu=1-\gamma_m.$$

Table 1 Summary of Product and Surge Demand Characteristics' Impact on the Optimal Planning Policy

	Impact on optimal planning policy		
Product/surge demand characteristic	SS level	RC level	Increase preference toward
Cost/objective			
Larger safety stock holding/purchasing costs	−	+	RC
Larger reactive capacity reservation/purchasing costs	+	−	SS
Larger budget/service level/shortage penalty	Depends	+	Depends
Magnitude			
Stochastically longer[a]	Depends	+	RC
More uncertain duration	Depends	+	RC
Stochastically more intense[a]	+	+	No change
More uncertain intensity	Depends	Depends	Depends
Shape			
Stochastically more compact	+	Depends	SS
Stochastically earlier peak position	+	Depends	SS
Stochastically more volatile	+	Depends	SS
Frequency			
More frequent surges	Depends	Depends	Depends

Notes. RC, Reactive capacity; SS, safety stock.
[a]Indicates maintaining Lorenz order equality.

Note that $E[D - B'(\mu^*)] = E[\mu^* T' + D(T) - D(T')]$ is the expected available inventory from reactive capacity, and $m^* + E[D - B'(\mu^*)]$ is the expected available total inventory to meet surge demand. Therefore, γ_m measures the proportion of expected available inventory one plans to procure from safety stock. As γ_m increases (or equivalently, as γ_μ decreases), we say that safety stock is increasingly attractive.

To foreshadow the findings of our sensitivity analyses, Table 1 provides a high-level summary of the results we will obtain, although we note that the nuances of some results are lost, and other results derived cannot be fully represented in the table. The general approach of our sensitivity analyses is to consider one metric/characteristic at a time, keeping the others constant. We also note that the effects summarized here are not mutually exclusive. For example, one can simultaneously face stochastically longer as well as less frequent surge demands. Our model can also be used to investigate the combination of multiple factors simultaneously on a case-by-case basis.

5.1. Product Characteristics

We first establish the impact of the cost parameters on the optimal policy.

Proposition 3. *μ^* is continuous and increasing in h, c, but decreasing in r, w, whereas m^* is continuous and decreasing in h, c, but increasing in r, w. Although μ^* is increasing in b, m^* and γ_m may be increasing or decreasing in b.*

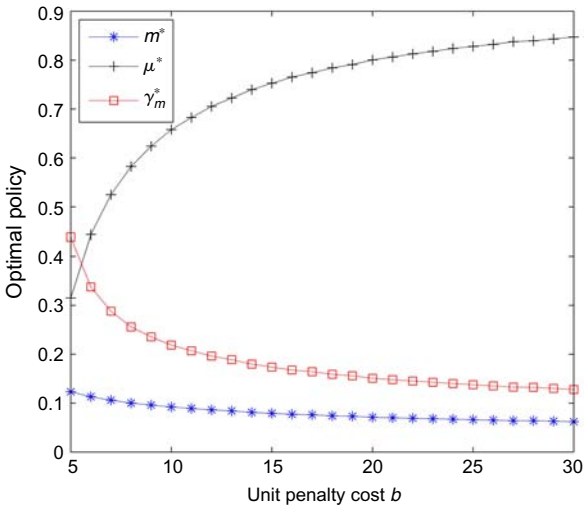
Parameters h, c, r, and w contain all cost information about the surge product, cost of production, and holding/reservation cost. Proposition 3 demonstrates that the optimal planning policy is affected by cost parameters in an intuitive way: each optimal level increases as its associated costs decrease relative to the other strategy. Specifically, one should increase safety stock m (and decrease reactive capacity μ) as the cost to hold safety stock h and the cost to procure safety stock c decrease. Similarly, one should increase reactive capacity μ (and decrease safety stock m) as the cost to reserve capacity r and the cost to produce from reactive capacity w decrease.

What is more interesting is that whereas μ^* is always increasing in the penalty cost b, the optimal safety stock level m^* may actually be decreasing in b. Put in another way, in light of Proposition 2, m^* may also be decreasing in α and $\bar{G}$; that is, it may be optimal to decrease the amount of stockpile if facing a higher service level requirement or a more generous budget. For an example of this phenomenon, see Figure 6. Here, X and $T \sim \text{Uniform}[0, 1]$, $\lambda = 0.05$, $c = 0.5$, $w = 1.7$, $h = 0.1$, $r = 0.04$, and λ is selected so that the ratio of the time spent in a surge $\tau/\nu = 2.5\%$. This result may seem somewhat surprising: if the cost of not meeting surge demand increases, why would one ever want to *decrease* the safety stock? The intuition is as follows: because the penalty cost is high, the planner seeks to obtain a higher service level. Although it may be cheaper to hold safety stock, reactive capacity may be better at securing a higher service level in a manner that outweighs the cost advantage of safety stock. We will examine the differences between the two strategies more closely in the rest of this section.

5.2. Demand Characteristics

Our demand model enables us to capture a variety of dimensions of surge demands as well as their uncer-

Figure 6 **(Color online) Example Illustrating the Possibility of the Optimal Safety Stock Level Decreasing in the Shortage Penalty Cost**



tainties. Recall from the introduction that there are several metrics by which we can characterize a realized surge trajectory: magnitude, duration, intensity, compactness, peak position, and volatility. Furthermore, from a planning perspective, surge demand trajectories are also *random*. Therefore, surge demands can also differ in terms of how *uncertain* one is about each metric. For example, the duration of one random surge demand can be more uncertain than the duration of another surge demand, even if the two surge demands have the same expected duration.

To examine such uncertainties, we will employ several notions of stochastic orders—the usual stochastic order, the convex order, the Lorenz order, and the dilation order. Note that with the convex order, if $Z \leq_{\text{cx}} Y$, then $E[Z] = E[Y]$. Thus, the convex order can only be used to compare the uncertainties of variables with the same mean. To compare the variables with different means, while holding uncertainty constant in a stochastic sense, we use two stochastic orders—Lorenz and dilation—that are less frequently used in the operations management literature. Roughly speaking, the Lorenz order is similar to comparing the coefficient of variances of two random variables. In fact, $Z =_{\text{Lorenz}} Y$ implies that Z and Y have equal coefficients of variance. We will also leverage the fact that if $Z \leq_{\text{st}} Y$ and $Z =_{\text{Lorenz}} Y$, then there exists $k \geq 1$ such that $Y \sim kZ$. Dilation order enables us to compare variables with different means by adding or subtracting a constant. If $Z \leq_{\text{st}} Y$ and $Z =_{\text{dil}} Y$, then there exists $k \geq 0$ such that $Y \sim Z + k$. See Muller and Stoyan (2002) for detailed information on these orders.

To investigate the first- and second-order effects of these metrics on the optimal planning policy, we organize metrics into two categories: those that affect the total *magnitude* of a surge and those that affect the *shape* of a surge. By stretching or shrinking trajectories in the time or demand axis while retaining the shape of trajectories, we first examine the effect of increasing or decreasing the magnitude of a surge by increasing its duration or intensity. Later, we fix the total magnitude of trajectories and examine the effects of changing the compactness, peak position, and volatility of surges.

5.2.1. Magnitude.

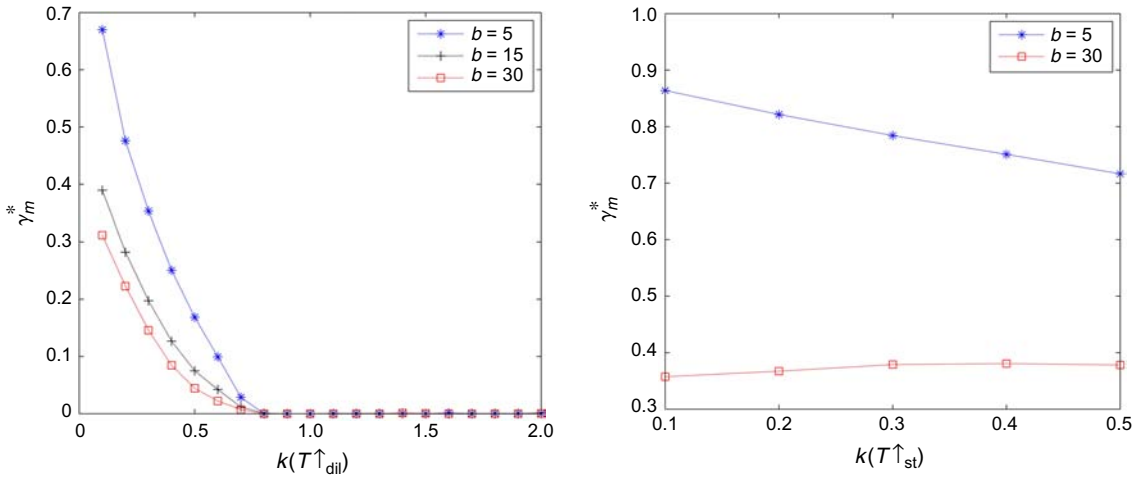
Duration. The following proposition shows the impact of a larger magnitude via a stochastically longer surge demand, controlling for duration uncertainty (using the Lorenz order) and the shape of the trajectory (by stretching trajectories along the time dimension).

Proposition 4. *Let T_1, T_2 be two surge durations such that $T_1 \leq_{st} T_2$ and $T_1 =_{\text{Lorenz}} T_2$. Let $X_1(t)$, $X_2(t)$ be their corresponding surge demands, with $X_2(t, \omega) = X_1((T_1(\omega)/T_2(\omega))t, \omega)$, $\forall\, \omega \in \Omega$. Then, safety stock is less attractive under T_2 than under T_1, i.e., $\gamma_m(T_2) \leq \gamma_m(T_1)$. Specifically, $\mu^*(T_2) \geq \mu^*(T_1)$, and there exists some $k \geq 1$ so that $T_2 \sim kT_1$ and $m^*(T_2) \leq km^*(T_1)$.*

The intuition behind Proposition 4 is as follows. Because reactive capacity is available during the entire surge duration, the expected unit reactive capacity cost $r\nu/\tau$ is lower for longer surge demands, leading to a larger μ^*. A stochastically longer duration also implies a larger total demand D, which may result in a higher optimal safety stock level, although this effect is smaller and bounded by $m^*(T_2) \leq km^*(T_1)$.

The problem structure bears a special relationship with the Lorenz order that makes the analysis tractable. However, we find similar results with numerical studies with other types of stochastically longer surge demands. Specifically, we consider the case when surge duration T increases under the dilation order. Under the dilation order, there exists $k \geq 0$ so that $T_2 \sim T_1 + k$. We set X and T to be Uniform[0, 1], and set $T = k + aX$, $0 \leq k \leq 1$, where k varies by increments of 0.1 and $a \in \{0.1, 1, 5\}$. We also vary λ to be $\{1/20,\ 1/10,\ 1/5,\ 1\}$ so that $p = \tau/\nu$ falls in [2.5%, 50%]. We consider the cost parameters $c \in \{0.1,\ 0.5,\ 1,\ 5\}$, $w \in \{0.1,\ 0.5,\ 1{,}5\}$, and $h \in \{0.01,\ 0.1,\ 0.5\}$. For each set of cost parameters, there exists $r_{\min}(h)$ and $r_{\max}(h)$ such that a joint strategy is optimal if and only if $r_{\min}(h) \leq r \leq r_{\max}(h)$. Because we want to focus on these cases, we choose r according to set $\{r_{\min}(h) + \frac{1}{4}(r_{\max}(h) - r_{\min}(h)),\ r_{\min}(h) + \frac{1}{2}(r_{\max}(h) - r_{\min}(h)),\ r_{\min}(h) + \frac{3}{4}(r_{\max}(h) - r_{\min}(h))\}$. We conduct a similar set of numerical experiments with exponential, normal, and log-normal demand distributions. All numerical results suggest that when T increases in dilation order, γ_m decreases, as shown in Figure 7 (left).

Figure 7 **(Color online) The Impact of Increasing Surge Magnitude via Stochastically Longer Duration in Dilation Order (Left) and an Counterexample Showing Stochastically Longer Duration Without Controlling for Duration Uncertainty (Right)**



Thus, we conclude that when surge duration increases, controlling for uncertainty, reactive capacity is more attractive. However, this is not necessarily the case when T increases without control of uncertainty. Figure 7 (right) shows a counter example. Here, $T \sim \text{Uniform}[k, 1]$, $k \in \{0, 0.2, 0.4\}$, $X \sim \text{Uniform}[0, 1]$, $\lambda = 0.05$, $c = 0.3$, $w = 0.8$, $r = 0.04$, and $h = 0.06$. In this example, as k increases, μ^* may decrease and γ_m may increase, especially when the required service level is high (b is large). The reason is that although longer surges increase the attractiveness of reactive capacity, second-order stochastic differences in durations may undermine that effect, as we will show next.

To investigate second-order effect of duration, we leverage the convex Order; that is, we study the impact of a more uncertain duration, controlling for the expected duration, magnitude, and surge trajectory shape.

Proposition 5. *Let $X_1(t)$ and $X_2(t)$ be two rectangular or triangular surge demands, where T_1 and T_2 are the respective surge durations such that $T_1 \le_{cx} T_2 = k(T_1 - E[T_1]) + E[T_2]$, $k > 1$. Safety stock is less attractive under T_2 than under T_1, i.e., $\gamma_m(T_2) \le \gamma_m(T_1)$. Moreover, $\mu^*(T_2) \ge \mu^*(T_1)$.*

Proposition 5 shows that reactive capacity becomes more favorable when the surge duration is more uncertain. The intuition is that in this case the risk of overage and underage associated with safety stock increases drastically compared to the risk associated with reactive capacity. Although we must assume a specific parametric form of surge demands to prove the result, we believe the result is likely to be more general. Some evidence along these lines is that if we define $\Delta = C(m_1, 0) - C(0, \mu_1)$ as the cost difference between a (forced) safety stock only strategy and a (forced) reactive capacity only strategy, then we can show that Δ is greater under T_2 than under T_1 if $T_2 \ge_{cx} T_1$.

If the durations are uniformly or normally distributed, then Proposition 5 holds for convex order. To further verify the robustness of our findings under convex order, we conduct numerical experiments for the log-normal distribution $T \sim \ln N$ with both rectangular and triangular surge demands, using parameters similar to those outlined earlier. These experiments yield consistent results: when T increases in the convex order, m^* decreases, but μ^* increases. Together with Proposition 5, these results suggest that reactive capacity is more advantageous when T is more variable in convex order.

Intensity. Next, we investigate the effect of a larger magnitude via a stochastically more intense surge demand, again controlling for intensity uncertainty (using the Lorenz order) and the shape of the trajectory (by stretching trajectories along the demand rate dimension). We obtain the following first-order effect.

Proposition 6. *Let $X_1(t)$ and $X_2(t)$ be surge demands such that $X_1(t) \le_{st} X_2(t)$ and $X_1(t) =_{Lorenz} X_2(t)$, $0 \le t \le T$. Then, safety stock and reactive capacity are equally attractive under $X_1(t)$ and $X_2(t)$, $\gamma_m(X_2) = \gamma_m(X_1)$. Specifically, there exists $k \ge 1$ such that the optimal hedging policy $(m^*(X_2), \mu^*(X_2)) = (km^*(X_1), k\mu^*(X_1))$.*

Proposition 6 states that when the intensities of surge demands increase, keeping Lorenz order equality (thus, keeping a constant coefficient of variation), the optimal stockpile and capacity levels increase proportionally. This reveals a special relationship between the structure of the optimal policy and surge demand uncertainty. If $X(t)$ increases in another fashion, e.g., by a constant, the change in the optimal policy may differ from Proposition 6. The underlying logic is that the extra kth unit of intensity

can be satisfied by reactive capacity without causing extra uncertainty. However, this extra kth unit of demand will incur extra kT demand uncertainty for safety stock. Thus, what strategy to favor depends on the trade-off between unit cost differences and this uncertainty difference. When the cost difference $w-c$ is small or when the required service level α is low, the unit cost difference becomes less important and reactive capacity's attractiveness increases. Otherwise, safety stock's attractiveness increases. Numerical experiments (omitted here) confirm the same trends under uniform, normal, exponential, and lognormal distributions with rectangular and triangular surge demands. If intensity uncertainty grows without changing the expected intensity, numerical experiments (omitted here) indicate no clear trend for strategy preference.

5.2.2. Shape. We now investigate the effect of metrics that have to do with the shape of a demand trajectory, controlling for the total magnitude of the demand surge. While fixing total expected demand, we conduct stochastic comparisons to reveal the effect of compactness, peak position, and volatility.

Compactness. First, we address the following question: How does the optimal policy change when the surge demand is more or less "compact" (i.e., more demand in a smaller amount of time)?

Proposition 7. *Let T_1 and T_2 be surge durations such that $T_2 \sim kT_1$, and define $X_2(t) = X_1(t/k), 0 \le t \le T_2$. Then, $\gamma_m(X_1, T_1) \ge \gamma_m(X_2, T_2)$. Specifically, $m^*(X_2, T_2) \le m^*(X_1, T_1)$, $k\mu^*(X_2, T_2) \ge \mu^*(X_1, T_1)$.*

Proposition 7 suggests that reactive capacity is preferred when the surge demand is of low intensity but has a long duration (keeping the total expected demand constant). On the other hand, quick but intense surges favor the safety stock strategy. This result is consistent with the intuition that reactive capacity is more valuable for surge demands that allow more time for production, while safety stock is more valuable for surges that demand an immediate response.

Peak position. Second, we consider the effect of peak position location (controlling for expected magnitude, duration, and intensity). Recall that demand process $\{X(t, \omega), 0 \le t \le T(\omega)\}$ is continuous and unimodal. Assume $X(t, \omega)$ is increasing in time t for each given ω. Construct the following demand group:

$$X'(t, \beta, \omega) = \begin{cases} X\left(\dfrac{t}{\beta}, \omega\right), & 0 \le t \le \beta T(\omega), \\ X\left(T - \dfrac{t-\beta T}{1-\beta}, \omega\right), & \beta T(\omega) < t \le T(\omega). \end{cases}$$

Here β captures the peak position: when $\beta = 0$, $X'(t, \beta, \omega) = X(T - t, \omega)$, the demand is decreasing and peaks at 0; when $\beta = 1$, $X'(t, \beta, \omega) = X(t, \omega)$, the demand is increasing and peaks at T. This construction isolates the effect of the peak position. In particular, the total demand $D(\beta, \omega)$ is constant with respect to β for each given ω. We have the following:

Proposition 8. *Let β_1 and β_2 be two peak times for surge demand $\{X'(t, \beta, \omega), 0 \le t \le T(\omega)\}$ such that $\beta_1 \le \beta_2$. Safety stock is more attractive under earlier peak times, $m^*(\beta_1) \ge m^*(\beta_2)$ and $\gamma_m(\beta_2) \le \gamma_m(\beta_1)$.*

Note that given a fixed total demand, if m^* decreases while the service level α^* increases, then μ^* must increase. Thus, Proposition 8 suggests that when surge demand peaks later, reactive capacity is more favorable. Interestingly, our numerical experiments show that it is possible for the optimal reactive capacity level to decrease if the unit penalty cost b is not large enough. This is because when the peak position is later, the demand ramps up slower, which makes the demand easier to satisfy from reactive capacity. If the service level α^* is low, peaking later gives reactive capacity a chance to reduce costs by reducing the ordering quantity while still maintaining a greater service level.

Volatility over duration. Third, we consider the volatility of surge demand by maintaining the same total demand while adjusting the demand variance over its duration. To capture volatility in a unimodal setting while controlling for magnitude, we assume the surge demand is a linear combination of rectangular and triangular surge demands. Consider a triangular surge demand $\{X(t, \omega), 0 \le t \le T(\omega)\}$ with zero end points $X(0, \omega) = X(T(\omega), \omega) = 0$, and define $X'(t, \rho, \omega) = X_1(t, \rho, \omega) + X_2(t, \rho, \omega)$, where $T_1(\omega) = T_2(\omega) = T(\omega)$, $X_1(t, \rho, \omega)$ is a rectangular surge demand and $X_2(t, \rho, \omega)$ is a triangular surge demand with $X_1(t, \rho, \omega) = (\rho/2)X(t_p(\omega), \omega)$, $X_2(t, \rho, \omega) = (1-\rho)X(t, \omega)$. Thus, $0 \le \rho \le 1$ captures the volatility of demand over duration. As ρ decreases, the total surge demand $D(\rho, \omega)$ remains constant, but the severity of the peak increases. In the extremes, this yields a rectangular surge demand with $\rho = 1$ and triangular surge demand with $\rho = 0$.

Proposition 9. *Let ρ_1 and ρ_2 be two volatility parameters for $X'(t, \rho, \omega)$ such that $\rho_1 \le \rho_2$. Safety stock is more attractive when the demand is more volatile $m^*(\rho_1) \ge m^*(\rho_2)$ and $\gamma_m(\rho_2) \le \gamma_m(\rho_1)$.*

This result implies that one should favor reactive capacity when surge demand is less volatile, but favor safety stock when surge demand is more volatile. The intuition is that safety stock provides more flexibility to meet sudden increases in demand, while reactive capacity increases in cost efficiency when demand is stable.

5.2.3. Frequency. Finally, we investigate the impact of surge frequency; that is, without changing any metrics of the demand surges themselves, what is the effect of making surges more or less frequent on the optimal planning policy?

For tractability, we focus on rectangular surge demands to derive analytical results. We define the ratio

$$\kappa = \frac{r(b-c)}{h(b-w)\tau} = \frac{r(v+\tau)}{(b-w)\tau} \Big/ \frac{h(v+\tau)}{b-c},$$

which captures the cost difference between safety stock and reactive capacity. Note that κ is increasing in r and w, but decreasing in h and c. When $\kappa > 1$, reactive capacity has a higher overage/underage cost ratio and hence is, in a sense, more expensive than safety stock.

Proposition 10. (i) *If* $\kappa \le 1$, *then* $\gamma_m = 0$ *for small* λ *and* $\gamma_m = 1$ *for large* λ. (ii) *If* $\kappa > 1$, *then* $\gamma_m = 1$ *for* λ *large or small, and* $\gamma_m < 1$ *otherwise*.

Proposition 10 states that if the surge is either rare enough or frequent enough, a single hedging strategy is optimal. The intuition is as follows: the surge frequency primarily impacts the ratio of time spent between the surge and the normal state. Thus, the normal state cost difference (i.e., the difference between inventory holding and reactive capacity costs) will dominate when λ is small, but the surge state cost difference (i.e., the difference between purchasing costs and shortage costs) will dominate when λ is large. Thus, if $\kappa \le 1$, when λ is small, the normal state cost difference dominates and safety stock is not used ($\gamma_m = 0$). On the other hand, when λ is big, the surge-state cost difference dominates and the reactive capacity is not used ($\gamma_m = 1$).

To further understand the impact of surge frequency and to complement Proposition 10, we conduct numerical experiments under four different types of distribution of X and T. We assume $X \sim$ Uniform$[0, 1]$, $T \sim a \cdot$ Uniform$[0, 1]$, and the same parameters for c, w, r, b, and a as in Subsection 5.2.1. For surge frequency λ, we similarly choose $p = \lambda/(\lambda+1)$ as in Subsection 5.2.1. Finally, to control for the impact of κ, we specifically choose parameter h such that $\kappa \in \{0.2, 0.4, 0.6, 0.8, 1, 2, 5, 10, 50\}$. We find that when $\kappa < 1$, γ_m is increasing in λ. This is because as λ increases, the safety stock's cost advantage in the surge state (because $c < w$) increases while its cost disadvantage in the normal state (because $\kappa < 1$) decreases. On the other hand, we find that when $\kappa \ge 1$, γ_m is typically first decreasing and then increasing in λ. This nonmonotonicity may seem counterintuitive. The key observation is that when surges are extremely rare or extremely common, the difference in cost parameters will tend to determine the strategy. When surges are extremely rare, the difference in h and r (which favor safety stock when $\kappa > 1$) will be the determining factor. When surges are extremely common, the difference in c and w (which also tends to favor safety stock when $\kappa > 1$) will be the determining factor. For intermediate surge demand rates, however, the safety stock's cost advantage in the surge state (because $c < w$) increases while its cost advantage in the normal state (because $\kappa \ge 1$) decreases. The behavior of γ_m is determined by which cost difference dominates.

6. Concluding Remarks

By developing a new demand model that captures demand surges and formulating an optimization problem with two important yet fundamentally different strategies to prepare for them, this paper provides an analytical framework to guide supply chain planning and design. The sample-path feature of the demand model also sheds light on how companies can leverage big data for coping with demand surges. We derived two switching curves that determine the optimal combination of reactive capacity and safety stock and show how to evaluate them through a system of equations. Our sensitivity analyses show the impact of the cost profile of each strategy as well as several first- and second-order characteristics of surge demands on the optimal joint strategy. Together, these results provide a set of insights and managerial guidelines for demand surge planning across different industries.

We conclude with some possible extensions and some opportunities for future research. First, we can extend the model to incorporate multiple normal and multiple surge states under certain assumptions. Multiple normal states can capture different levels of risks of entering different types of surges, each with its own set of possible sample paths. If the planner is well informed about the system state, and demand surge is rare enough to allow planning policy adjustments in every normal state, then a state dependent planning policy can be solved in the same way as in the main model by expanding the set Ω. At the other extreme, if the planning policy cannot be changed over many demand surges, then a long-run state-independent planning strategy can also be solved in the same way using steady state analysis. Other extensions of the model may require significant further analyses. For example, we make Assumption 1 primarily so that the details of the optimal deployment policy during a surge demand are simplified and we can focus on the planning policy. However, it is not always the case that forecast updating adheres to this assumption, and one could investigate a more sophisticated dynamic program that simultaneously

updates demand forecasts to make optimal deployment decisions. Similarly, one could add more features and caveats associated with reactive capacity and safety stock (e.g., fixed lead times, ramp-up times for safety stock, etc.). Such analysis would be useful in that it would aid practitioners in determining the exact reactive capacity and safety stock levels. We hope that future research will be able to leverage our model to pursue these directions.

Acknowledgments

The authors would like to thank Paul Zipkin, Fernando Bernstein, Robert Swinney, and Peng Sun for their valuable conversations and feedback. Additionally, the authors are grateful for the helpful comments and questions from the seminar participants at 2014 MSOM and POMS, Singapore University of Technology and Design, and Fudan University. Finally, the authors thank the editor as well as the anonymous associate editor and reviewers for their constructive feedback. This research was partially supported by the Chinese Natural Science Foundation [Grant 71390331].

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2016.0583.

References

Abhyankar H, Graves S (2001) Creating an inventory hedge for Markov-modulated Poisson demand: An application and model. *Manufacturing Service Oper. Management* 3(4):306–320.

Anupindi R, Akella R (1993) Diversification under supply uncertainty. *Management Sci.* 39(8):944–963.

Aviv Y, Federgruen A (1997) Stochastic inventory models with limited production capacity and periodically varying parameters. *Probab. Engrg. Informational Sci.* 11(1):107–135.

Babich V, Burnetas AN, Ritchken PH (2007) Competition and diversification effects in supply chains with supplier default risk. *Manufacturing Service Oper. Management* 9(2):123–146.

Bhat S, Krishnamurthy A (2014) Dynamic control policies for multiclass systems with seasonal demands. Working paper, University of Wisconsin–Madison, Madison.

Chen F, Song J-S (2001) Optimal policies for multi-echelon inventory problems with Markov-modulated demand. *Oper. Res.* 49(2):226–234.

Chopra S, Reinhardt G, Mohan U (2007) The importance of decoupling recurrent and disruption risks in a supply chain. *Naval Res. Logist.* 54(5):544–555.

Ciarallo FW, Akella R, Morton TE (1994) A periodic review, production planning model with uncertain capacity and uncertain demand: Optimality of extended myopic policies. *Management Sci.* 40(3):320–332.

Donohue KL (2000) Efficient supply contracts for fashion goods with forecast updating and two production modes. *Management Sci.* 46(11):1397–1411.

Eppen GD, Iyer AV (1997) Backup agreements in fashion buying—The value of upstream flexibility. *Management Sci.* 43(11):1469–1484.

Erdem O, Oumlil AB, Tuncalp S (1999) Consumer values and the importance of store attributes. *Internat. J. Retailing Distribution Management* 27(4):137–144.

Ergun O, Keskinocak P, Swann J, Villarreal M (2012) Disaster preparedness and retail operations. *Handbook of Operations Research for Homeland Security* (Springer, New York).

Federgruen A, Yang N (2009) Optimal supply diversification under general supply risks. *Oper. Res.* 57(6):1451–1468.

Fisher M, Raman A (1996) Reducing the cost of demand uncertainty through accurate response to early sales. *Oper. Res.* 44(1):87–99.

Henig M, Gerchak Y (1990) The structure of periodic review policies in the presence of random yield. *Oper. Res.* 38(4):634–643.

Henig M, Gerchak Y, Ernst R, Pyke DF (1997) An inventory model embedded in designing a supply contract. *Management Sci.* 43(2):184–189.

Hou J, Zeng AZ, Zhao L (2010) Coordination with a backup supplier through buy-back contract under supply disruption. *Transportation Res. Part E* 46(6):881–895.

Iglehardt D, Karlin S (1962) Optimal policy for dynamic inventory process with nonstationary stochastic demands. Arrow K, Karlin S, Scarf H, eds. *Studies in Applied Probability and Management Science* (Stanford University Press, Stanford, CA), 127–147.

Kapuściński R, Tayur S (1998) A capacitated production-inventory model with periodic demand. *Oper. Res.* 46(6):899–911.

Manjoo F (2013) The only problem with the greatest hoodie ever made. *Slate* (March 21), http://www.slate.com/articles/technology/technology/2013/03/american_giant_hoodie_the_only_problem_with_the_world_s_greatest_sweatshirt.html.

Muharremoglu A, Tsitsiklis JN (2008) A Single-unit decomposition approach to multi-echelon inventory systems. *Oper. Res.* 56(5):1089–1103.

Muller A, Stoyan D (2002) *Comparison Methods for Stochastic Models and Risks* (John Wiley & Sons Inc., Chichester, UK).

National Association of State Procurement Officials (2013) Emergency preparedness for state procurement officers. Accessed September 1, 2015, http://www.naspo.org/dnn/portals/16/documents/EmergencyPreparednessforStateProcurementOfficials.pdf.

Piger J (2009) Econometrics: Models of regime changes. Meyers RA, ed. *Encyclopedia of Complexity and Systems Science* (Springer, New York), 2744–2757.

Parlar M, Perry D (1996) Inventory models of future supply uncertainty with single and multiple suppliers. *Naval Res. Logist.* 43(2):191–210.

Saghafian S, Oyen VM (2012) The value of flexible suppliers and disruption risk information: Newsvendor analysis with recourse. *IIE Trans.* 44(10):834–867.

Sethi S, Cheng F (1997) Optimality of (s, S) policies in inventory models with Markovian demand. *Oper Res.* 45(6):931–939.

Sheffi Y (2005) *The Resilient Enterprise: Overcoming Vulnerability for Competitive Advantage* (MIT Press, Cambridge, MA).

Shontell A (2010) 10 products that went crazy viral and became marketing phenomena. *Business Insider* (September 29), http://www.businessinsider.com/10-products-that-went-crazy-viral-and-became-youtube-phenomenons-2010-9#.

Song J-S, Zipkin P (1993) Inventory control in a fluctuating demand environment. *Oper. Res.* 41(2):351–370.

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

Tomlin B (2006) On the value of mitigation and contingency strategies for managing supply-chain disruption risks. *Management Sci.* 52(5):639–657.

Van Wassenhove LN (2006) Humanitarian aid logistics: Supply chain management in high gear. *J. Oper. Res. Soc.* 57(5):475–489.

Yarow J (2014) Apple's iPhone 6 sales are going to be stronger for even longer than people are expecting. *Business Insider* (November 20), http://www.businessinsider.co.id/analyst-raises-price-target-2014-11/#.VOvMx_nF9qI.