Manufacturing & Service Operations Management

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

Strategic Safety Stocks in Supply Chains with Evolving Forecasts

Tor Schoenmeyr, Stephen C. Graves,

To cite this article:
Tor Schoenmeyr, Stephen C. Graves, (2009) Strategic Safety Stocks in Supply Chains with Evolving Forecasts. Manufacturing & Service Operations Management 11(4):657-673. http://dx.doi.org/10.1287/msom.1080.0245

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 11, No. 4, Fall 2009, pp. 657–673
ISSN 1523-4614 | EISSN 1526-5498 | 09 | 1104 | 0657
informs
DOI 10.1287/msom.1080.0245

Strategic Safety Stocks in Supply Chains with Evolving Forecasts

Tor Schoenmeyr
OptiSolar, Inc., Hayward, California 94544, tschoenmeyr@optisolar.com
Stephen C. Graves
Leaders for Manufacturing Program and A. P. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, sgraves@mit.edu

We examine the placement of safety stocks in a supply chain for which we have an evolving demand forecast. Under assumptions about the forecasts, the demand process, and the supply chain structure, we show that safety-stock placement for such systems is effectively equivalent to the corresponding well-studied problem for systems with stationary demand bounds and base-stock policies. Hence, we can use existing algorithms to find the optimal safety stocks. We use a case study with real data to demonstrate that there are significant benefits from the inclusion of the forecast process when determining the optimal safety stocks. We also conduct a computational experiment to explore how the placement and size of the safety stocks depend on the nature of the forecast evolution process.

Key words: evolving forecast; safety stock; supply chain; MRP; guaranteed service
History: Received: March 14, 2008; accepted: September 26, 2008. Published online in *Articles in Advance* April 8, 2009.

1. Introduction

Most firms plan their supply chain operations based on a forecast of future demand over some planning horizon. Furthermore, firms regularly update and revise these forecasts based on observed sales, advanced orders, and market intelligence. With each forecast revision, a firm revises its supply chain plans, in terms of its master schedules for production, procurement, and transportation. Indeed, this update and revision process is central to any supply chain planning function and is facilitated by the widespread deployment of material requirements planning (MRP) systems.

The intent of this paper is to examine the optimal placement of safety-stock inventory in a supply chain that is subject to a dynamic, evolving demand forecast. In particular we strive to develop models and algorithms that have the potential to determine safety stocks in real-world supply chains. We assert that this paper makes five contributions.

First, we incorporate a forecast evolution process into the safety-stock placement models developed by Simpson (1958) for a serial-system supply chain, and by Graves and Willems (2000, 2003) for supply chains with spanning-tree topologies. In particular, we use the forecast evolution process and model that has been used previously by Graves et al. (1986), Heath and Jackson (1994), and Graves et al. (1998).

Second, we show for a serial-system supply chain with an evolving forecast that the optimal placement of safety stocks satisfies the all-or-nothing property, i.e., each stage either holds a decoupling safety stock or no safety stock. Thus, we can determine the optimal safety stocks by enumeration, as suggested by Simpson (1958), or by an efficient dynamic programming algorithm (Graves and Willems 2000).

Third, for an assembly supply chain with an evolving forecast, we show that its safety-stock optimization problem has the same structure as the safety-stock optimization for an assembly system operating with a base-stock ordering policy. Graves and Willems (2000) developed an efficient dynamic programming algorithm to determine the optimal safety stocks for this latter system. Thus, we can also use this algorithm to solve the safety-stock optimization problem for assembly systems with an evolving forecast.

Fourth, based on an industrial study and on a computational experiment, we demonstrate the potential

value of incorporating the forecast evolution process into the safety-stock optimization. We find that substantial reductions in inventory are possible, where the size of the reduction depends on how the forecast improves over time; not surprisingly, the better the forecast, the less safety stock is required. However, prior safety-stock optimization methods were unable to extract the value from an improving forecast.

Fifth, we demonstrate that we can use our forecast evolution process to model a wide class of demand processes introduced by Aviv (2003); for instance, this class includes autoregressive integrated moving average (ARIMA) processes. The significance of this result is that all of the developments in this paper also apply to a supply chain whose product demand comes from one of these demand processes. For instance, consider an assembly system that is subject to demand from an ARIMA (p, d, q) process for any specification of the parameters (p, d, q). We can infer a forecast process for this supply chain and then use this forecast process to determine the supply chain safety stocks, using the models and methods developed in this paper.

This paper consists of seven sections. In the remainder of this section, we provide a brief review of related literature. In §2, we introduce the forecast evolution process and show the equivalence between it and a class of demand processes introduced by Aviv (2003). In §3, we define the ordering policy for a supply chain with an evolving forecast, and use this to model the inventory dynamics for a serial-system supply chain. We also establish the safety stock required at each stage to satisfy the guaranteed service constraint. In §4, we establish the all-or-nothing property for the optimal solution for a serial system, and show how to determine the safety stocks for an assembly system. We report on an industrial case study in §5, and on a set of computational experiments in §6. We conclude the paper in §7. An online appendix provides the detailed development for several of the results in this paper.

1.1. Literature Review

This paper adds to the literature on multiechelon inventory systems. For a general overview of this research area, we refer to review articles by Axsäter (1993), Federgruen (1993), Inderfurth (1994), and Diks et al. (1996). This paper contributes to three bodies of work in particular.

First, our model uses a dynamic model for forecast evolution, and is related to other work on forecasting and advanced demand information in supply chains. Our forecast evolution model (§2) is a generalization of the one used by Graves et al. (1986), Heath and Jackson (1994), and Graves et al. (1998). We show that there is a close relationship between this forecast model and popular time-series demand models, such as ARIMA. Therefore, this paper is related to the growing body of work that assumes such demand models in supply chains. In particular, the demand model we use is based on the framework introduced by Aviv (2003). We refer the reader to Zhang (2004) for results and references on supply chain dynamics, Aviv (2004) for an overview of forecasts and collaboration, and Gallego and Özer (2001), Karaesmen et al. (2002), and Dong and Lee (2003) for results on the value of advanced demand information in the supply chain.

Second, the underlying supply chain model (§3) and optimization procedure (§4) follow closely the work of Simpson (1958) and Graves and Willems (2000). These authors assume that each stage or node of the supply chain operates under a base-stock policy with bounded demand and guaranteed service. They then find the least-cost service times and inventory placement that are guaranteed to meet any demand realization within these bounds. This approach provides a way to find the optimal safety stocks in quite general supply chains, and it has been successfully deployed to industry (e.g., Billington et al. 2004, Willems 2008).

In the aforementioned work, the authors assume stationary demand (or more accurately, a stationary demand bound), and assume that orders are placed according to base-stock policies in response to realized demand. In a recent paper, Graves and Willems (2008) consider the optimization of safety stock in a supply chain with a nonstationary demand bound and with guaranteed service. They assume an adaptive base-stock policy and demonstrate the effectiveness of a constant-service-time policy in which the locations of the safety stocks are fixed, but the amount of safety stock varies with the demand bound. They also show that the determination of the safety-stock locations, under the assumption of constant service times, is equivalent to the optimization

for the stationary demand case studied by Graves and Willems (2000).

Our work shares many important aspects with this line of research, but one significant difference is that we assume that each stage places orders in response to an evolving forecast of future demand. As noted, the aforementioned work generally assumes that the stages operate according to (local) base-stock policies by which orders are placed in response to realized demand at the customer-facing stages. In our work, the guaranteed service is made contingent on forecast errors, rather than on demand variation. Because of these differences, the new safety-stock strategies are more applicable for firms that already operate in a forecast- or schedule-driven way, and who seek a comprehensive safety-stock strategy.

Third, our assumed ordering policy is similar to that for an MRP system; thus, we can relate our work to the research literature on safety stocks in MRP systems. For an overview of MRP literature in general, see Baker (1993). Guide and Srivastava (2000) have reviewed buffering in particular, and list a comprehensive table of various approaches and results.

More specifically, Lambrecht et al. (1984) propose a dynamic programming algorithm for determining the safety-stock levels; however, because this algorithm is computationally infeasible for realistic problem sizes, they develop and test an extension to the approximate procedure from Clark and Scarf (1962). Buzacott and Shanthikumar (1994) analyze and compare safety stocks and safety times in a single-stage system. Yano and Carlson (1987) consider a two-stage system under either fixed or flexible scheduling; the analysis in this paper corresponds most closely to flexible scheduling. Lagodimos and Anderson (1993) consider the maximum service level achievable for a given safety-stock quantity. Molinder (1997) studies a number of systems using simulation, and finds optimal solutions with simulated annealing.

Most of the aforementioned work is limited to small systems, typically one or two stages. The lack of solutions for larger systems in particular has been highlighted in the overview paper by Guide and Srivastava (2000). Moreover, these papers do not model dynamically evolving forecasts and nonstationary demand.

We note that the ordering policy in our paper is a special case of the class of policies considered by Graves et al. (1998), who model a supply chain with a dynamically evolving forecast and with an objective to smooth operations to reduce the variability of production. However, Graves et al. (1998) do not attempt to optimize the supply chain safety stocks, which is the primary goal of our paper. Similar to Graves et al. (1998), Aviv (2007) develops a model of a two-stage supply chain with a dynamically evolving forecast. He also incorporates production smoothing and schedule/forecast changes into his objective function. But Aviv's (2007) emphasis is understanding the benefits of collaborative forecasting. In contrast, we assume an ordering policy and accept the resulting variability from the induced schedule changes, and seek only to reduce the safety-stock costs across the supply chain.

2. Forecast Model

We use a forecast evolution model based on Graves et al. (1986) and Heath and Jackson (1994). In period t we denote the forecast for period $t+i$ as $f_t(t+i)$ for $i \in \{1, 2, \ldots, H\}$, where H is the forecast horizon. By convention, we set $f_t(t) = D_t$, where D_t is the demand in period t. We will not make any notational distinction between D_t for future times, which are random variables, and for past times, which are realized scalar values. We assume that in each period t we make an initial forecast for the demand in period $t+H$, that is $f_t(t+H)$; we also assume that in each period we revise the nearer-term forecasts, where we define the *forecast revision* as

$$\Delta f_t(t+i) = f_t(t+i) - f_{t-1}(t+i) \quad \text{for } i \in \{0, 1, \ldots, H-1\}.$$

We can express demand as follows:

$$D_t = f_{t-H}(t) + \sum_{i=1}^{H} \Delta f_{t-H+i}(t). \tag{1}$$

We let $\underline{\Delta f}_t$ be the vector of H forecast revisions. We assume that $\underline{\Delta f}_t$ is a random, independent and identically distributed (i.i.d.) vector with $\mathrm{E}[\Delta f_t(j)] = 0$ for all t and j. With these assumptions, Graves et al. (1986), Heath and Jackson (1994), and Graves et al. (1998) have established several properties for this forecast evolution model: $f_t(t+i)$ is a martingale; $f_t(t+i)$ is an unbiased estimate of D_{t+i}; and the variance of the forecast error $(D_{t+i} - f_t(t+i))$ increases in i. Furthermore, the variance of the random variable

D_t is the trace of the covariance matrix for $\underline{\Delta f}_t$, which we denote by Σ.

The aforementioned papers assumes that the initial forecast is $f_t(t+H)=\mu$ for all t. Under this assumption, the demand process D_t has mean μ and variance given by the trace of Σ. We depart from this earlier work in that we do not make *any* assumptions about $f_t(t+H)$. In particular, we permit $f_t(t+H)$ to be generated by a nonstationary process of arbitrary complexity, or to be user-specified.

Thus, we can apply this forecast model to contexts in which the initial forecast $f_t(t+H)$ contains information of future orders or advanced demand information. For instance, consider the planning process used at Teradyne, Inc., a manufacturer of semiconductor test equipment with which we have worked (see also Abhyankar and Graves 2001 for more about Teradyne's planning process). For many of its product lines, Teradyne is a make-to-order operation. But the supply chain lead time (the longest procurement time for a piece part plus the internal assembly and test lead times) exceeds the customer lead time (the delivery lead time requested by customers). Hence, Teradyne must plan much of its procurement and upstream production activities prior to receiving an order. Teradyne does this by means of a master production schedule (MPS) that covers a planning horizon that corresponds to the length of the supply chain lead time. In effect, this MPS is its demand forecast. At any point in time, the master schedule consists of a mix of open orders, identified orders, and booked orders. An open order corresponds to a traditional forecast of what the sales force plans to sell, an identified order is associated with a potential customer and is based on some preliminary discussions with the customer, and a booked order is a firm customer order. As time moves forward, an open order is converted into an identified order as the sales force obtains tentative commitments and product specifications from a customer. Similarly, an identified order is converted into a booked order once (and if) the product specifications and due date become a firm order.

From our experiences, this process describes many make-to-order firms. In these cases, the initial forecast conveys the progress at identifying customers and in securing advanced orders. Subsequently, the forecast revisions correspond to changes of the master schedule, which reflect the success at converting the forecast (open orders) into demand (booked orders).

2.1. Relationship with Demand Models

In the previous section, we defined a forecast process and discussed how the framework arises in practice. Moreover, because $D_t = f_t(t)$, defining a forecast process also gives us a demand process; that is, the specification of the covariance matrix Σ and the initial forecast $f_t(t+H)$ determines a demand process D_t.

In this section, we start with a demand model D_t and show that we can infer a forecast process by setting the forecast to the expected value of demand. That is, for a given demand model D_t, we set

$$f_t(t+s) \equiv \mathrm{E}[D_{t+s} \mid D_t, D_{t-1}, D_{t-2}, \ldots], \quad \text{and}$$
$$\Delta f_t(t+s) \equiv \mathrm{E}[D_{t+s} \mid D_t, D_{t-1}, D_{t-2}, \ldots] - \mathrm{E}[D_{t+s} \mid D_{t-1}, D_{t-2}, \ldots].$$

A question of interest is whether the forecast revisions generated in this way are i.i.d. and have a mean of zero, because these assumptions were made in the forecast evolution model, and in fact are necessary for the supply chain work to follow.

We find that these i.i.d. and a mean of zero properties hold quite generally. In particular, suppose that we model demand by the general state space framework proposed by Aviv (2003):

$$\begin{aligned} X_t &= FX_{t-1} + V_t, \\ \Psi_t &= HX_t, \\ D_t &= \mu + R\Psi_t, \end{aligned} \tag{2}$$

where X_t is the state vector (with a dimension that depends on the complexity of the demand model), Ψ_t the vector of observations, F, H, and R constant matrices, and V_t an i.i.d., multivariate random variable with a mean of zero. Demand can, in general, also be a vector (if there are multiple demand streams), but presently we will consider the case where demand is scalar and R is a row vector. Assume further that the system is *observable*, which, loosely speaking, means that the system state X_t can be inferred from the observations Ψ_t, or more specifically for the model (2),

that $\mathrm{E}[X_t \mid \Psi_t] = X_t$. Then, we show in the online appendix that

$$\begin{aligned}\Delta f_t(t+s) &= \mathrm{E}[D_{t+s} \mid \Psi_t, \Psi_{t-1}, \ldots] \\ &\quad - \mathrm{E}[D_{t+s} \mid \Psi_{t-1}, \Psi_{t-2}, \ldots] \\ &= RHF^{s+1}V_t. \end{aligned} \tag{3}$$

Because V_t is i.i.d. with a mean of zero, so is $\Delta f_t(t+s)$ and thus, $\underline{\Delta f}_t$ is i.i.d. with a mean of zero. Given the covariance matrix for V_t, we easily find from (3) the covariance matrix for the forecast revision $\underline{\Delta f}_t$. We note that most common time-series models of demand, including ARIMA models, can be written in this state space framework and are observable. Thus, we find that even demand models that are quite complex and nonstationary often have i.i.d. forecast revisions. Also, we will see that whereas the initial forecast $f_t(t+H) = \mathrm{E}[D_{t+H} \mid D_t, D_{t-1}, D_{t-2}, \ldots]$ might be quite complex, this has no bearing on our safety-stock analysis.

This equivalence between the forecast evolution model and this broad class of demand models means that we can apply our results, for example, to a make-to-stock supply chain with a time-series demand model. We refer the reader to the specialized literature (for example, Hamilton 1994) on how to estimate a time-series demand model based on historical data. Once this is done, one can use (3) to find the properties of $\Delta f_t(t+s)$, which are needed for the safety-stock optimization to follow.

3. Supply Chain Model and Ordering Policy

In this section we develop our inventory model. This model is closely related to the supply chain model developed by Simpson (1958) for a serial system. Simpson (1958) assumes that each node in the supply chain operates with guaranteed service times. He then assumes that each node uses a base-stock ordering policy, which allows him to model the inventory at each node as a function of the service times. Simpson then introduces an assumption of bounded demand, which permits him to find the minimum base-stock levels necessary to assure that the service times are guaranteed.

We consider the same system but with an evolving forecast process. We first state the assumption of guaranteed service times, which is the same as in Simpson (1958). We then present and discuss a forecast-based ordering policy, rather than a simpler base-stock policy for each node. The forecast-based ordering policy is parameterized by a safety-stock target, as opposed to a base-stock level for the Simpson (1958) model. Given this ordering policy and the evolving forecast, we can then model the inventory at each node as a function of the service times. To determine the minimum safety-stock target that will guarantee the service times, we introduce another assumption that is analogous to Simpson's (1958) assumption of bounded demand. In particular, we assume that we have a bound on the forecast revision process. This allows us to model the safety stocks as a function of the guaranteed service times for each node in the system.

We develop the model for a serial-system supply chain, and we discuss the extension to assembly systems in the optimization section. For a serial system, we index the nodes by k, and we designate the customer-facing stage as node 1, and the most upstream stage as node N. A stage might represent the procurement of a raw material, the production of a component, the manufacture of a subassembly, the assembly and test of a finished good, or the transportation of a finished product from a distribution center to a warehouse. Each stage k is a potential location for holding a safety-stock inventory of the item processed at the stage.

For each stage, we assume a known deterministic production lead time, denoted as T_k. When a stage reorders, the production lead time is the time from when all of the inputs are available until production is completed and available to serve demand. The production lead time includes the waiting and processing time at the stage, plus any transportation time to put the item into inventory. We assume that there is no capacity constraint and, thus, the lead time is not affected by the size of the order.

3.1. Ordering Policy

We assume that each stage shares a common review period and places an order in each period. We denote the order placed in period t on stage $k+1$ by stage k as $P_k(t)$; each stage orders after observing the forecast revision for period t.

We assume that node 1 promises a *guaranteed service time* S_1 by which it will satisfy customer demand.

That is, the customer demand D_t at time t must be filled by time $t + S_1$. Furthermore, we assume that stage 1 provides 100% service whereby it delivers exactly D_t to the customer at time $t + S_1$.

Similarly, we assume that each upstream stage k ($k \neq 1$) quotes and guarantees a service time S_k to its internal customer, namely, stage $k-1$. At period t, the order placed on stage k by stage $k-1$ is $P_{k-1}(t)$; stage k delivers exactly this amount to stage $k-1$ at time $t+S_k$.

We assume that the service time for the external customer, S_1, is determined exogenously and is a given input; without loss of generality we can assume $S_1 = 0$. (We can model a nonzero S_1 by shifting the forecast process so that the forecast is perfect over the time between customer order and delivery.) The remaining internal service times are *decision variables* and determine where we place safety stocks. The replenishment time at node k is the service time to get its inputs, plus its processing time, namely $S_{k+1} + T_k$. We define τ_k to be the *net replenishment time* for node k, equal to its replenishment time net of its service time:

$$T_k + S_{k+1} - S_k = \tau_k. \tag{4}$$

We will see that the net replenishment time determines the safety stock at node k.

Graves and Willems (2000, 2003, 2008) provide motivation and justification for the assumption of guaranteed service times. The benefits of the guaranteed service assumption are at least twofold. First, it permits significant analytical tractability for characterizing and determining the safety stocks in supply chains. Second, guaranteed service has practical value in that it greatly facilitates coordination in assembly contexts, and along with the assumption of bounded demand, provides a more amenable way for managers to frame the service tradeoffs in a supply chain. Nevertheless, an assumption of guaranteed service is restrictive and comes with a cost, namely somewhat higher inventories (see Aviv 2007).

Our assumptions to this point are identical to those made by Simpson (1958) and Graves and Willems (2000). These authors assume that each stage uses a base-stock ordering policy; that is, in each period, each stage observes and orders the customer demand:

$$P_k^{\text{Base stock}}(t) = f_t(t) = D_t \quad \forall k. \tag{5}$$

By contrast, in this paper we assume that each stage places an order based on the forecast of future demand. Specifically, we define

$$L_k = S_{k+1} + \sum_{j=1}^{k} T_j = \sum_{j=1}^{k} \tau_j \tag{6}$$

as the cumulative lead time for node k. This represents the shortest time for an order on stage k to reach the final stage and become available to meet customer demand. The cumulative lead time for stage k consists of the service time to obtain the raw material or input (S_{k+1}) plus the processing time at stage k and all downstream nodes ($\sum_{j=1}^{k} T_j$). Given the cumulative lead time L_k, we denote the order placed by stage k at time t as follows:

$$P_k(t) = f_t(t + L_k) + \sum_{i=0}^{L_k - 1} \Delta f_t(t + i). \tag{7}$$

We note that this ordering mechanism assumes that in each period the forecast is shared among all the nodes. We will term (7) to be the *forecast-based ordering policy*. Intuitively, if forecasts were perfect ($\Delta f_t \equiv 0$), then $P_k(t) = f_t(t + L_k)$; in each period, each node of the supply chain places an order so as to push forward exactly what is necessary to meet customer demand in the future, and there is no need for safety stocks. We can view the base-stock policy as a special case of the forecast-based ordering policy (7) in which the initial forecast $f_t(t+H) = \mu$, there are no forecast revisions until the period of demand realization, and then $\Delta f_t(t) = f_t(t) - f_{t-1}(t) = D_t - \mu$. With these assumptions, it is easy to show that for each stage (7) reduces to the base-stock policy $P_k(t) = D_t$.

The forecast-based ordering policy given by (7) can result in negative orders. In the following analysis we ignore this possibility, and in effect assume that $P_k(t)$ is nonnegative, corresponding to the situation in which the forecast revisions in each period are small relative to the forecasted demand. In practice, if there were sufficient downward revisions in the forecast to suggest a negative order, then we presume that no order ($P_k(t) = 0$) would be placed, and that any surplus order would be deducted from the next positive order. We do not account for this dynamic in the following analyses.

We show in the online appendix that the forecast-based ordering policy (7) is equivalent to a policy in which each node k in each period t places an order so

as to keep constant the expected inventory at node k at time $t + T_k + S_{k+1}$. We note that the forecast-based ordering policy (7) is the multinode extension of the installation-based order-up-to policy from Aviv (2003).

Moreover, the forecast-based ordering policy is analogous to what one might expect in practice, because it represents the orders from applying MRP logic to a serial system with no lot sizing and no yield uncertainties. In particular, if we denote node k's on-hand inventory at the end of time t with $I_k(t)$, we show in the online appendix that we can write (7) recursively as

$$P_0(t) = D_t$$

$$P_k(t) = \underbrace{\sum_{i=1}^{T_k+S_{k+1}} \mathrm{E}_t[P_{k-1}(t+i-S_k)]}_{\text{scheduled downstream demand}} - \underbrace{\sum_{i=1}^{T_k+S_{k+1}-1} P_k(t-i)}_{\text{inventory on order}} - \underbrace{I_k(t)}_{\substack{\text{inventory}\\ \text{on hand}}} + \underbrace{I_k^0}_{\substack{\text{desired safety}\\ \text{stock}}}. \quad (8)$$

The first term on the right-hand side represents the order schedule that node k needs to fulfill over its replenishment lead time. The second term represents what is currently on order to node k, namely, the inbound orders from node $k+1$ and the orders currently in process at node k. The third term is inventory on hand. The final term, I_k^0, is a constant safety-stock target, which is set to maintain an inventory buffer for the eventuality of higher than expected demand. Thus, from (8) we see that the order placed by stage k in time t equals the forecast of requirements on stage k over its replenishment lead time, net of the inventory that it will have available over this time period, plus the safety-stock target.

In some simple settings, we can show that the forecast-based ordering policy is optimal with respect to certain criteria. As noted, the policy is optimal when the forecasts are perfect, because the inventory variation is zero in this case, and no safety buffers are needed. Moreover, Aviv (2003) shows by induction that each stage should follow such a policy to deliver on orders received and minimize a quadratic cost function. These results all assume that the service times are zero. Our contribution is to consider nonzero service times in a global optimization problem.

3.2. Inventory Dynamics

Given a forecast-based ordering policy, we now investigate the dynamics of the inventories $I_k(t)$. Under the guaranteed service time assumption, we have the inventory balance equations

$$I_k(t+1) = I_k(t) - P_{k-1}(t+1-S_k) + P_k(t+1-S_{k+1}-T_k). \quad (9)$$

We show in the online appendix that by combining (7) and (9), we have

$$I_k(t+T_k+S_{k+1}) = I_k^0 - \sum_{i=t+1}^{t+\tau_k} \sum_{j=i}^{t+L_k} \Delta f_i(j), \quad (10)$$

where we choose the time $t + T_k + S_{k+1}$ on the left-hand side for ease of exposition, and I_k^0 is the target safety stock. The expression (10) shows that the current inventory level is a function of recent forecast revisions. Because we assume that $\underline{\Delta f}_t$ is i.i.d. with a mean of zero and known covariance function, we see from (10) that the properties of the inventory random variable at a stage do not depend on the specific time period but only on its cumulative lead time and its net replenishment time.

To get some insight on this characterization of the inventory, we can re-express the forecast revision summation in terms of the forecast errors:

$$\begin{aligned}
&\sum_{i=t+1}^{t+\tau_k} \sum_{j=i}^{t+L_k} \Delta f_i(j) \\
&= \sum_{i=t+1}^{t+L_k} \sum_{j=i}^{t+L_k} \Delta f_i(j) - \sum_{i=t+\tau_k+1}^{t+L_k} \sum_{j=i}^{t+L_k} \Delta f_i(j) \\
&= \sum_{j=t+1}^{t+L_k} \sum_{i=t+1}^{j} \Delta f_i(j) - \sum_{j=t+\tau_k+1}^{t+L_k} \sum_{i=t+\tau_k+1}^{j} \Delta f_i(j) \\
&= \sum_{j=t+1}^{t+L_k} (D_j - f_t(j)) - \sum_{j=t+\tau_k+1}^{t+L_k} (D_j - f_{t+\tau_k}(j)). \quad (11)
\end{aligned}$$

The first term on the right-hand side of (11) is the cumulative forecast error for the forecast made at time t for the next L_k periods. The second term is the cumulative forecast error for the forecast made at time $t+\tau_k$ for the next $L_{k-1} = L_k - \tau_k$ periods. Thus, from (11) we need to set the safety-stock target to cover the forecast revisions that are made to the L_k-period cumulative forecast over the next $\tau_k = L_k - L_{k-1}$ periods.

We now assume that we have a bound $B(L_{k-1}, L_k)$ on the forecast revisions to the L_k-period cumulative

forecast over the next $\tau_k = L_k - L_{k-1}$ periods. That is, we identify $B(L_{k-1}, L_k)$ such that

$$\sum_{i=t+1}^{t+\tau_k} \sum_{j=i}^{t+L_k} \Delta f_i(j) \le B(L_{k-1}, L_k) \quad \forall t. \tag{12}$$

If we set

$$I_k^0 \leftarrow B(L_{k-1}, L_k), \tag{13}$$

then it is clear from (10) that the inventory is nonnegative, and thus we fulfill the guaranteed service constraint. This development is directly analogous to the base-stock model in which the base-stock level is set equal to the demand bound over the net replenishment time (see Kimball 1988, Simpson 1958).

A natural question is how to determine the bound function. We might obtain this bound based on historical data; if we have enough observations of the forecast revisions, then we can develop an empirical distribution for the left-hand side of (12) and use this to determine bounds for setting the safety stocks.

An alternate way to obtain the bound is to suppose that management specifies that the safety stock is to protect against some maximum level of forecast error. (For a discussion and justification of this perspective, see Simpson 1958, Graves and Willems 2000.) In particular, suppose we can measure the standard deviation of the cumulative forecast error for each possible cumulative lead time. Then, for the purposes of setting the safety stocks, we might set the maximum forecast error analogous to a service-level bound; that is, we define $F(L)$ to be the maximum cumulative forecast error for any interval of length L as follows:

$$F(L) = z\sigma\left(\sum_{j=t+1}^{t+L} (D_j - f_t(j))\right), \tag{14}$$

where z is a safety factor and $\sigma(\,)$ is the standard deviation. Thus, we want the safety stock to provide 100% protection as long as the forecast errors are within $F(L)$ for all lead times L.

With this specification, we show in the online appendix that the bound function is simply

$$\begin{aligned} B(L_{k-1}, L_k) &= z\sigma\left(\sum_{i=t+1}^{t+\tau_k} \sum_{j=i}^{t+L_k} \Delta f_i(j)\right) \\ &= \sqrt{F^2(L_k) - F^2(L_{k-1})}. \end{aligned} \tag{15}$$

Hence, if we are given the maximum allowable level of forecast errors (14) for each possible L, we can then determine the bound function (15). From the bound function we can determine the safety-stock level (13) that is necessary to assure the guaranteed service for all forecast/demand realizations within the maximum forecast errors.

We note that (15) is a fairly simple and workable form. We just need to characterize the variance of the cumulative forecast error over all relevant time horizons. From this function, we can directly compute the bound function as given by (15). In the next section we show how we use this bound function to choose the optimal service times S_k (and consequently L_k) to minimize total inventory costs.

4. Optimization

Given the bound function on the forecast revisions, we can formulate the optimization problem. The objective is to minimize the expected inventory holding costs in the system. From (10) we observe that the expected inventory level at each stage is given by

$$\mathrm{E}[I_k] = I_k^0.$$

We assume that we set the safety-stock target I_k^0 according to (13), and thus,

$$\mathrm{E}[I_k] = I_k^0 = B(L_{k-1}, L_k).$$

Finally, we assume that each stage k incurs holding costs at a rate h_k proportional to the average inventory level I_k^0. In addition to this safety stock, the supply chain has pipeline inventory, which is directly proportional to the production lead times. We do not consider pipeline inventory in the optimization because this inventory does not depend at all on the choice of service times.

By changing the service times S_k we can find different safety-stock configurations; we seek the least-cost solution. The optimization problem for a serial supply chain is

$$\begin{aligned} \min_{S_k} \quad & \sum_{k=1}^{N} h_k B(L_{k-1}, L_k), \\ \text{s.t.} \quad & S_{k+1} + T_k \ge S_k \quad \forall k, \\ & L_k = S_{k+1} + \sum_{j=1}^{k} T_j \quad \forall k, \\ & S_k \ge 0 \quad \forall k, \\ & S_1,\ S_{N+1},\ L_0 = 0. \end{aligned} \tag{16}$$

The first set of constraints assures that the net replenishment time is nonnegative for each stage, and the second set of constraints defines the cumulative lead times for each stage.

Simpson (1958) posed and analyzed a similar problem for a serial system operating with a base-stock policy. He assumes that for any time interval $(0, \tau]$ there is a bound on the demand given by

$$B(\tau) = \mu\tau + z\sigma\sqrt{\tau},$$

where μ is the average demand rate and σ corresponds to the standard deviation of demand. We can interpret his assumptions and his analysis as a special case of optimization problem (16) in which the bound function for each stage k is given by

$$\begin{aligned} B(L_{k-1}, L_k) &= z\sigma\sqrt{L_k - L_{k-1}} \\ &= z\sigma\sqrt{T_k + S_{k+1} - S_k} \quad \forall k. \end{aligned} \tag{17}$$

Thus, for the base-stock policy, the objective function is a sum of terms, each of which is concave in the service times. Consequently, the solution is on the corners of the solution space, which implies "all-or-nothing" solutions: Either a node keeps no safety stock ($S_{k+1} + T_k = S_k$), or it keeps so much safety stock that it is decoupled from the downstream nodes ($S_k = 0$). It also means that we can find the optimal solution for a serial system through enumeration.

Now suppose we assume an evolving forecast and the forecast-based ordering policy, with the bound given by (15). We will demonstrate that the optimization (16) for the general case is no more difficult than that solved by Simpson (1958) for the special case of a base-stock policy. For ease of notation, we define

$$g(L) = F^2(L) = z^2 \operatorname{var}\left[\sum_{j=t+1}^{t+L} (D_j - f_t(j))\right]. \tag{18}$$

We can rewrite the optimization problem as

$$\begin{aligned} \min_{S_k} \quad & \sum_{k=1}^{N} h_k\sqrt{g(L_k) - g(L_{k-1})}, \\ \text{s.t.} \quad & S_{k+1} + T_k \geq S_k \quad \forall k, \\ & L_k = S_{k+1} + \sum_{j=1}^{k} T_j \quad \forall k, \\ & S_k \geq 0 \quad \forall k, \\ & S_1,\ S_{N+1},\ L_0 = 0. \end{aligned}$$

Without loss of generality, we add $\sum_{j=1}^{k-1} T_j$ (a constant) to both sides of the first set of constraints and to the nonnegativity constraints:

$$\begin{aligned} \min_{S_k} \quad & \sum_{k=1}^{N} h_k\sqrt{g(L_k) - g(L_{k-1})}, \\ \text{s.t.} \quad & S_{k+1} + T_k + \sum_{j=1}^{k-1} T_j \geq S_k + \sum_{j=1}^{k-1} T_j \quad \forall k, \\ & L_k = S_{k+1} + \sum_{j=1}^{k} T_j \quad \forall k, \\ & S_k + \sum_{j=1}^{k-1} T_j \geq \sum_{j=1}^{k-1} T_j \quad \forall k, \\ & S_1,\ S_{N+1},\ L_0 = 0. \end{aligned}$$

Now we can rewrite everything, including the decision variable, in terms of L_k:

$$\begin{aligned} \min_{L_k} \quad & \sum_{k=1}^{N} h_k\sqrt{g(L_k) - g(L_{k-1})}, \\ \text{s.t.} \quad & L_k \geq L_{k-1} \quad \forall k, \\ & L_k \geq \sum_{j=1}^{k} T_j \quad \forall k, \\ & L_0 = 0. \end{aligned}$$

At this point, we need a mild technical assumption: The variance of the cumulative forecast error $g(L)$ is strictly increasing in L. Then, we apply $g(\,)$ to both sides of the constraint equations:

$$\begin{aligned} \min_{L_k} \quad & \sum_{k=1}^{N} h_k\sqrt{g(L_k) - g(L_{k-1})}, \\ \text{s.t.} \quad & g(L_k) \geq g(L_{k-1}) \quad \forall k, \\ & g(L_k) \geq g\left(\sum_{j=1}^{k} T_j\right) \quad \forall k, \\ & g(L_0) = 0. \end{aligned}$$

Finally, we define $Z_k = g(L_k)$ and use this as a scalar decision variable. We can do this because, by assumption, $g(\,)$ is strictly increasing, and hence it is a bijective (one-to-one) mapping. Every solution in terms of Z_k corresponds to a unique solution in terms of L_k and S_k, and the feasibility and objective value are

unaffected by the mapping (note that $g(\sum_{j=1}^{k} T_j)$ is a constant). The final program is

$$\begin{aligned} \min_{Z_k} \quad & \sum_{k=1}^{N} h_k \sqrt{Z_k - Z_{k-1}}, \\ \text{s.t.} \quad & Z_k \geq Z_{k-1} \quad \forall k, \\ & Z_k \geq g\left(\sum_{j=1}^{k} T_j\right) \quad \forall k, \\ & Z_0 = 0. \end{aligned} \tag{19}$$

This program has the same concavity properties found by Simpson (1958), which implies that the optimal solution is found on a corner of the feasible region: for each stage either $Z_k = g(\sum_{j=1}^{k} T_j)$ or $Z_k = Z_{k-1}$, corresponding to $S_{k+1} = 0$ or $S_{k+1} + T_k - S_k = 0$, respectively. Thus, *the all-or-nothing property of the optimal solution still holds when there is an evolving forecast and the forecast-based ordering policy*. Moreover, we can find the solution by enumeration, although, as we will see next, faster dynamic programming methods are available as well.

4.1. Pure Assembly Systems

Graves and Willems (2000) introduce a dynamic programming algorithm to solve the safety-stock optimization problem for systems with base-stock ordering. This approach is not only faster than that of Simpson (1958), but it can also be used on supply chains with more general spanning-tree topology. Moreover, it does not rely on concavity properties of the bound function $B(L_{k-1}, L_k)$.

We now show how to use the forecast-based results for pure assembly structures. We discuss the more general spanning tree problem in the last section as an area for future research.

For the assembly system, we let $k = 1$ be the customer-facing node as before, and introduce the function $a(k)$ to denote the node that is immediately downstream (after) of k. We set $a(1) = 0$. Each node can now have multiple upstream supply nodes. Because processing at a node cannot start until material from *all* of its supply nodes is available, we define the inbound service time SI_k as the longest service time from the set of supply nodes:

$$SI_k = \max_{\{j:\, a(j)=k\}} S_j \quad \forall k. \tag{20}$$

We then can define the cumulative lead time for each stage by the recursion

$$\begin{aligned} L_0 &= 0, \\ L_k &= SI_k + T_k - S_k + L_{a(k)}. \end{aligned}$$

Given the cumulative lead time L_k, we assume that in each period t each node k places an order on its supply nodes for delivery at time $t + SI_k$ with the forecast-based ordering policy, namely,

$$P_k(t) = f_t(t + L_k) + \sum_{i=0}^{L_k - 1} \Delta f_t(t + i).$$

Analogous to (8), we can re-express the ordering policy in the following form:

$$\begin{aligned} P_0(t) &= D_t \\ P_k(t) &= \underbrace{\sum_{i=1}^{T_k + SI_k} \mathrm{E}_t[P_j(t + i - S_k)]}_{\text{scheduled downstream demand}} - \underbrace{\sum_{i=1}^{T_k + SI_k - 1} P_k(t - i)}_{\text{inventory on order}} \\ &\quad - \underbrace{I_k(t)}_{\substack{\text{inventory} \\ \text{on hand}}} + \underbrace{I_k^0}_{\substack{\text{desired} \\ \text{safety stock}}}, \end{aligned} \tag{21}$$

where $j = a(k)$. (For any supply node i for which $S_i < SI_k$ and $k = a(i)$, we delay each order from node k by $SI_k - S_i$ periods so as to avoid early delivery and excess inventory.)

As in Graves and Willems (2000), the inventory dynamics at each node k depend on the node's outbound and inbound service times, namely, S_k, SI_k. By following the same development as for the serial system, we can express the expected inventory at each stage k as

$$\mathrm{E}[I_k] = I_k^0 = B(L_{a(k)}, L_k).$$

We can then write the optimization problem as

$$\begin{aligned} \min_{S_k,\, SI_k} \quad & \sum_{k=1}^{N} h_k B(L_{a(k)}, L_k), \\ \text{s.t.} \quad & SI_k + T_k \geq S_k \quad \forall k, \\ & L_k = L_{a(k)} + SI_k + T_k - S_k \quad \forall k, \\ & SI_k \geq S_j \quad \forall j,\ k = a(j), \\ & S_1, L_0 = 0, \\ & SI_k, S_k \geq 0 \quad \forall k. \end{aligned} \tag{22}$$

As for the optimization for a serial system, the first set of constraints assures that the net replenishment time is nonnegative and the second set specifies the cumulative lead time. For the assembly systems we need to add a third set of constraints to relate the inbound service time for each stage to the outbound service times for the adjacent upstream stages.

The optimization problem (22) is the same problem solved by Graves and Willems's (2000) algorithm for a base-stock system, except that we now have a different bound function. Graves and Willems (2000) use a demand bound in their objective function, whereas here we use the bound function on the forecast revisions, given by (12); that is, we set the bound function (with modifications for the assembly system) as

$$\sum_{i=t+1}^{t+\tau_k} \sum_{j=i}^{t+L_k} \Delta f_i(j) \le B(L_{a(k)}, L_k) \quad \forall t, \tag{23}$$

where the net replenishment time is $\tau_k = L_k - L_{a(k)}$. Because Graves and Willems's (2000) dynamic programming algorithm does not rely on any special properties of the bound function, we can solve (22) by their algorithm.

5. Case Study

To test the results from the previous sections, we performed a case study of the supply chain for an electronic testing system manufactured by Teradyne, Inc. At the time, Teradyne had large safety stocks and a high service level, but was looking at ways to reduce inventories. It was thus a good match for our research. This test case also allowed us to develop some intuition for how the new method manifests itself in terms of the locations and quantities of safety stocks. To do these things, we implemented Graves and Willems's (2000) dynamic programming algorithm in the PERL programming language, after modifying it with the new bound function.

The supply chain produces a family of semiconductor test equipment. The actual product sold is customized to meet the requirements of the customer's application. This customization is accomplished by the selection of options from a large set of alternatives, where there is an electronic subassembly for each option. Nevertheless, except for this choice of options, the rest of the product is standard for all customers. For our test, we consider only the standard bill of material, and the corresponding supply chain that is common for all products. This supply chain entails 3,866 stages or nodes and a single end item, where each node represents one specific part at one specific location. The supply chain extends over multiple locations. Many of the production steps are not done by Teradyne, but by subcontractors. Because of close cooperation and strong relationships with the suppliers, Teradyne has considerable influence over safety stocks at their locations as well. We used real data from the bill of material to characterize the different parts and locations. We assumed that the holding cost was directly proportional to the value of the parts, which were already calculated by Teradyne. We plot the supply chain topology in Figure 1.

Teradyne forecasts the demand for future weeks in a master schedule. Orders are first entered as open or "preliminary" orders, representing perhaps an early discussion with an interested customer or a sales target. Eventually the customer has to commit and the order is booked. In this way, the master schedule can be seen as a forecast. The next few weeks are quite accurate (booked orders cannot be cancelled, and new orders are usually not allowed), whereas further into the future the schedule is bound to undergo more changes, and hence it is less reliable.

Figure 1 **Schematic View of Supply Chain for the Studied Product**

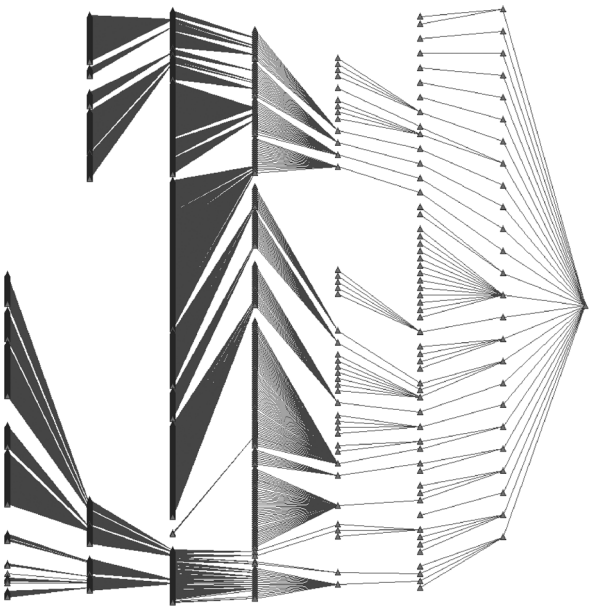
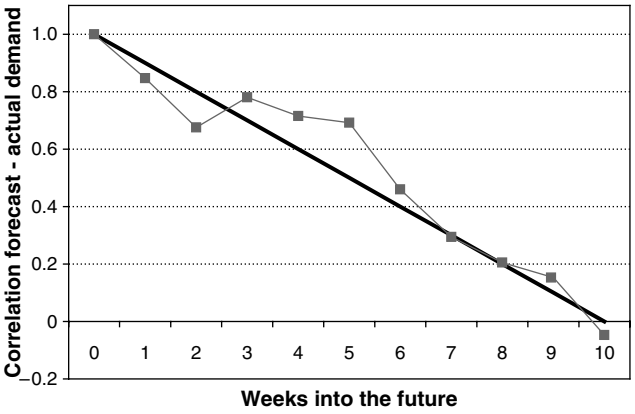


Figure 2 **Forecast Quality (Correlation with What Was Actually Produced) as a Function of Time into the Future for an Electronic Test System**



We collected data on schedules and their revisions for one year, and compared the schedules with actual demand. For each week, we had data for the forecasts made for a 16-week horizon into the future. In total, we had about 50 observations for each of the 16 forecasts in the forecast horizon; that is, we had 50 observations for the one-week-ahead forecast, for the two-week ahead forecast, up to the 16-week-ahead forecast.

As shown in Figure 2, we measured the correlation between each forecast and demand and found that this correlation decreased approximately linearly over the next 10 weeks. Beyond 10 weeks, we found that there was effectively no correlation, which implies that the forecast had no predictive power. In the subsequent experiments we use the linear fit for the first 10 weeks, and assume zero correlation beyond that. Under the assumption of i.i.d. forecast revisions we find that the forecast correlation is equivalent to the standard deviation of the forecast, normalized with respect to the standard deviation of demand:

$$\begin{aligned}\rho(D_t, f_i(t)) &= \frac{\operatorname{cov}(D_t, f_i(t))}{\sigma(D_t)\sigma(f_i(t))} \\ &= \frac{\operatorname{cov}\left(f_i(t) + \sum_{j=i+1}^{t} \Delta f_j(t), f_i(t)\right)}{\sigma(D_t)\sigma(f_i(t))} \\ &= \frac{\operatorname{cov}(f_i(t), f_i(t)) + \operatorname{cov}\left(\sum_{j=i+1}^{t} \Delta f_j(t), f_i(t)\right)}{\sigma(D_t)\sigma(f_i(t))} \\ &= \frac{\sigma^2(f_i(t)) + 0}{\sigma(D_t)\sigma(f_i(t))} = \frac{\sigma(f_i(t))}{\sigma(D_t)}. \end{aligned} \tag{24}$$

We can relate this forecast quality measure to the variance of the forecast error, which we use to calculate the bound function in Equation (15); we show in the online appendix that

$$\operatorname{var}(D_t - f_i(t)) = (1 - \rho^2(D_t, f_i(t)))\operatorname{var}(D_t). \tag{25}$$

To make the initial test simple, we assume that $\Delta f_t(t+j)$ is independent for different values of j. With this additional assumption, and Equations (25) and (17), we have (see the online appendix)

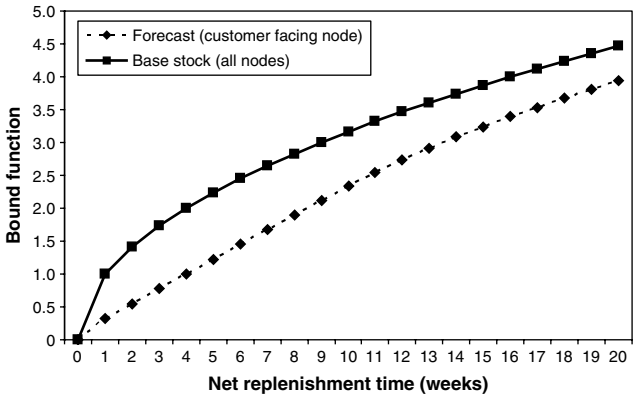
$$\begin{aligned} &B(L_{a(k)}, L_k) \\ &= z\sigma(D_t)\sqrt{L_k - L_{a(k)} - \sum_{j=t+L_{a(k)}+1}^{t+L_k} \rho^2(D_j, f_t(j))} \\ &= z\sigma(D_t)\sqrt{T_k + SI_k - S_k - \sum_{j=t+L_{a(k)}+1}^{t+L_k} \rho^2(D_j, f_t(j))}. \end{aligned} \tag{26}$$

This is similar to the expression for the base stock problem (17) with $SI_k = S_{k+1}$, but now the square root has an additional term. In particular, we reduce the net replenishment time for node k ($\tau_k = L_k - L_{a(k)}$) by a measure of the forecast quality over the time window $(t + L_{a(k)}, t + L_k]$. This measure depends on the correlation between the forecasts and demand; thus, a forecast that is positively correlated with demand always leads to less inventory relative to the base-stock policy.

We note that in both cases the bound is proportional to the term $z\sigma(D_t)$; this can be seen as a constant with which the objective function is multiplied, but which does not affect the optimal solution or the relative performance between the base-stock policy and the forecast-based policy. Making the arbitrary assignment $z\sigma(D_t) \leftarrow 1$, we compared the bound functions for the base-stock ordering policy versus the forecast-based ordering policy, using the straight regression line from the correlation terms measured at Teradyne and illustrated in Figure 2. In Figure 3, we plot the bound $B(0, L)$ for both the base-stock system (17) and for the forecast system (26). Because we assume $S_1 = 0$, the bound function $B(0, L)$ equals the required safety stock for node $k = 1$ if its net replenishment time was $L = T_1 + SI_1$.

We see in Figure 3 that for node 1, the forecast-based ordering policy results in significantly less inventory than the base-stock ordering policy, especially

Figure 3 Normalized Bound Functions for Systems with Forecasts and with Base-Stock Policies



when its net replenishment time is only a few weeks. This is because the forecasts are relatively accurate in the short term. As the net replenishment time increases, the accuracy of the forecast decreases; as a consequence, the forecast has less value and the inventory savings decline.

We solved the optimization problem (22) for four cases. For three cases, we assume a forecast-based ordering policy but with different forecast properties. Specifically, we assumed that the correlation between the forecast and actual demand drops linearly from one to zero over a five-week, 10-week, or 20-week horizon, and then remains at zero beyond this horizon. As illustrated in Figure 2, the 10-week horizon closely matches the actual situation at Teradyne and is our base case. The five-week and 20-week horizons were hypothetical cases included for comparative purposes. For the fourth case, we assume a base-stock ordering policy, i.e, the Graves and Willems (2000) optimization; this case ignores the evolving forecast.

The safety-stock holding cost for the forecast-based ordering policy with a 10-week horizon is 25% less than that for the base-stock ordering policy. Thus, for this supply chain, there seems to be substantial benefit from accounting for the forecast evolution when setting the safety stocks.

We also find that the safety stocks depend on the quality of the forecast process. For the actual case, the forecast improves steadily over a 10-week period as seen in Figure 2. In comparison, we considered the supply chain assuming that the forecast improves over a 20-week horizon. This is a higher-quality forecast because all of the forecasts in a 20-week horizon are more accurate relative to the 10-week case. When we optimize this case, we find that there is a reduction of 21% in the safety-stock holding costs relative to the optimal solution for the 10-week case. Similarly, we considered a lower-quality forecast in which the forecast improves over a five-week horizon. Here, the optimal safety-stock holding costs were 17% higher than for the case with a 10-week horizon.

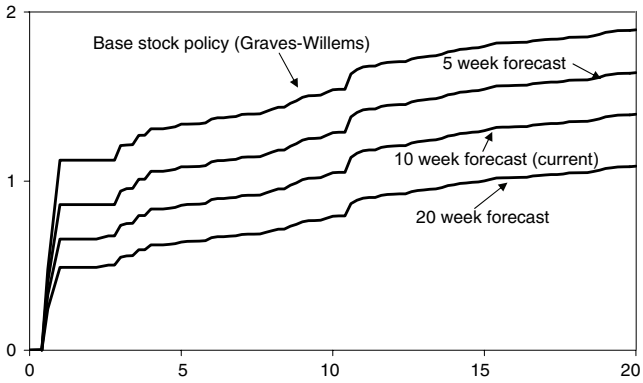
To get some intuition for how the four solutions differ, we calculated how the inventory was distributed in the supply chain for each case. To make this comparison, we defined the *minimal cumulative lead time* $\check{L}_k$ to be the cumulative lead time L_k when $SI_k = 0$. That is,

$$\begin{aligned} \check{L}_1 &= T_1, \\ \check{L}_j &= \check{L}_k + T_j, \quad k = a(j), \ \forall j. \end{aligned} \tag{27}$$

By construction, $\check{L}_k$ is a property of each node that does not depend on the service times, and so it serves as a measure by which we can compare the inventory placement for different solutions.

In Figure 4, we plot the total holding cost (on the y axis) for all nodes k such that $\check{L}_k \le x$ (on the x axis). For example, at 10 weeks on the x axis, the curve represents the total inventory holding cost for all inventory that can, in theory, be processed into finished products within 10 weeks. We only plot the holding costs for the first 20 weeks, which accounts for 97% of the total holding cost of the supply chain for the base stock case. Beyond 20 weeks, the cumulative holding costs for the four cases grow at the same rate.

Figure 4 Total Safety-Stock Inventory Costs for All the Nodes with Less Than a Certain Lead Time



From Figure 4, we see that the difference in the solutions is found primarily in those parts of the supply chain whose distance is less than the effective range of the forecasts. For example, the safety stock for the base-stock policy initially grows much faster than that for the 10-week case. As explanation, the 10-week case requires very little safety stock in the downstream parts of the supply chain, because it can take advantage of the accuracy of the short-term forecasts. In contrast, the base-stock ordering policy does not use these forecasts and so cannot realize this gain; this policy must have inventories commensurate with the temporal variations of demand, which are considerable. However, beyond 10 weeks, the safety stocks for the 10-week case and for the base-stock case grow at approximately the same rate. This is because the stages with cumulative lead times greater than 10 weeks use the same bound functions for both the forecast case and the base-stock case, and hence, the forecast case has no advantage over the base-stock case.

The computations for each of these supply chain problems took about one minute on a mobile computer (Intel®Core™2 CPU, 2.33 GHz, 1 GB RAM); no doubt this time can be reduced by implementation in a compiled programming language.

It is somewhat difficult to compare the solution directly to the actual situation at Teradyne, because at the time there were multiple layers of buffering and it was not clear what the actual service level was. Because Teradyne had not performed a global optimization to set its stocks, we suspect that the optimal solution for the base-stock ordering policy is a fairly conservative proxy for their current state. Hence, we believe there is substantial opportunity for improvement from the consideration of the evolving forecast.

6. Additional Numerical Examples

To examine the impact of an evolving forecast for various supply chains, we performed a set of numerical experiments. We used the same supply chain and cost structures as in Graves and Willems (2008). Specifically, we considered a serial system with $N = 5$ nodes, and with three alternatives for both the cost accumulation and the production lead time as follows.

The terms "increasing" and "decreasing" should be understood in terms of going upstream starting from the customer-facing stage 1. In the case of cost accumulation, the values stated in Table 1 represent the cost added at each stage. For example, for the increasing cost scenario, the cost at stage 5 is 36, the cost at stage 4 is $36 + 28 = 64$, the cost at stage 3 is $36 + 28 + 20 = 84$, etc. For all three scenarios, the cost of the finished good at stage 1 is 100.

Table 1 Alternative Structures for Supply Chain Lead Time and Cost Accumulation

Stage	5	4	3	2	1
Increasing	36	28	20	12	4
Constant	20	20	20	20	20
Decreasing	4	12	20	28	36

For the production lead times, the values for each scenario represent the values for T_k. For each scenario, the cumulative lead time for the supply chain is 100.

We assume that the length of the forecast horizon is 100 periods, and we use (26) as a bound on the forecast revisions, where we assume $z = 2$ and $\sigma = 20$. Similar to the Teradyne example, we assume that the correlation between the forecast and realized demand goes linearly from 0 to 1 over a horizon of 0, 25, 50, 75, or 100 periods. The first case thus represents no useful forecasts and is equivalent to the base-stock policy. The remaining four cases represent increasing improvements in the quality of the forecast.

The combination of three cost structures, three lead time structures, and five forecast horizons results in a total of 45 experiments. In Table 2, we state the optimal holding cost for a 10% holding cost rate for the zero-horizon case (the base-stock policy). For the other forecast-horizon cases, we report the optimal cost as a percentage of the zero-horizon case. We report the structure of the optimal solution in Table 3. We denote a solution by a binary code, whereby a "1" in the kth position denotes a decoupling inventory at stage k, whereas a "0" denotes no inventory. For example, "00001" represents inventory at stage 1 only.

To interpret these results, we make two observations. First, we note that by design, we always hold a safety stock at stage 1, the customer-facing node. Second, we note from (26) that the forecast has value at a stage only if the cumulative lead time for the stage's customer is within the forecast horizon. For instance, if the forecast horizon is 25 periods, then the forecast has no impact on any stage whose customer's

Table 2 Total Costs for Various Supply Chains and Forecast Horizons

		Forecast horizon				
Cost	Lead time	0	25 (%)	50 (%)	75 (%)	100 (%)
Increasing	Increasing	40.0	96.0	90.8	84.5	78.3
	Constant	40.0	96.0	91.6	86.9	82.0
	Decreasing	40.0	96.0	91.6	86.9	82.0
Constant	Increasing	36.8	87.2	79.7	72.2	66.0
	Constant	39.4	95.4	90.3	84.8	79.0
	Decreasing	40.0	96.0	91.6	86.9	82.0
Decreasing	Increasing	26.8	79.2	66.7	58.2	52.0
	Constant	34.6	93.9	85.0	76.6	69.7
	Decreasing	39.2	95.5	90.5	85.2	79.4

cumulative lead time is more than 25 periods. As a consequence of these observations, we find that the forecast is most relevant for the downstream stages, and we can get much of our insight by examining the impact on stage 1, the stage at which inventory is most expensive.

First consider the test cases for the increasing cost scenario. For this scenario the costs added are highest at the upstream stages. With no forecast, for each lead time scenario the optimal policy is to hold a single safety stock at stage 1 that protects against the entire lead time of 100 periods. As we incorporate a 25-period forecast into the supply chain, the structure of the optimal policy remains the same; but there is a benefit as the forecast permits a reduction in the size of this safety stock at stage 1, as seen from (26). For both the constant- and decreasing-lead-time cases, this pattern continues as we improve the forecast and extend it over a longer horizon; we still hold one inventory at stage 1, but can reduce its size as we have a better forecast. For the case of increasing lead time, we observe a change in the structure of the solution. As we extend the forecast beyond 50 periods, we introduce a safety stock at the most upstream stage. This safety stock reduces the net replenishment time for stage 1, resulting in less inventory held there. This change in structure occurs because the improved forecast reduces the size of the safety stock at the most upstream stage so that its cost is outweighed by the benefit from reducing the safety stock at stage 1.

Table 3 Structure of Optimal Solution

		Forecast horizon				
Cost	Lead time	0	25	50	75	100
Increasing	Increasing	00001	00001	10001	10001	10001
	Constant	00001	00001	00001	00001	00001
	Decreasing	00001	00001	00001	00001	00001
Constant	Increasing	01001	10011	10011	10101	10101
	Constant	10001	10001	10001	10001	10001
	Decreasing	00001	00001	00001	00001	00001
Decreasing	Increasing	11101	11011	11111	11111	11111
	Constant	11001	11001	10101	10101	10101
	Decreasing	11001	11001	11001	11001	10101

Note. 1 represents inventory at a node.

At the other extreme, consider the test cases for the decreasing-cost scenario. For this scenario, the costs added are highest at the downstream stages. Thus, it is relatively cheap to hold a safety-stock buffer at the upstream stages. Again, the safety stock at the upstream stages reduces the net replenishment time and inventory for stage 1, where holding inventory is most expensive. With no forecast, we see that the optimal policy holds safety stocks at two or three upstream stages, depending on the lead-time settings. For the constant- and decreasing-lead-time cases, the structure of the optimal policy remains pretty much the same as we add and improve the forecast. The amount of inventory, though, reduces significantly as the forecast improves. The cost savings occur primarily at stage 1, which can take full advantage of the improving forecast as its net replenishment time is well within the forecast horizon because of the upstream buffers. For the increasing-lead-time case, we see that as we improve the forecast, we have both a change in the structure of the solution and the largest savings in inventory costs. This is the most favorable case for using the forecast. The lead times are longest at the upstream stages, for which the cost added is the smallest. As we improve the forecast, the optimal solution places inventory at all stages so as to take advantage of the forecast. For each pair of stages, the additional cost of holding an inventory at the upstream stage is small relative to the inventory savings from reducing the net replenishment time at the downstream stage.

Finally, we note that for a particular cost and lead-time scenario, the structure of the solution is relatively stable across the different forecast scenarios. This is important because it suggests that the optimal location of safety stocks depends primarily on how the

holding costs and lead times are set across the supply chain, rather than on the specifics of the forecast process. However, we note that there are limited degrees of freedom in our set of experiments. As noted earlier, all solutions put a safety stock at stage 1, and the forecast has an impact on the upstream stages, only if its horizon is sufficiently long, depending on the lead times. Thus, it is primarily the middle stages on which we might see an impact from the forecasts.

7. Conclusions and Future Directions

In spite of the ubiquity of forecast-based planning systems (e.g., MRP systems), the analysis of safety stocks has been limited to simple special cases, such as one or two nodes, i.i.d. demand processes, or perfect forecasts (Guide and Srivastava 2000). In this paper, we develop an approach that extends the framework of Simpson (1958) and Graves and Willems (2000) to include an evolving forecast. We can then apply the dynamic programming algorithm from Graves and Willems (2000) to solve for the safety stocks in assembly-system supply chains. Furthermore, we demonstrate that accounting for the forecast evolution process results in less safety stock, where the magnitude of the savings depends on the quality of the forecasts. We expect that our approach is computationally fast enough to solve assembly systems of any size likely to arise in practice.

In the literature, there is debate over where to place safety stocks in MRP systems, and the type of buffer to use (Guide and Srivastava 2000). If one accepts the specific assumptions made in this paper, then the optimal placement of supply chain safety stocks is driven by three different (and sometimes conflicting) principles. The first two points are the same as for systems with base-stock ordering policies.

- Statistical economies of scale, as manifested in the (strict) concavity of the bound functions, encourage the use of fewer, larger, safety-stock buffers.
- Value-adding activities (holding costs that increase downstream in the supply chain) encourage the use of more numerous, smaller, and distributed safety stocks.
- When using a forecast-based ordering policy (e.g., MRP logic), the overall size of the safety stocks depends on the size of the forecast errors, rather than the variability of demand. To the extent that we have a meaningful forecast, we expect the forecast errors to be smaller than demand variability, resulting in less safety stock. These reductions in safety stock will primarily be downstream in the supply chain, at the stages whose cumulative lead times correspond to the horizon of useful forecasts.

Finally we remind the reader of the limitations of this work, and the opportunities that this suggests. In this paper, we have only considered serial and pure assembly structures. Graves and Willems (2000), whose algorithm we have modified, consider more general spanning-tree systems, which may have multiple customer stages. The primary challenge in such systems is how to determine the bound function (12) for stages that serve multiple demand processes. If one can specify this function, then one can extend the analysis in this paper to supply chains with spanning-tree topologies. The challenges involved in establishing such bounds, and some remedies, are discussed in Schoenmeyr (2008). We leave even more general (cyclic) networks for future research. Another limitation is the simplicity of the ordering policy. We do not consider many features that are typically incorporated in MRP systems, such as lot sizing, capacity constraints, and supply uncertainty; we do not account for these considerations in the present framework for strategic safety stocks. Finally, we note our assumption of deterministic procurement and production lead times. It would be most valuable to determine how to extend this approach to accommodate stochastic lead times, as is common in practice.

Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

Acknowledgments

This research was supported in part by the MIT Leaders for Manufacturing Program (a partnership between MIT and major manufacturing firms) and by the Singapore-MIT Alliance (an engineering education and research collaboration among the National University of Singapore, Nanyang Technological University and MIT), and by Marcus Wallenberg's Foundation for Education in International Industrial Management. The authors thank the editors and referees for their very helpful and constructive feedback on earlier versions of this paper.

References

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

Aviv, Y. 2003. A time-series framework for supply-chain inventory management. *Oper. Res.* **51**(2) 210–227.

Aviv, Y. 2004. Collaborative forecasting and its impact on supply chain performance. D. Simchi-Levi, S. D. Wu, Z. J. Shen, eds. *Handbook of Quantitative Supply Chain Analysis*, Chap. 10. Kluwer Academic Publishers, Norwell, MA.

Aviv, Y. 2007. On the benefits of collaborative forecasting partnerships between retailers and manufacturers. *Management Sci.* **53**(5) 777–794.

Axsäter, S. 1993. Continuous review policies for multi-level inventory systems with stochastic demand. S. C. Graves, A. H. Rinnooy Kanm, P. H. Zipkin, eds. *Logistics of Production and Inventory*, Chap. 4. North-Holland Publishing Company, Amsterdam, 175–197.

Baker, K. R. 1993. Requirements planning. S. C. Graves, A. H. Rinnooy Kan, P. H. Zipkin, eds. *Logistics of Production and Inventory*, Chap. 3. North-Holland Publishing, Amsterdam, 571–627.

Billington, C., G. Callioni, B. Crane, J. D. Ruark, J. U. Rapp, T. White, S. P. Willems. 2004. Accelerating the profitability of Hewlett-Packard's supply chains. *Interfaces* **34**(1) 59–72.

Buzacott, J. A., J. G. Shanthikumar. 1994. Safety stock versus safety time in MRP controlled production systems. *Management Sci.* **40**(12) 1678–1689.

Clark, A., H. Scarf. 1962. Approximate solutions to a simple multiechelon inventory problem. K. Arrow, S. Karlin, H. Scarf, eds. *Studies in Applied Probability and Management Science*, Chap. 5. Stanford University Press, Stanford, CA, 88–100.

Diks, E. B., A. G. de Kok, A. G. Lagodimos. 1996. Multi-echelon systems: A service measure perspective. *Eur. J. Oper. Res.* **95** 241–263.

Dong, L., H. L. Lee. 2003. Optimal policies and approximations for a serial multiechelon inventory system with time-correlated demand. *Oper. Res.* **51**(6) 969–980.

Federgruen, A. 1993. Centralized planning models for multiechelon inventory systems under uncertainty. S. C. Graves, A. H. Rinnooy Kan, P. H. Zipkin, eds. *Logistics of Production and Inventory*, Chap. 3. North-Holland Publishing, Amsterdam, 133–173.

Gallego, G., Ö. Özer. 2001. Integrating replenishment decisions with advance demand information. *Management Sci.* **47**(10) 1344–1360.

Graves, S. C., S. P. Willems. 2000. Optimizing strategic safety stock placement in supply chains. *Manufacturing Service Oper. Management.* **2**(1) 68–83.

Graves, S. C., S. P. Willems. 2003. Supply chain design: Safety stock placement and supply chain configuration. A. G. de Kok, S. C. Graves, eds. *Supply Chain Management: Design, Coordination and Operation*, Chap. 3. North-Holland Publishing Company, Amsterdam, 95–132.

Graves, S. C., S. P. Willems. 2008. Strategic inventory placement in supply chains: Non-stationary demand. *Manufacturing Service Oper. Management* **10**(2) 278–287.

Graves, S. C., D. B. Kletter, W. B. Hetzel. 1998. A dynamic model for requirements planning with application to supply chain optimization. *Oper. Res.* **46**(3) S35–S49.

Graves, S. C., H. C. Meal, S. Dasu, Y. Qiu. 1986. Two-stage production planning in a dynamic environment. S. Axsäter, C. Schneeweiss, E. Silver, eds. *Multi-Stage Production Planning and Inventory Control*. Springer, Berlin, 9–43.

Grimmett, G., D. Stirzaker. 2001. *Probability and Random Processes*. Oxford University Press, Oxford, UK.

Guide, Jr., V. D. R., R. Srivastava. 2000. A review of techniques for buffering against uncertainty in MRP systems. *Production Planning Control* **11**(3) 223–233.

Hamilton, J. D. 1994. *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Heath, D. C., P. L. Jackson. 1994. Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. *IIE Trans.* **26**(3) 17–30.

Inderfurth, K. 1994. Safety stocks in multistage, divergent inventory systems: A survey. *Internat. J. Production Econom.* **35** 321–329.

Karaesmen, F., J. A. Buzacott, Y. Dallery. 2002. Integrating advance order information in make-to-stock production systems. *IIE Trans.* **34** 649–662.

Kimball, G. E. 1988. General principles of inventory control. *J. Manufacturing Oper. Management* **1** 119–130.

Lagodimos, A. G., E. J. Anderson. 1993. Optimal positioning of safety stocks in MRP. *Internat. J. Production Res.* **31**(8) 1797–1813.

Lambrecht, M. R., J. A. Muckstadt, R. Luyten. 1984. Protective stocks in multistage production systems. *Internat. J. Production Res.* **22**(6) 1001–1025.

Molinder, A. 1997. Joint optimization of lot-sizes, safety stocks, and safety lead times in an MRP system. *Internat. J. Production Res.* **35**(4) 983–994.

Schoenmeyr, T. 2008. Strategic inventory placement in multiechelon supply chains: Three essays. Ph.D. thesis, MIT Sloan School of Management, Cambridge, MA.

Simpson, K. F. 1958. In-process inventories. *Oper. Res.* **6**(6) 863–873.

Willems, S. P. 2008. Real-world multiechelon supply chains used for inventory optimization. *Manufacturing Service Oper. Management* **10**(1) 19–23.

Yano, C. A., R. C. Carlson. 1987. Interaction between frequency of rescheduling and the role of safety stock in material requirements planning systems. *Internat. J. Production Res.* **25**(2) 221–232.

Zhang, X. 2004. Evolution of ARMA demand in supply chains. *Manufacturing Service Oper. Management* **6**(2) 195–198.