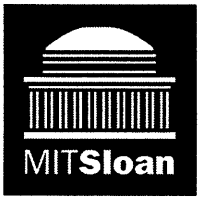

Massachusetts Institute of Technology

Sloan School of Management

Working Paper

Analysis of a Forecasting-Production-Inventory System with Stationary Demand

L. Beril Toktay
Lawrence M. Wein

April 1999
Working Paper Number 4070

Contact Address:
Prof. Lawrence M.Wein
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02142-1343

Lwein@mit.edu

Analysis of a Forecasting-Production-Inventory System with Stationary Demand

L. Beril Toktay
INSEAD
77305 Fontainebleau, France

and

Lawrence M. Wein
Sloan School of Management, MIT
Cambridge, MA 02142

Abstract

We consider a production stage that produces a single item in a make-to-stock manner. Demand for finished goods is stationary. In each time period, an updated vector of demand forecasts over the forecast horizon becomes available for use in production decisions. We model the sequence of forecast update vectors using the Martingale Model of Forecast Evolution developed by Graves et al. (1986, 1998) and Heath and Jackson (1994). The production stage is modeled as a single-server discrete-time continuous-state queue. We focus on the stationary version of a class of policies that is shown to be optimal in the finite time horizon, deterministic-capacity case, and use an approximate analysis rooted in heavy traffic theory and random walk theory to obtain a closed-form expression for the (forecast-corrected) base-stock level that minimizes the expected steady-state inventory holding and backorder costs. This expression, which is shown to be accurate under certain conditions in a simulation study, sheds some light on the interrelationships among safety stock, stochastic correlated demand, inaccurate forecasts, and random and capacitated production in forecasting-production-inventory systems.

April 1999

1. Introduction

In make-to-stock environments, manufacturers produce goods according to a forecast of future demands. Typically, within the confines of a materials requirement planning (MRP) system, future demands over a specific horizon are forecasted, these forecasts are revised each period in a rolling-horizon fashion, and production plans are updated accordingly. To better understand such forecasting-production-inventory settings, we analyze a discrete-time single-item make-to-stock queue facing a stationary demand process and rolling-horizon forecast updates. In our model, we envision forecasting and production-inventory control as decentralized activities: Forecasts are generated by a forecaster using some process (e.g., time-series methods, advance order information, expert judgement) that is unknown to the production manager. The production manager only observes a stream of forecast updates, and must convert these updates into a production policy that minimizes the expected steady-state holding and backorder costs of finished goods inventory. To perform this conversion in an effective manner, the production manager needs to have a characterization of both the demand and forecasting processes.

Graves et al. (1986, 1998) and Heath and Jackson (1994) have made great strides in enabling this characterization by independently developing a model for how forecasts evolve in time, which is dubbed the Martingale Model of Forecast Evolution (MMFE) by Heath and Jackson. We employ this simple but deceptively powerful tool to model the inputs to our make-to-stock queue. Readers are referred to Heath and Jackson for an excellent literature review, where the MMFE is placed in the context of earlier attempts at modeling forecast evolution and alternative modeling approaches (Bayesian, time-series, power approximations) to managing inventory systems with uncertain demand.

The first MMFE paper was by Graves et al. (1986), who constructed a single-item version of the MMFE with independent and identically distributed (iid) demand and applied it in a two-stage setting, focusing on production smoothing and the disaggregation of an aggregate production plan across multiple items. Graves et al. (1998)

embedded the single-item MMFE with serial demand correlation within the linear systems model of Graves (1986), which corresponds to an uncapacitated system with deterministic lead times. For the case of a single-stage model and iid demand, they analytically optimized the tradeoff between production smoothing (i.e., standard deviation of production) and safety stock (i.e., standard deviation of inventory). They also showed how to use the single-stage model as a building block for a multistage acyclic system. Unaware of Graves et al. (1986), Heath and Jackson developed an MMFE for a multiproduct system that accounts for correlations in demands across products and across time periods. They used the MMFE to generate forecast updates in their simulation of an existing LP-based production-distribution system and demonstrated how improved forecasts could reduce safety stocks without affecting service performance. Using a special case of the MMFE, Güllü (1996) considered a single-item system with instantaneous but capacitated production and uncorrelated demand, and showed that the system performs better when a demand forecast for one period into the future is employed. Chen et al. (1999) illustrated their stochastic dynamic programming algorithm based on experimental design and regression splines by numerically computing the optimal ordering quantities for an inventory system with instantaneous replenishment that is driven by the MMFE.

A related paper that does not employ the MMFE is Buzacott and Shanthikumar (1994), who analyzed the safety stock versus safety time tradeoff in a capacitated continuous-time MRP system with iid demands that are known exactly over a fixed time horizon. Karaesmen et al. (1999) developed a dynamic programming formulation of a discretized, Markovian version of the Buzacott-Shanthikumar model with unit demand and productions in each period, and showed computationally that the value of advance information decreases with system utilization. There is also a stream of literature that ignores the capacitated nature of the production environment and uses alternative models to incorporate forecasts of stationary demand in inventory management decisions (e.g., Veinott 1965, Johnson and Thompson 1975, Miller 1986, Badinelli 1990, Lovejoy 1992, Drezner et al. 1996, Chen et al. 1997, Aviv 1998). To our knowledge, this paper contains the first analysis of a capacitated production-

inventory model facing a general stationary stochastic demand process and dynamic forecast updates.

The remainder of the paper is organized as follows. The MMFE is described in §2.1 and the production-inventory model is presented in §2.2. Dynamic programming is used in §2.2 to show that the optimal policy in the finite time horizon, deterministic-capacity case is a modified (by capacity restrictions) base-stock policy with respect to the forecast-corrected inventory level, which is the inventory level minus the total expected demand over the forecast horizon; we consider a stationary version of this policy in §3. Heavy traffic analysis and tail asymptotics for random walks are combined in a heuristic manner in §3.1 to obtain a closed-form expression for the forecast-corrected base-stock level that minimizes the expected steady-state inventory holding and backorder costs. In §3.2, we numerically assess the accuracy of the derived base-stock level in some special cases. The managerial implications of this analysis are provided in §4, where we interpret the results in §3.1 and use them to address the following questions: How does forecast information impact base-stock levels? How can forecast quality in the context of production-inventory management be characterized? What is the relative value of correctly specifying a time-series forecast model versus optimally using the available forecast information? Possible extensions of the model are briefly discussed in §5.

2. The Model

The single-item version of the MMFE is described in §2.1. In §2.2, we formulate the production control problem of minimizing expected steady-state inventory holding and backorder costs in a single-stage production-inventory system that is driven by forecast updates, and motivate our use of a modified forecast-corrected base-stock policy.

2.1. The Martingale Model of Forecast Evolution

In each period, as additional information becomes available, the forecaster generates a new demand forecast for a single item for all periods in the forecast horizon. The difference between this vector of demand forecasts and the one that was generated in

the previous period is called the *forecast update vector.* The MMFE is a descriptive model that characterizes the resulting sequence of forecast update vectors.

Let D_t denote the actual demand in period t. The demand process $\{D_t\}$ is assumed to be stationary with $E[D_t] = \lambda$. Let $D_{t,t+i}$ be the forecast made in period t for demand in period $t+i$, $i = 0, 1, \ldots$; we assume that forecasts are made after current demand information is revealed, so that $D_{t,t} = D_t$. Let H be the forecast horizon over which nontrivial forecasts are available: We assume $D_{t,t+i} = \lambda$ and $Cov(D_t, D_{t+i}) = 0$ for $i > H$. Then $\varepsilon_{t,t+i} = D_{t,t+i} - D_{t-1,t+i}$ is the forecast update for period $t+i$ demand recorded at the beginning of period t (with $\varepsilon_{t,t+i} = 0$ for $i > H$), and $\varepsilon_t = (\varepsilon_{t,t}, \varepsilon_{t,t+1}, \ldots, \varepsilon_{t,t+H})$ is the forecast update vector recorded at the beginning of period t.

Heath and Jackson essentially assume that the forecast represents the conditional expectation of demand given all available information, which implies forecasts are unbiased and forecast updates are uncorrelated (for brevity's sake, we refer readers to pages 21-22 of Heath and Jackson for details). Under these relatively mild assumptions, the MMFE posits that $\{\varepsilon_t\}$ forms an iid $N(0, \Sigma)$ sequence of random vectors. To use the MMFE in modeling the observed forecast update process, the production manager only has to estimate the components of the $(H+1) \times (H+1)$ covariance matrix Σ from historical forecast updates (see page 23 of Heath and Jackson for a discussion of parameter estimation); for convenience, we index the elements of Σ by σ_{ij}, $i, j = 0, \ldots, H$. Note that the MMFE is a model of an existing forecasting process, and not a forecasting tool.

The following proposition demonstrates the correspondence between the parameters of the MMFE and the autocovariance structure of demand. All propositions are proved in the Appendix.

Proposition 1 *If the MMFE assumptions are satisfied, then the true autocovariance structure of demand can be recovered from the covariance matrix Σ via*

$$\gamma_i \triangleq Cov(D_t, D_{t+i}) = \sum_{j=0}^{H-i} \sigma_{j,j+i} \quad for \quad i = 0, 1, \ldots, H. \tag{1}$$

Graves et al. (1986) assumed that the covariance matrix Σ was diagonal, which is equivalent to assuming that demand is iid with $Var(D_t) = \sum_{j=0}^{H} \sigma_{jj}$. Graves et al. (1998) and Heath and Jackson allowed a general covariance matrix, which arises when demand exhibits correlation. Note that this structure can capture correlations among forecast updates for future periods that are made in a given period. As noted by Heath and Jackson, the MMFE is preferable to a direct time-series approach (Box and Jenkins, 1970) in the production-inventory setting because it can model forecast updates arising from both an autoregressive moving average (ARMA) forecast model and from a variety of realistic informational structures (e.g., forecasts are generated by expert judgment, demand is known a fixed number of periods in advance, or total demand for the next quarter is known with more certainty than the breakdown by month).

The assumptions of the MMFE may fail to hold in practice. For example, if the forecaster is using an ARMA(p, q) model to forecast stationary demand, but does not specify its parameters correctly, the resulting sequence $\{\varepsilon_t\}$ of forecast update vectors will be correlated, thereby violating the MMFE assumptions. Notice that "satisfying the MMFE assumptions" is a more general concept than "correctly specifying an ARMA model" because the MMFE does not limit the forecaster to the use of time-series models only. In this section and the next, we assume that the assumptions of the MMFE are satisfied; at the end of §4, we investigate the impact of equation (1) failing to hold due to the misspecification of a time-series model by the forecaster.

2.2. The Production-Inventory Model

Our production-inventory model is a single-server discrete-time continuous-state make-to-stock queue that is driven by a forecast update process modeled by the MMFE introduced in §2.1. Let C_t be the production capacity (i.e., maximum number of service completions) in period t. We assume that the sequence $\{C_t\}$ is iid $N(\mu, \sigma_C^2)$, with $\mu > \lambda$ to ensure stability. Define I_t to be the inventory level at the end of period t. At the beginning of period t, the current demand is observed and the forecasts are updated. Based on this information, the production quantity $P_t \in [0, C_t]$ is determined. Demands are satisfied using on-hand inventory and the newly man-

ufactured units. Hence, the end-of-period inventory level evolves according to the relation $I_t = I_{t-1} + P_t - D_t$. The one-period cost is $hI_t^+ + bI_t^-$, where h and b are the unit inventory holding and backorder costs, respectively. Our goal is to find a production policy $\{P_t\}$ to minimize the total expected steady-state inventory holding and backorder costs, $hE[I_\infty^+] + bE[I_\infty^-]$.

To motivate the form of the production policy analyzed in §3, we briefly consider the dynamic programming formulation of the finite time horizon problem with deterministic capacity. It is convenient to define a new state variable $\tilde{I}_t = I_t - \sum_{i=1}^{H} D_{t,t+i}$, which we call the *forecast-corrected inventory level*; it is the current inventory level minus the total forecasted demand over the forecast horizon. This gives $\tilde{I}_t = \tilde{I}_{t-1} + P_t - (\lambda + e^T \varepsilon_t)$, which is a state evolution equation with iid random variables (throughout the paper, e is a column vector of ones). This transformation allows us to prove the following result using methods from Federgruen and Zipkin (1986b), who analyzed the production-inventory problem with iid demand.

Proposition 2 *For the finite-time horizon problem with deterministic capacity, the optimal production policy is characterized by a sequence of scalars* $\{B_1, \ldots, B_T\}$ *and has the form*

$$P_t^*(\tilde{I}_{t-1}) = \begin{cases} C_t & \text{if} \quad B_t > \tilde{I}_{t-1} + C_t; \\ B_t - \tilde{I}_{t-1} & \text{if} \quad \tilde{I}_{t-1} \le B_t \le \tilde{I}_{t-1} + C_t; \\ 0 & \text{if} \quad B_t < \tilde{I}_{t-1}. \end{cases} \tag{2}$$

The quantity B_t in (2) is an order-up-to, or base-stock, level with respect to the forecast-corrected inventory level, $\tilde{I}_{t-1}$, and the optimal production policy is a modified base-stock policy, where the modification is due to the capacity constraint $P_t \le C_t$.

In the long-run average cost case with discrete iid demand and deterministic capacity, Federgruen and Zipkin (1986a) show that a stationary modified base-stock policy is optimal. Federgruen and Zipkin (1986b) show that this policy is optimal in the infinite-horizon, discounted-cost case with continuous iid demand. We have

not attempted to generalize Proposition 2 to these cases. However, Proposition 2, Federgruen and Zipkin's results, and the attractiveness (with respect to analytical tractability and ease of implementation) of a stationary base-stock policy lead us to consider a stationary version of the modified forecast-corrected base-stock policy for the steady-state, random-capacity problem.

As is typical for make-to-stock queues (e.g., Chapter 4 of Buzacott and Shanthikumar 1993), we define this production policy in terms of a release policy. Let R_t denote the number of order releases to the production stage at the beginning of period t and Q_t be the number of items at the production stage at the end of period t, called the work-in-process (WIP) inventory. The production quantity is characterized by the release policy via $P_t = \min(Q_{t-1} + R_t, C_t)$.

As we show below, a modified base-stock policy with respect to the forecast-corrected inventory process $\tilde{I}_t$ is constructed by setting R_t equal to the aggregate forecast update over the forecast horizon plus the new demand forecast for the last period of the horizon (see Buzacott and Shanthikumar 1994 for a similar policy). Using the above notation,

$$R_t = \sum_{i=0}^{H-1} \varepsilon_{t,t+i} + D_{t,t+H} = \lambda + \sum_{i=0}^{H} \varepsilon_{t,t+i} = \lambda + e^T \varepsilon_t.$$

Note that $\{R_t\}$ is an iid sequence and a dip in projected demand can cause negative releases, but these negative releases are not of grave concern because the release policy is simply a convenient way to define the production policy.

Suppose $Q_0 = 0$ and $I_0 = s_H + \sum_{i=1}^{H} D_{0i}$, so that we initially stock enough items to satisfy the forecasted demand over the forecast horizon H plus a safety stock, s_H. Then

$$Q_t + I_t - \sum_{i=1}^{H} D_{t,t+i} = Q_t + \tilde{I}_t = s_H \quad \text{for} \quad t = 1, 2, \ldots .$$

We call s_H the *forecast-corrected base stock level.* Under this policy and the above initial conditions, our production control problem is to find the forecast-corrected base-stock level s_H^* that minimizes $hE[I_\infty^+(s_H)] + bE[I_\infty^-(s_H)]$, where we have introduced the dependence of the steady-state inventory level on the forecast-corrected base-stock level.

As a benchmark in our subsequent analysis, we also consider a *myopic policy,* where forecasts satisfying the MMFE assumption are available, but are not exploited in determining releases; that is, the production manager sets $R_t = D_t$. Note that the autocovariance structure of $\{R_t\}$ can be determined using equation (1). Under this myopic policy, and initial conditions $Q_0 = 0$ and $I_0 = s_m$, the relation $Q_t + I_t = s_m$ holds in every period. Let s_m^* denote the corresponding optimal base-stock level within this class of policies.

3. Main Results

In §3.1, we heuristically combine results from random walks and heavy traffic approximations of queues to determine closed-form expressions for s_m^* and s_H^* in heavy traffic. The accuracy of our approximate analysis is investigated using simulation in §3.2.

3.1. Analytical Results

We start by analyzing the steady-state WIP distribution using heavy traffic theory, which is an asymptotic method based on the system utilization $\rho \triangleq \lambda/\mu$ approaching one.

Proposition 3 *In heavy traffic, Q_∞ has an exponential distribution with parameter*

$$\nu = \frac{2(\mu - \lambda)}{e^T \Sigma e + \sigma_C^2} = \frac{2(\mu - \lambda)}{\gamma_0 + 2\sum_{i=1}^{H} \gamma_i + \sigma_C^2}$$

under both the forecast-corrected base-stock policy and the myopic policy.

A complementary asymptotic approach to discrete-time make-to-stock queues has been developed by Glasserman and co-workers. More specifically, Glasserman and Liu (1997) use Siegmund's (1979) method for a corrected diffusion approximation of random walks to analyze a discrete-time make-to-stock queue with iid demand. They show that $P(Q_\infty > x) = e^{-\tilde{\nu}(x+\beta)} + o(\tilde{\nu}^2)$ as $\lambda \to \mu$, $x \to \infty$ and $(\lambda - \mu)x \to$ *constant*, where $\tilde{\nu} = \frac{2(\mu-\lambda)}{\sigma_D^2 + \sigma_C^2}$ (which is consistent with ν in Proposition 3 in the iid demand case) and the correction term (specialized to the case where demand is iid normal with

variance σ_D^2) is $\beta = 0.583\sqrt{\sigma_D^2 + \sigma_C^2}$. Comparing Proposition 3 with Glasserman and Liu's result allows us to see the relative strengths of both approaches: Heavy traffic analysis yields the entire distribution rather than just the right tail, and appears to more easily incorporate non-iid demand, whereas the random walk approximation is more accurate because it incorporates a higher-order correction term.

Hereafter, we employ the following approximation to the steady-state WIP distribution:

$$P(Q_\infty = 0) = 1 - e^{-\nu\beta} \quad \text{and} \quad P(Q_\infty > x) = e^{-\nu(x+\beta)} \quad \text{for} \quad x \geq 0. \tag{3}$$

Equation (3) combines the relative strengths of both approaches in a heuristic manner. From Proposition 3, it uses the shape of the distribution (except for a pulse at the origin) and the heavy traffic term $e^T \Sigma e$, even for the myopic policy where releases are correlated. Equation (3) also uses the corrected diffusion term from Glasserman and Liu. Simulation results in Toktay (1998) show that (3) is indeed more accurate than the heavy traffic approximation in Proposition 3 in the iid demand case.

Our main analytical results are collected in Proposition 4. Because the characterization of the base stock level in Proposition 4b is difficult to work with, we consider the asymptotic case where $b >> h$ in Proposition 4c (the "a" in s_H^a is mnemonic for "asymptotic").

Proposition 4 *Using approximation (3),*

a. The optimal base-stock level under the myopic policy is $s_m^* = F_{Q_\infty}^{-1}(\frac{b}{b+h}) = \frac{1}{\nu}\ln(1 + \frac{b}{h}) - \beta$. *The optimal cost is* $C_m^* = hs_m^* + \frac{h(1-e^{-\nu\beta})}{\nu}$.

b. The optimal base-stock level under the forecast-corrected base stock policy with forecast horizon H is $s_H^* = F_W^{-1}(\frac{b}{b+h})$, *where* $W = \max\{Q_\infty + Y_0, \max_{1\leq k\leq H} Y_k\}$ *and* $Y_k = -k\lambda + \sum_{i=k+1}^{H}\sum_{j=i}^{H}\varepsilon_{t-H+i,t-H+j} - \sum_{i=1}^{k}\sum_{j=H+1}^{H+i}\varepsilon_{t-H+i,t-H+j} - \sum_{i=k+1}^{H} C_{t-H+i}$.

c. For $b >> h$, s_H^* *is well approximated by* $s_H^a = s_m^* + \mu_{Y_0} + \frac{1}{2}\sigma_{Y_0}^2\nu$, *and the optimal cost* $C_H^* = hE[I_\infty^+(s_H^*)] + bE[I_\infty^-(s_H^*)]$ *is well approximated by* $C_H^a = h(s_H^* + \frac{1}{\nu} - E[W])$, *where* μ_{Y_0} *and* $\sigma_{Y_0}^2$ *are the mean and variance of* Y_0.

3.2. Accuracy of s_H^a

Proposition 4c is the central result of the paper. However, to obtain this reasonably tractable form for the forecast-corrected base-stock level, we made a series of approximations: The heavy traffic approximation in Proposition 3, the heuristic combination of heavy traffic and random walk asymptotics in (3), and Clark's (1961) approximation and several $b >> h$ approximations in estimating $F_W(w)$ in the proof of Proposition 4c. Hence, we expect the accuracy of s_H^a in Proposition 4c to improve as the cost ratio b/h and the system utilization ρ increase, and - to a lesser extent - as the forecast horizon H decreases (see the comment under (11)).

To assess the accuracy of s_H^a, we use discrete-event simulation to estimate the steady-state cost incurred by the forecasting-production-inventory system when it is operating under a forecast-corrected base-stock policy. Let s^* denote the optimal base-stock level determined by an exhaustive search via simulation, and let $C(s)$ be the simulated cost of the forecast-corrected base-stock policy with base-stock level s. All scenarios in this paper were simulated until the 95% confidence interval width fell below 1.0% of the average cost.

b/h	2			10		
r	-0.3	0	0.3	-0.3	0	0.3
$\rho = 0.90$	4.88%	3.02%	4.41%	0.43%	0.00%	0.04%
$\rho = 0.95$	0.18%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 1: The simulated cost suboptimality of the derived base-stock level, s_H^a, for $MA(1)$ demands. Parameter values: $\sigma = \sigma_C = 10$, $\mu = 100$, $h = 1$.

For concreteness, we assume that demand follows a moving average process of lag 1, which is abbreviated by $MA(1)$. Thus, $D_t = \lambda + e_t - \theta_1 e_{t-1}$, where the e_t's are iid $N(0, \sigma^2)$. The forecast horizon is $H = 1$, and forecasts are given by $D_{t,t+1} = \lambda - \theta_1 e_t$, $D_{t,t+i} = \lambda$, $i > 1$, giving rise to forecast update vectors of the form $\varepsilon_t = (1, -\theta_1)e_t$. Table 1 compares the cost suboptimality, $\frac{C(s_1^a) - C(s^*)}{C(s^*)}$, for various values of the backorder-to-holding cost ratio b/h, the system utilization ρ and the demand correlation $r = -\theta_1$. This table reveals that s_1^a is very accurate when $\rho =$

0.95, and the cost suboptimality is below 5.0% even when $\rho = 0.9$ and $b/h = 2$. To investigate the robustness with respect to the forecast horizon H, we also simulated a case with a $MA(5)$ demand process (where $H = 5$) with $\theta_i = -0.3$ for $i = 1, 2, \ldots, 5$, and $b/h = 10$; the cost suboptimality was 5.62% when $\rho = 0.9$ and 0.00% when $\rho = 0.95$. Hence the derived base-stock level appears to be reasonably robust with respect to the forecast horizon H.

4. Discussion

In this section we record some observations about Proposition 4.

iid demand, no advance demand information. In the traditional case where demand is iid with variance σ_D^2 and no advance demand information is available, the forecast horizon $H = 0$, $Y_0 = 0$ and $W = Q_\infty$. The optimal base stock level in Proposition 4b is $s_0^* = \nu^{-1} \ln(1 + \frac{b}{h}) - \beta$, where ν reduces to $\frac{2(\mu-\lambda)}{\sigma_D^2+\sigma_C^2}$. The optimal base-stock level increases with system utilization, the variability of service capacity and the variability of demand. This result coincides with Glasserman's (1997) asymptotic result, which uses approximations to tail probabilities for random walks (e.g., Siegmund 1985) but does not use Glasserman and Liu's corrected diffusion approximation.

Correlated demands, myopic policy. If demands are correlated, but the myopic production policy is used, then equation (3) and Proposition 4a imply that positively (negatively, respectively) correlated demands increase (decrease, respectively) the base-stock level and – according to a second-order Taylor series approximation – the resulting cost.

Interpretation of Y_0. When $b >> h$, the forecast-corrected base-stock level in Proposition 4c is expressed as the myopic base stock level s_m^* plus several terms that incorporate the random variable Y_0. To interpret Y_0, let us define $F_{t,t+i} = D_{t+i} - D_{t,t+i}$, which is the error in the forecast made at time t in estimating time $t + i$ demand. It can be shown that $Y_0 = \sum_{i=1}^{H} F_{t-H,t-H+i} - \sum_{i=1}^{H} C_{t-H+i}$. The stationarity of the underlying demand and production processes implies that Y_0 is independent of t, and thus it can be interpreted as the error in the forecast of total demand over any forecast horizon of H periods minus the total production capacity over this forecast horizon.

By construction, $\{D_{t,t+i},\ i = 1, 2, \ldots, H\}$ and $\{F_{t,t+i},\ i = 1, 2, \ldots, H\}$ are given in terms of $\{\varepsilon_{t-H+i},\ i = 1, 2, \ldots, H\}$ and $\{\varepsilon_{t+i},\ i = 1, 2, \ldots, H\}$, respectively. Thus, $\sum_{i=1}^{H} D_{t,t+i}$ and $\sum_{i=1}^{H} F_{t,t+i}$ are independent. The total demand over a forecast horizon of H periods can be written as $\sum_{i=1}^{H} D_{t+i} = \sum_{i=1}^{H} D_{t,t+i} + \sum_{i=1}^{H} F_{t,t+i}$. By independence, $Var(\sum_{i=1}^{H} D_{t+i}) = Var(\sum_{i=1}^{H} D_{t,t+i}) + Var(\sum_{i=1}^{H} F_{t,t+i})$. Thus, $Var(\sum_{i=1}^{H} F_{t,t+i})$ represents the total demand variability over an H-period forecast horizon that has not been resolved as of the beginning of that horizon. The total system (demand and production) variability over this horizon is $Var(\sum_{i=1}^{H} D_{t+i}) + Var(\sum_{i=1}^{H} C_{t+i})$, so $Var(Y_0)$ is the portion of total system variability over an H-period forecast horizon that is as yet unresolved at the beginning of that horizon.

Forecast quality. The interpretation of Y_0 suggests the following definition of forecast quality from the production manager's viewpoint.

Definition 1 *Let Σ_A and Σ_B correspond to two different forecasting schemes for the same demand process that both satisfy the MMFE assumptions. Then forecasting scheme A is better than forecasting scheme B if*

$$Var(\sum_{i=1}^{H} F_{t,t+i}^{A}) \leq Var(\sum_{i=1}^{H} F_{t,t+i}^{B}) \quad \forall t. \tag{4}$$

Condition (4) states that forecast quality is characterized by the unresolved demand variability over the forecast horizon. We can express this quantity in terms of the covariance matrix by $Var(\sum_{i=1}^{H} F_{t,t+i}) = \sum_{i=1}^{H} f_i^T \Sigma f_i$, where f_i is a column vector with i ones followed by $H + 1 - i$ zeroes.

Proposition 5 *If forecasting scheme A is better than forecasting scheme B according to (4), then $(s_H^a)_A \leq (s_H^a)_B$ and $(C_1^a)_A \leq (C_1^a)_B$.*

Thus, systems using a better forecast (even within the restriction of the forecast already being "good" in the sense that it is unbiased and correctly captures the covariance matrix) will hold less safety stock. A better forecast also leads to a lower cost when $H = 1$ (i.e., the demand autocovariance function has one lag); the $H > 1$ case remains open.

Interpretation of Proposition 4c: The iid demand case. Suppose demand is iid with variance σ_D^2 but advance demand information may be available. The set of covariance matrices corresponding to all possible forecast structures consistent with iid demand is given by $\sum_{iid} = \{\sum | \sum = diag(\sigma_0^2, \sigma_1^2, \ldots, \sigma_H^2), \sum_{i=0}^{H} \sigma_i^2 = \sigma_D^2\}$. In this setting, $D_t = \lambda + \sum_{i=0}^{H} \varepsilon_{t-i,t}$ where the forecast updates $\varepsilon_{t-i,t}$ are independent and $N(0, \sigma_i^2)$ for $i = 0, \ldots, H$, and $\varepsilon_{t-i,t}$ represents the number of units ordered (or cancelled) i periods in advance for delivery in period t. Here, $Var(\sum_{i=1}^{H} D_{t,t+i}) = \sum_{i=1}^{H} i\sigma_i^2$, $Var(\sum_{i=1}^{H} F_{t,t+i}) = \sum_{i=1}^{H}(H-i)\sigma_i^2$ and $Var(\sum_{i=1}^{H} D_{t+i}) = H\sigma_D^2$.

Let us consider the two extreme cases: If the forecaster has no advance information about demands then $\sum = diag(\sigma_H^2, 0, \ldots, 0)$, $Var(\sum_{i=1}^{H} F_{t,t+i}) = H\sigma_D^2$ and $Var(\sum_{i=1}^{H} D_{t,t+i}) = 0$. Although this situation corresponds to a natural forecast horizon of zero, this general framework allows us to make a consistent (i.e., common forecast horizon) comparison of the forecast-corrected base stock levels under forecasting processes with $\sum \in \sum_{iid}$. At the other extreme, if the forecaster receives exact information about demand H periods ahead then $\sum = diag(0, 0, \ldots, \sigma_H^2)$, $Var(\sum_{i=1}^{H} F_{t,t+i}) = 0$ and $Var(\sum_{i=1}^{H} D_{t,t+i}) = H\sigma_D^2$.

By Proposition 4c, $s_H^a = s_m^* + \mu_{Y_0} + \frac{1}{2}\sigma_{Y_0}^2 v = s_m^* - H\mu + [Var(\sum_{i=1}^{H} F_{t,t+i}) + Var(\sum_{i=1}^{H} C_{t+i})]\frac{(\mu-\lambda)}{e^T\Sigma e + \sigma_C^2}$. If demand is iid then $s_m^* = s_0^*$, $e^T\Sigma e = \sigma_D^2$ and

$$s_H^a = s_0^* - H\lambda - H(\mu - \lambda)\left(\frac{\sum_{i=1}^{H} i\sigma_i^2}{H\sigma_D^2 + H\sigma_C^2}\right). \tag{5}$$

Equation (5) reduces to $s_H^a = s_0^* - H\lambda$ and $s_H^a = s_0^* - H\lambda - H(\mu - \lambda)(\frac{H\sigma_D^2}{H\sigma_D^2 + H\sigma_C^2})$, respectively, in the two extreme cases of no advance information and full advance information. In all cases, $0 \leq \frac{\sum_{i=1}^{H} i\sigma_i^2}{H\sigma_D^2 + H\sigma_C^2} \leq 1$; the right inequality is tight in the full information case with deterministic capacity, which leads to the minimum base stock level, $s_H^a = s_0^* - H\mu$.

Equation (5) is our most revealing result. It shows that the reduction in the base-stock level from the case of no advance information is the product of two factors: The first factor is $H(\mu - \lambda)$, which is the total expected excess production capacity over the forecast horizon. The second factor in the reduction term is a fraction: The numerator is the demand variability over the forecast horizon that has been resolved already,

and the denominator is the total system (i.e., demand and production) variability over the forecast horizon. Thus the reduction is that fraction of expected excess production capacity over the forecast horizon that equals the proportion of the total variability over the forecast horizon that has already been resolved. This highlights the interchangeability of excess production capacity and safety stock as alternative resources in make-to-stock queues: The earlier the demand variability is resolved, the more the system can rely on excess production capacity (as opposed to safety stock) to counter demand variability.

Interpretation of Proposition 4c: The general case. For the non-iid demand case, repeating the calculations leading to (5) yields

$$s_H^a = s_m^* - H\lambda - H(\mu-\lambda)\left(1 - \frac{Var(\sum_{i=1}^{H} F_{t,t+i}) + H\sigma_C^2}{He^T\Sigma e + H\sigma_C^2}\right). \quad (6)$$

Here, $s_m^* - H\lambda$ does not represent an appropriate benchmark for comparison when demand is not iid, because it assumes $\sum_{i=1}^{H} D_{t,t+i} = H\lambda$ (which violates the MMFE assumption) and $\sum_{i=1}^{H} D_{t,t+i}$ is independent of Q_t (which ignores information about demand correlation contained in previous demands). Moreover, $\frac{Var(\sum_{i=1}^{H} F_{t,t+i})+H\sigma_C^2}{He^T\Sigma e+H\sigma_C^2}$ is not a proper fraction because $e^T\Sigma e$ is the variance of the limiting Brownian motion corresponding to the demand process, and hence $He^T\Sigma e$ is a measure of – but does not equal – the total demand variance over the forecast horizon. Nonetheless, by (6), it still follows that the optimal base-stock level increases with the amount of unresolved system variability over the forecast horizon.

Heavy traffic limit. From equations (5) and (6), we see that $s_H^a \to s_m^* - H\lambda$ as the system utilization $\rho \to 1$. Hence, the value of using forecast information vanishes as the traffic intensity approaches unity. In this case, there is very little excess capacity (i.e., $H(\mu - \lambda)$ is small) to satisfy unresolved forecasted demand, and safety stock must be used instead. This observation is similar in spirit to an observation in §5.6 of Markowitz and Wein (1998), who show analytically that reducing due-date lead times in a make-to-stock system does not improve system performance in the heavy traffic limit. This observation is also consistent with the computational results in Karaesmen et al.

Forecast bias and temporal forecast aggregation. The MMFE assumes that the sequence $\{\varepsilon_t\}$ of forecast update vectors is iid $N(0, \Sigma)$. The remainder of this section explores the two natural violations of this assumption by an erroneous forecaster: The mean of ε_t is not zero (due to forecast bias) and the sequence $\{\varepsilon_t\}$ is not iid (which can occur because the forecaster has not used all available information about past demands).

Proposition 4 implies that if the forecaster overestimates (underestimates, respectively) the average demand λ, then the production manager using this forecast will carry too much (too little, respectively) safety stock, leading to suboptimal performance. Here we discuss a more subtle type of forecast bias, where the forecast update vector ε_t is such that $E[e^T\varepsilon_t] = 0$ but $E[\varepsilon_{t,t+i}] = m_i \neq 0, \quad i = 0, 1, \ldots, H$. That is, the mean demand λ has been identified correctly but the forecast made at time t for demand at time $t+i$ is biased (i.e., $E[D_{t,t+i}] = \lambda + \sum_{j=i}^{H} m_i$).

If the production manager recognizes the existence of the bias, then he can reduce the base-stock level s_H^a in Proposition 4c (which is optimal for unbiased forecasts) by the bias in the forecast of total demand over the forecast horizon H, $\sum_{i=1}^{H} im_i$, and performance will not suffer. If this bias is not recognized then the base stock level and the resulting cost will be suboptimal. However, in this case we suspect that temporal aggregation of the forecast might improve performance by decreasing the forecast bias; for example, if a forecaster has an unbiased estimate of the total demand over the next quarter, but biased estimates of the individual monthly demands, then performance might be improved by using a quarterly forecast in production decisions rather than a monthly forecast. However, a discrete-time model – where releases and demands occur with the same frequency and costs are assessed at the end of the period – is unable to accurately capture the cost tradeoff between improved forecast accuracy and the loss of information and increased system variability due to temporal aggregation.

Relative value of forecast model specification versus forecast use. In the remainder of this section, we assume that the forecaster takes a time-series approach, the first step of which is forecast model specification. The effective management of

our forecasting-production-inventory system requires the correct specification of the forecast model by the forecaster and the appropriate use of the forecast information by the production manager. However, each of these steps may be performed incorrectly in practice, and we conclude this section by using discrete-event simulation to investigate this possibility. More specifically, we consider three different scenarios to compare the impact of misspecifying the forecast model versus the impact of failing to incorporate forecast information from a correctly specified model into the production decisions. In all three scenarios, we assume that demand has an autocovariance function of H lags. In the first scenario, the forecaster incorrectly thinks that demand is iid and no advance demand information is available. In this case, the forecaster's update vector ε_t contains only the element $\varepsilon_{tt} = D_t - \lambda$. The production manager sets the forecast horizon $H = 0$ and recovers only the unconditional demand variance. (Notice that, unbeknownst to the production manager, the sequence $\{\varepsilon_t\}$ is correlated and does *not* satisfy the iid assumption of the MMFE.) By an adaptation of Proposition 4a to this case, we see that the production manager who receives this incorrect forecast employs a base stock policy with respect to the actual inventory level with the suboptimal base stock level $s_{iid} = \frac{1}{\nu_{iid}} \ln(1 + \frac{b}{h}) - \beta_{iid}$, where $\nu_{iid} = \frac{2(\mu-\lambda)}{\sigma_D^2+\sigma_C^2}$ and $\beta_{iid} = 0.583\sqrt{\sigma_D^2 + \sigma_C^2}$.

To assess the cost ramifications of the forecaster's erroneous iid assumption, we consider a second scenario where a different forecaster correctly specifies the forecast model, but the production manager who receives this forecast fails to use it in an optimal manner, instead employing the myopic production policy defined at the end of §2, with the base stock level s_m^* specified in Proposition 4a. Note that here the production manager recovers the true autocovariance structure of demand and uses it in calculating s_m^*. A comparison of the first two scenarios allows us to isolate the cost of forecast model misspecification because both production managers are using the release policy $R_t = D_t$.

Finally, we consider a third scenario where the forecast-corrected base-stock policy is used, with base-stock level s_1^a from Proposition 4c. Comparison of the second and third scenarios allows us to evaluate the impact of using forecast information to determine the production policy: The autocovariance structure of the demand process

is correctly identified in both cases, but the myopic policy fails to use the forecast information in its production decisions.

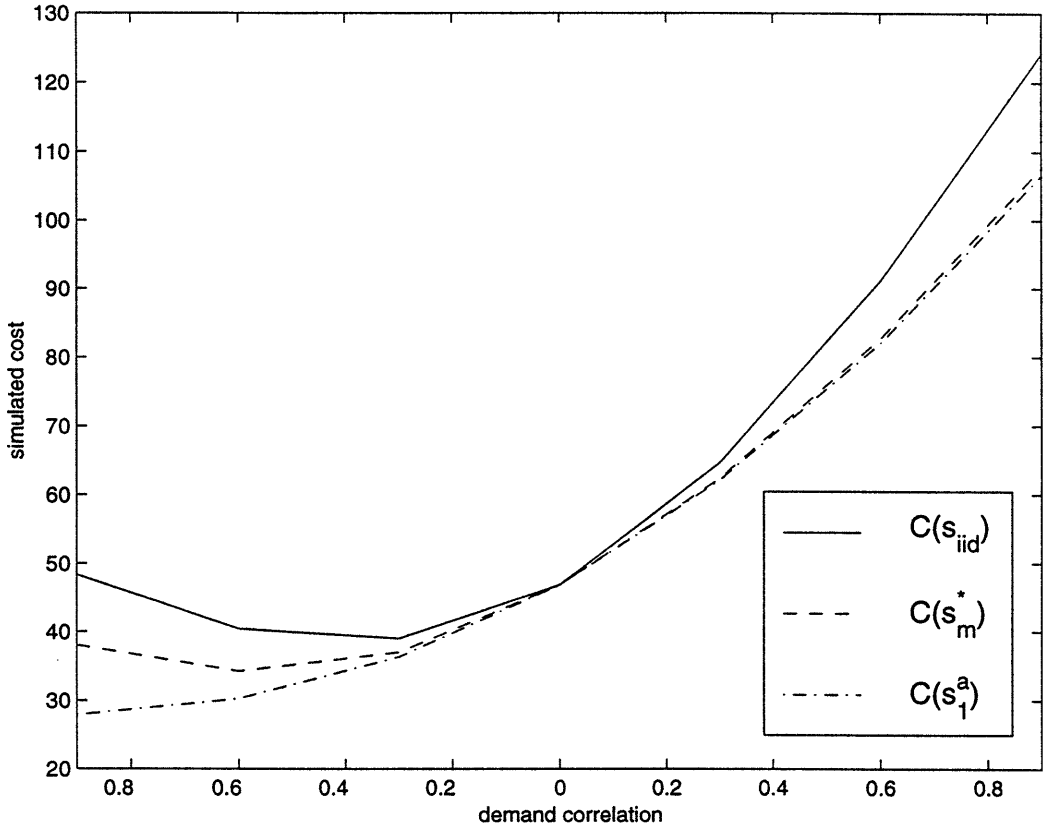

Figure 1: Comparison of simulated costs under three scenarios for $MA(1)$ demands. In the top curve, the forecaster incorrectly believes that demand is iid; in the middle curve, the forecast model is correctly specified but the myopic policy is used; in the lower curve, the forecast model is correctly specified and the forecast-corrected base-stock policy is used. Parameters: $\sigma = \sigma_C = 10$, $\lambda = 95$, $\mu = 100$, $h = 1$ and $b = 10$.

The simulated costs $C(s_{iid})$, $C(s_m^*)$ and $C(s_1^a)$ are plotted as a function of the demand correlation r in Figure 1 for a $MA(1)$ demand process. Comparing the first two scenarios (the top two curves in Figure 1), we see that when the demand correlation $r = 0$, demand is iid and $C(s_m^*) = C(s_{iid})$. As expected, the difference between the cost curves increases as $|r| \to 1$; in fact, using Proposition 4a we can show that the cost difference between the top two curves is locally convex in r at

$r = 0$.

A comparison of the bottom two curves in Figure 1 shows that the forecast-corrected base-stock policy outperforms the myopic policy, as expected. For a fixed value of $|r|$, the cost deviation between the bottom two curves in Figure 1 is more severe when demand is negatively correlated. This can be explained by equation (6), which reduces in this case to $s_1^a = s_m^* - \mu + \frac{\sigma^2+\sigma_\epsilon^2}{2(1+r)\sigma^2+\sigma_\epsilon^2}(\mu - \lambda)$. Simulation results in Toktay demonstrate more generally that the lowest cost is achieved when the production manager uses forecast information about demands over the entire forecast horizon, rather than just a portion of the horizon.

A comparison of all three curves in Figure 1 demonstrates that greater cost increases are incurred if the forecaster incorrectly specifies the forecast model than if the production manager fails to use forecasts from a correctly-specified model in his production decisions. We have been unable to assess analytically whether this claim holds in broad generality. However, a similar set of curves (not shown here) for a $MA(5)$ demand process gives the same qualitative results. Hence, our results suggest that under reasonably high system utilization, the main value of forecasting stems from the specification stage, which arguably is not carried out thoroughly in practice. It would thus be of considerable value to design production-inventory policies that are robust to forecast model misspecification.

5. Extensions

Two natural extensions of this work are to multiple items and multiple stages. Heavy traffic theory for continuous-time multiclass queues is well developed and Heath and Jackson have developed a multi-item MMFE, suggesting that the multi-item extension should not be very difficult. However, care must be taken in defining – and analyzing in heavy traffic – a queueing discipline in discrete time. An interesting empirical and theoretical question in this context is the relative contribution to the safety stock of the demand correlation across time versus the demand correlation across products. Extending our analysis to a multistage model would be much more difficult, in light of the analytical intractability of Markovian tandem make-to-stock queueing systems

(e.g., Veatch and Wein 1994). One tractable approach (see Rubio and Wein 1996) is to employ a CONWIP policy that maintains the total forecast-corrected WIP plus finished goods inventory (i.e., $Q_t^s + I_t - \sum_{i=1}^{H} D_{t,t+i}$, where Q_t^s is the system-wide WIP) at a constant base-stock level.

Appendix

Proof of Proposition 1: The MMFE assumptions imply that $D_t = \lambda + \sum_{j=0}^{H} \varepsilon_{t-H+j,t}$. Therefore

$$\begin{aligned}
\gamma_i &\triangleq Cov(D_t, D_{t+i}) = E[(D_t - \lambda)(D_{t+i} - \lambda)] \\
&= E[(\sum_{j=0}^{H} \varepsilon_{t-H+j,t})(\sum_{j=0}^{H} \varepsilon_{t+i-H+j,t+i})] \\
&= \sum_{j=0}^{H-i} E[\varepsilon_{t+i-H+j,t}\varepsilon_{t+i-H+j,t+i}] \text{ because forecast updates are uncorrelated} \\
&= \sum_{j=0}^{H-i} \sigma_{H-i-j,H-j} = \sum_{j=0}^{H-i} \sigma_{j,j+i}.
\end{aligned}$$

Proof of Proposition 2: The dynamic programming algorithm for the finite time horizon problem is

$$\begin{aligned}
J_{T+1}(\tilde{I}_T) &= 0 \text{ (assuming no salvage value or disposal cost)} \\
J_t(\tilde{I}_{t-1}) &= \min_{0 \le P_t \le C_t} \{E_{\varepsilon_t}[h(\tilde{I}_{t-1} + P_t + \sum_{i=1}^{H-1} D_{t-1,t+i} - \varepsilon_{tt})^+ \\
&\quad + b(\tilde{I}_{t-1} + P_t + \sum_{i=1}^{H-1} D_{t-1,t+i} - \varepsilon_{tt})^-] \\
&\quad + E_{\varepsilon_t}[J_{t+1}(\tilde{I}_{t-1} + P_t - (\lambda + e^T\varepsilon_t))]\}, \quad t \in \{1, 2, ..., T\}.
\end{aligned}$$

Here, $\tilde{I}_{t-1} + P_t + \sum_{i=1}^{H-1} D_{t-1,t+i} - \varepsilon_{tt} = I_{t-1} + P_t - D_t = I_t$ and $\tilde{I}_{t-1} + P_t - (\lambda + e^T\varepsilon_t) = \tilde{I}_t$. If we define $y_t = \tilde{I}_{t-1} + P_t$, $w_t = \varepsilon_{tt} - \sum_{i=1}^{H-1} D_{t-1,t+i}$, $H_t(y_t) = hE_{\varepsilon_t}[(y_t - w_t)^+] + bE_{\varepsilon_t}[(y_t - w_t)^-]$ and $G_t(y_t) = H_t(y_t) + E_{\varepsilon_t}[J_{t+1}(y_t - (\lambda + e^T\varepsilon_t)]$, then the dynamic programming algorithm can be expressed as

$$\begin{aligned}
J_{T+1}(\tilde{I}_t) &= 0 \\
J_t(\tilde{I}_{t-1}) &= \min_{\tilde{I}_{t-1} \le y_t \le \tilde{I}_{t-1} + C_t} G_t(y_t), \quad t \in \{1, 2, ..., T\}.
\end{aligned}$$

It can be shown inductively as in Federgruen and Zipkin (1986b) that $G_t(y_t)$ is convex and $\lim_{|y_t|\to\infty} G_t(y_t) = \infty$. This implies $G_t(y_t)$ has an unconstrained minimum with respect to y_t for all t. Let $B_t = \arg\min_{y_t} G_t(y_t)$ and $y_t^* = \arg\min_{\tilde{I}_{t-1} \le y_t \le \tilde{I}_{t-1}+C_t} G_t(y_t)$. Then

$$y_t^* = \begin{cases} \tilde{I}_{t-1} + C_t & \text{if} \quad \tilde{I}_{t-1} < B_t - C_t; \\ B_t & \text{if} \quad B_t - C_t \le \tilde{I}_{t-1} \le B_t; \\ \tilde{I}_{t-1} & \text{if} \quad \tilde{I}_{t-1} > B_t, \end{cases}$$

which yields (2).

The induction proof uses the facts that for all t, $H_t(y_t)$ is convex in y_t, $\lim_{|y_t|\to\infty} H_t(y_t) = \infty$, $J_t \ge 0$, and if B_t minimizes G_t then

$$J_t(\tilde{I}_{t-1}) = \begin{cases} G_t(\tilde{I}_{t-1} + C_t) & \text{if} \quad \tilde{I}_{t-1} < B_t - C_t; \\ G_t(B_t) & \text{if} \quad B_t - C_t \le \tilde{I}_{t-1} \le B_t; \\ G_t(\tilde{I}_{t-1}) & \text{if} \quad \tilde{I}_{t-1} > B_t \end{cases}$$

is a convex function.

Proof of Proposition 3: Let $E_t = R_t - C_t$ and $X_n = \sum_{t=1}^n E_t$. Then $Q_n = X_n - \inf_{1\le t\le n} X_t$. Take a sequence of systems indexed by k with $\lambda^{(k)} \to \lambda$ and $\mu^{(k)} \to \mu$ such that $\sqrt{k}(\lambda^{(k)} - \mu^{(k)}) \to c < 0$, where $c = O(1)$. Note that $Q_n^{(k)} = X_n^{(k)} - \inf_{1\le t\le n} X_t^{(k)} = X_n^{(k)} - nm^{(k)} + nm^{(k)} + Y_n^{(k)}$ where $m^{(k)} = \lambda^{(k)} - \mu^{(k)}$ and $Y_n^{(k)} = -\inf_{1\le t\le n} X_t^{(k)}$. If we define $\tilde{Q}^k(t) = \frac{Q^{(k)}_{\lfloor kt \rfloor}}{\sqrt{k}}$, $\tilde{X}^k(t) = \frac{X^{(k)}_{\lfloor kt \rfloor}}{\sqrt{k}}$ and $\tilde{Y}^k(t) = \frac{Y^{(k)}_{\lfloor kt \rfloor}}{\sqrt{k}}$, then $\tilde{Q}^k(t) = \tilde{X}^k(t) + \tilde{Y}^k(t)$ where $\tilde{X}^k(t) = \frac{X^{(k)}_{\lfloor kt \rfloor} - m^{(k)}\lfloor kt \rfloor}{\sqrt{k}} + \frac{m^{(k)}\lfloor kt \rfloor}{\sqrt{k}}$. By the functional central limit theorem for $\varphi-$mixing processes (Thm. 20.1 in Billingsley 1968), $\frac{X^{(k)}_{\lfloor kt \rfloor} - m^{(k)}\lfloor kt \rfloor}{\sqrt{k}} \Rightarrow \sigma B(t)$, where $\sigma^2 = Var(R_0 - C_0) + 2\sum_{t=1}^\infty Cov(R_0 - C_0, R_t - C_t)$ and B is standard Brownian motion. Since the $\{C_t\}$ are iid and independent of the releases, $\sigma^2 = Var(R_0) + Var(C_0) + 2\sum_{t=1}^\infty Cov(R_0, R_t)$. Under the forecast-corrected base-stock policy, releases are iid with $R_t \sim N(\lambda, e^T\Sigma e)$ and $\sigma^2 = e^T\Sigma e + \sigma_C^2$. Under the myopic policy, $R_t = D_t$ and $Var(R_0) + 2\sum_{t=1}^\infty Cov(R_0, R_t) = Var(D_0) + 2\sum_{t=1}^H Cov(D_0, D_t) = \gamma_0 + 2\sum_{t=1}^H \gamma_t = e^T\Sigma e$, where the last equality follows from equation (1). Furthermore, the assumption $\sqrt{k}(\lambda^{(k)} - \mu^{(k)}) \to c$ implies that $\frac{m^{(k)}\lfloor kt \rfloor}{\sqrt{k}} \to ct$. Therefore, under both release policies, $\tilde{X}^k \Rightarrow BM(c, e^T\Sigma e + \sigma_C^2)$, which is a Brownian motion

with drift θ and variance $e^T\Sigma e + \sigma_C^2$. If we define $\phi(x)(t) = -\inf_{0\le s\le t}\{x(s)\}$ and $\psi(x)(t) = x + \phi(x)(t)$, where x is a right-continuous left-limit process with $x(0) = 0$, then $(\tilde{Q}^k(t), \tilde{Y}^k(t)) = (\psi, \phi)(\tilde{X}^k(t))$. Since we have established that $\tilde{X}^k \Rightarrow X^*$, where X^* is $B(c, e^T\Sigma e + \sigma_C^2)$, a process with continuous sample paths, the continuous mapping theorem (Thm. 5.1 in Billingsley) implies that $(\psi, \phi)(\tilde{X}^k)$ converges to $(\psi, \phi)(X^*)$, i.e, $\tilde{Q}^k \Rightarrow Q^*$, where Q^* is $RBM(c, e^T\Sigma e + \sigma_C^2)$ on [0,∞), which is a reflected Brownian motion on the nonnegative halfline. The steady-state distribution of $RBM(c, \sigma^2)$ is exponential with parameter $\frac{\sigma^2}{2|c|}$ (Harrison 1985). Reversing the heavy traffic scaling, we estimate the steady-state WIP Q_∞ by an exponential random variable with parameter $\sqrt{k}\frac{e^T\Sigma e+\sigma_C^2}{2|c|} = \frac{e^T\Sigma e+\sigma_C^2}{2(\mu-\lambda)}$.

Proof of Proposition 4:

4a. Since $I_t = s_m - Q_t$ for all t, we have $I_\infty = s_m - Q_\infty$. Let $C(s_m) = hE[I_\infty^+(s_m)] + bE[I_\infty^-(s_m)]$, which is strictly convex in s_m. Setting $C'(s_m) = 0$ yields $s_m^* = F_{Q_\infty}^{-1}(\frac{b}{h+b}) = \frac{1}{\nu}\ln(1+\frac{b}{h}) - \beta$. Furthermore,

$$\begin{aligned} C(s_m^*) &= h\int_{x=0}^{s_m^*}(s_m^* - x)f_{Q_\infty}(x)dx + b\int_{x=s_m^*}^{\infty}(x - s_m^*)f_{Q_\infty}(x)dx \\ &= hs_m^* + h\frac{e^{-\nu(s_m^*+\beta)} - e^{-\nu\beta}}{\nu} + b\frac{e^{-\nu(s_m^*+\beta)}}{\nu} \\ &= hs_m^* + \frac{h(1-e^{-\nu\beta})}{\nu}. \end{aligned}$$

4b. As in part 4a, we need to find the distribution of I_∞. Recall that $I_t = s_H - (Q_t - \sum_{i=1}^H D_{t,t+i})$ for all t. Since Q_t and $\sum_{i=1}^H D_{t,t+i}$ are dependent, I_∞ cannot be expressed as a function of Q_∞ and the unconditional distribution of $\sum_{i=1}^H D_{t,t+i}$. Hence, our objective is to write $Q_t - \sum_{i=1}^H D_{t,t+i}$ as the sum of independent random variables.

The queue length process $\{Q_t\}$ is a reflected random walk on $[0,\infty)$ with unrestricted step sizes $R_t - C_t$, and can be expressed as

$$Q_t = \max\{Q_{t-H} + \sum_{k=t-H+1}^{t}(R_k - C_k), \sum_{k=t-H+2}^{t}(R_k - C_k), \ldots, R_t - C_t, 0\}. \quad (7)$$

We can write $R_{t-H+i} = \lambda + f_{H+1}^T\varepsilon_{t-H+i}$, $i = 1, 2, \ldots, H$ and $\sum_{i=1}^H D_{t,t+i} = \lambda H + \sum_{i=1}^H g_i^T\varepsilon_{t-H+i}$, where f_i (g_i, respectively) denotes an $(H+1)$-dimensional

column vector whose first (last, respectively) i elements are one and the rest are zero. Substituting (7) into $I_t = s_H - (Q_t - \sum_{i=1}^{H} D_{t,t+i})$ yields $I_t = s_H - \max\{Q_{t-H} + Y_{t0}, \max_{1\le k\le H} Y_{tk}\}$, where $Y_{tk} = -k\lambda + \sum_{i=k+1}^{H} f_{H+1-i}^T \varepsilon_{t-H+i} - \sum_{i=1}^{k} g_i^T \varepsilon_{t-H+i} - \sum_{i=k+1}^{H} C_{t-H+i}$, $k = 0, 1, \ldots, H$. Hence, $Y_t = (Y_{t0}, Y_{t1}, \ldots, Y_{tH})$ is an $(H+1)$-dimensional multivariate normal vector with $E[Y_{tk}] = -k\lambda - (H-k)\mu$, $Var(Y_{tk}) = (H-k)\sigma_C^2 + \sum_{i=1}^{k} g_i^T \Sigma g_i + \sum_{i=k+1}^{H} f_{H+1-i}^T \Sigma f_{H+1-i}$ and $Cov(Y_{tk}, Y_{tl}) = (H-l)\sigma_C^2 + \sum_{i=1}^{k} g_i^T \Sigma g_i - \sum_{i=k+1}^{l} f_{H+1-i}^T \Sigma g_i + \sum_{i=l+1}^{H} f_{H+1-i}^T \Sigma f_{H+1-i}$ for $k < l$. Note that the distribution of Y_t is independent of t; let Y be a generic random variable with this distribution. Since Q_{t-H} depends only on ε_k and C_k for $k \le t-H$, Y_t is independent of Q_{t-H}. These properties yield the characterization $I_\infty = s_H - \max\{Q_\infty + Y_0, \max_{1\le k\le H} Y_k\}$. If we define $W = \max\{Q_\infty + Y_0, \max_{1\le k\le H} Y_k\}$, then $I_\infty = s_H - W$ and by an argument similar to that in part 4a, we have $s_H^* = F_W^{-1}(\frac{b}{b+h})$.

4c. Let us approximate $Z \triangleq \max_{1\le k\le H} Y_k$ by a $N(\mu_Z, \sigma_Z^2)$ random variable as in Clark. If we define $\alpha = Cov(Y_0, Z)/(\sigma_{Y_0}\sigma_Z)$ then

$$
\begin{aligned}
P(W \le w) &= \int_{y_0=-\infty}^{w}\int_{z=-\infty}^{w} (1 - e^{-\nu(w+\beta-y_0)}) f_{Y_0 Z}(y_0, z)\,dz\,dy_0 \\
&= \int_{y_0=-\infty}^{w}\int_{z=-\infty}^{w} f_{Y_0 Z}(y_0, z)\,dz\,dy_0 \\
&\quad -e^{-\nu(w+\beta)} \int_{z=-\infty}^{w} f_Z(z) \int_{y_0=-\infty}^{w} e^{\nu y_0} f_{Y_0|Z}(y_0|z)\,dy_0\,dz \\
&\approx P(\max\{Y_0, Z\} \le w) \\
&\quad -e^{-\nu(w+\beta)} \int_{z=-\infty}^{w} f_Z(z) e^{(\mu_{Y_0} + \alpha\frac{\sigma_{Y_0}}{\sigma_Z}(z-\mu_Z))\nu + \frac{\sigma_{Y_0}^2(1-\alpha^2)}{2}\nu^2} \\
&\quad \Phi\left(\frac{w - (\mu_{Y_0} + \alpha\frac{\sigma_{Y_0}}{\sigma_Z}(z-\mu_Z) + \sigma_{Y_0}^2(1-\alpha^2)\nu)}{\sigma_{Y_0}\sqrt{1-\alpha^2}}\right) dz. \qquad (8)
\end{aligned}
$$

Let us approximate $P(W \le w)$ for large w. To this end, approximate the $\Phi(\cdot)$

term in (8) by one for large w. Then

$$
\begin{aligned}
& e^{-\nu(w+\beta)} \int_{z=-\infty}^{w} f_Z(z) e^{(\mu_{Y_0}+\alpha\frac{\sigma_{Y_0}}{\sigma_Z}(z-\mu_Z))\nu+\frac{\sigma_{Y_0}^2(1-\alpha^2)}{2}\nu^2} dz \\
= \; & e^{-\nu(w+\beta)} e^{\mu_{Y_0}\nu-\alpha\frac{\sigma_{Y_0}}{\sigma_Z}\mu_Z\nu+\frac{\sigma_{Y_0}^2(1-\alpha^2)}{2}\nu^2} \int_{z=-\infty}^{w} e^{\alpha\frac{\sigma_{Y_0}}{\sigma_Z}\nu z} f_Z(z) dz \\
= \; & e^{-\nu(w+\beta)} e^{\mu_{Y_0}\nu-\alpha\frac{\sigma_{Y_0}}{\sigma_Z}\mu_Z\nu+\frac{\sigma_{Y_0}^2}{2}\nu^2-\frac{\sigma_{Y_0}^2\alpha^2}{2}\nu^2} \\
& e^{\mu_Z\alpha\frac{\sigma_{Y_0}}{\sigma_Z}\nu+\frac{1}{2}\alpha^2\frac{\sigma_{Y_0}^2}{\sigma_Z^2}\sigma_Z^2\nu^2} \Phi\left(\frac{w-(\mu_Z+\alpha\frac{\sigma_{Y_0}}{\sigma_Z}\nu\sigma_Z^2)}{\sigma_Z}\right) \\
= \; & e^{-\nu(w-\mu_{Y_0}-\frac{\sigma_{Y_0}^2}{2}\nu+\beta)} \Phi\left(\frac{w-(\mu_Z+\alpha\frac{\sigma_{Y_0}}{\sigma_Z}\nu\sigma_Z^2)}{\sigma_Z}\right). \qquad (9)
\end{aligned}
$$

Equations (8)-(9) imply that

$$
\begin{aligned}
P(W \leq w) & \approx P(\max\{Y_0, Z\} \leq w) - e^{-\nu(w-\mu_{Y_0}-\frac{\sigma_{Y_0}^2}{2}\nu+\beta)} \Phi\left(\frac{w-(\mu_Z+\alpha\frac{\sigma_{Y_0}}{\sigma_Z}\nu\sigma_Z^2)}{\sigma_Z}\right) \\
& = P(\max_{0\leq k\leq H} Y_k \leq w) - e^{-\nu(w-\mu_{Y_0}-\frac{1}{2}\sigma_{Y_0}^2\nu+\beta)} \Phi\left(\frac{w-(\mu_Z+\alpha\frac{\sigma_{Y_0}}{\sigma_Z}\nu\sigma_Z^2)}{\sigma_Z}\right). \qquad (10)
\end{aligned}
$$

Equation (10) and Clark's algorithm provide an approximate numerical solution to $s_H^* = F_W^{-1}(\frac{b}{b+h})$, but because we are interested in a closed-form approximation, we make a further "large w" approximation that yields

$$
P(W \leq w) \approx 1 - e^{-\nu(w-\mu_{Y_0}-\frac{\sigma_{Y_0}^2}{2}\nu+\beta)} \quad \text{for large } w. \qquad (11)
$$

Note that the goodness of the approximation $P(\max_{0\leq k\leq H} Y_k \leq w) \approx 1$ for large w decreases with H, and thus the accuracy of our base-stock level should degrade for large forecast horizons.

In the proof of part 4b, we showed that s_H^* solves $F_W(s_H^*) = \frac{b}{b+h}$. Hence for $b >> h$, we use (11) to approximate s_H^* by s_H^a, where $\frac{b}{b+h} = 1 - e^{-\nu(s_H^a-\mu_{Y_0}-\frac{1}{2}\sigma_{Y_0}^2\nu+\beta)}$. Also, by Proposition 4a, $\frac{b}{b+h} = 1 - e^{-\nu(s_m^*+\beta)}$. Equating these two expressions gives $s_H^a = s_m^* + \mu_{Y_0} + \frac{1}{2}\sigma_{Y_0}^2\nu$.

Turning to the cost, we obtain

$$\begin{aligned} C_H^* &= hE[I_\infty^+(s_H^*)] + bE[I_\infty^-(s_H^*)] \\ &= h\int_{-\infty}^{s_H^*}(s_H^* - w)f_W(w)dw + b\int_{s_H^*}^{\infty}(w - s_H^*)f_W(w)dw \\ &= -h\int_{-\infty}^{s_H^*} w f_W(w)dw + b\int_{s_H^*}^{\infty} w f_W(w)dw. \end{aligned} \tag{12}$$

Adding and subtracting $h\int_{s_H^*}^{\infty} w f_W(w)dw$ to (12) and collecting terms yield

$$C_H^* = (b+h)\int_{s_H^*}^{\infty} w f_W(w)dw - hE[W]. \tag{13}$$

The interval $[s_H^*, \infty)$ is on the tail of the distribution of W for $b >> h$, so we use (11) to approximate $\int_{s_H^*}^{\infty} w f_W(w)dw$ by $\int_{s_H^a}^{\infty} w\nu e^{-\nu(w-\mu_{Y_0}-\frac{1}{2}\sigma_{Y_0}^2\nu+\beta)}dw = \frac{h}{h+b}(s_H^a + \frac{1}{\nu})$. Substituting this expression into (13) gives our approximation for C_H^*, which is $C_H^a = h(s_H^* + \frac{1}{\nu} - E[W])$.

Proof of Proposition 5: Recall that $s_H^a = s_m^* + \mu_{Y_0} + \frac{1}{2}\sigma_{Y_0}^2\nu = \frac{1}{\nu}\ln(1+\frac{b}{h}) - \beta - H\mu + \frac{1}{2}\sigma_{Y_0}^2\nu$, where $\nu = \frac{2(\mu-\lambda)}{e^T\Sigma e+\sigma_C^2}$ and $\beta = 0.583\sqrt{e^T\Sigma e + \sigma_C^2}$. Since Σ_A and Σ_B satisfy (1), we have $e^T\Sigma_A e = e^T\Sigma_B e = \gamma_0 + 2\sum_{i=1}^{H}\gamma_i$. Hence $\nu_A = \nu_B$ and $\beta_A = \beta_B$. Because $\sigma_{Y_0}^2 = Var(\sum_{i=1}^{H} F_{t,t+i}) + H\sigma_C^2$, if condition (4) holds then $(\sigma_{Y_0}^2)_A \le (\sigma_{Y_0}^2)_B$. We conclude that $(s_H^a)_A \le (s_H^a)_B$.

By Proposition 4c, $C_1^a = h(s_1^* + \frac{1}{\nu} - E[W])$ where $W = \max\{Q_\infty + Y_0, Y_1\}$. $E[\max\{Y_0, Y_1\}] = \mu_{Y_0}\Phi(\alpha) + \mu_{Y_1}\Phi(-\alpha) + a\varphi(\alpha)$, where $a^2 = \sigma_{Y_0}^2 + \sigma_{Y_1}^2 - 2\sigma_{Y_0Y_1}$, $\alpha = (\mu_{Y_0} - \mu_{Y_1})/a$ and $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal pdf and cdf, respectively (Clark). In our case, $a^2 = \sigma_C^2 + e^T\Sigma e$ and $\alpha = (\lambda - \mu)/a$. Thus, $E[\max\{Y_0, Y_1\}]_A = E[\max\{Y_0, Y_1\}]_B$. Since $(s_1^a)_A \le (s_1^a)_B$ and $\nu_A = \nu_B$, we conclude that $E[W_A] = E[W_B]$ and $(C_1^a)_A \le (C_1^a)_B$.

References

Aviv, Y. 1998. The Effect of Forecasting Capabilities on Supply Chain Coordination. Working Paper, John M. Olin School of Business, Washington University, St. Louis, MO.

Badinelli, R. D. 1990. The Inventory Costs of Common Misspecifications of Demand Forecasting Models. *Int. J. Prod. Res.* **28** 2321-2340.

Billingsley, P. 1968. *Convergence of Probability Measures.* John Wiley & Sons, New York.

Box, G. E. P., G. M. Jenkins. 1970. *Times Series Analysis: Forecasting and Control.* Holden-Day, San Francisco, California-London-Amsterdam.

Buzacott, J. A., J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems.* Prentice Hall, Englewood Cliffs, NJ.

——, ——. 1994. Safety Stock versus Safety Time in MRP Controlled Production Systems. *Management Sci.* **40** 1678-1689.

Chen, V. C. P., D. Ruppert, C. A. Shoemaker. 1999. Applying Experimental Design and Regression Splines to High-Dimensional Continuous-State Stochastic Dynamic Programming. *Oper. Res.* **47** 38-53.

Chen, R. Y., J. K. Ryan, D. Simchi-Levi. 1997. The Impact of Exponential Smoothing Forecasts on the Bullwhip Effect. Working Paper, Department of Industrial Engineering and Management Science, Northwestern University, Evanston, IL.

Clark, E. C. 1961. The Greatest of a Finite Set of Random Variables. *Oper. Res.* **9** 145-162.

Drezner, Z., J. K. Ryan, D. Simchi-Levi. 1996. Quantifying the Bullwhip Effect in a Simple Supply Chain: The Impact of Forecasting, Lead Times and Information. Working Paper, Department of Industrial Engineering and Management Science, Northwestern University, Evanston, IL.

Federgruen, A., P. Zipkin. 1986a. An Inventory Model with Limited Production Capacity and Uncertain Demands. I. The Average-Cost Criterion. *Math. Oper. Res.* **11**(2) 193-207.

——, ——. 1986b. An Inventory Model with Limited Production Capacity and Uncertain Demands. II. The Discounted-Cost Criterion. *Math. Oper. Res.* **11**(2) 208-215.

Glasserman, P. 1997. Bounds and Asymptotics for Planning Critical Safety Stocks. *Oper. Res.* **45** 244-257.

——, T.-W. Liu. 1997. Corrected Diffusion Approximations for a Multistage Production-Inventory System. *Math. Oper. Res.* **22** 186-201.

Graves, S. C. 1986. A Tactical Planning Model for a Job Shop. *Oper. Res.* **34** 522-533.

——, H. C. Meal, S. Dasu, Y. Qiu. 1986. Two-Stage Production Planning in a Dynamic Environment. S. Axsäter, C. Schneeweiss and E. Silver, eds. *Multi-Stage Production Planning and Control.* Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin. **266** 9-43.

——, D. B. Kletter, W. B. Hetzel. 1998. A Dynamic Model for Requirements Planning with Application to Supply Chain Optimization. *Oper. Res.* Supplement to **46** S35-S49.

Güllü, R. 1996. On the Value of Information in Dynamic Production/Inventory Problems under Forecast Evolution. *Naval Res. Logist.* **43** 289-303.

Harrison, J. M. 1985. *Brownian Motion and Stochastic Flow Systems.* John Wiley, New York.

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.

Johnson, G. D., H. E. Thompson. 1975. Optimality of Myopic Inventory Policies for Certain Dependent Demand Processes. *Management Sci.* **21** 1303-1307.

Karaesmen, F., J. A. Buzacott, Y. Dallery. 1999. Integrating Advance Information in Pull Type Control Mechanisms for Multi-Stage Production. Working Paper, Laboratoire d'Informatique de Paris 6, Université Pierre et Marie Curie, Paris, France.

Lovejoy, W. S. 1992. Stopped Myopic Policies in Some Inventory Models with Generalized Demand Processes. *Management Sci.* **38** 688-707.

Markowitz, D. M., L. M. Wein. 1998. Heavy Traffic Analysis of Dynamic Cyclic Policies: A Unified Treatment of the Single Machine Scheduling Problem. Sloan School of Management, MIT, Cambridge, MA.

Miller, B. L. 1986. Scarf's State Reduction Method, Flexibility, and a Dependent Demand Inventory Model. *Oper. Res.* **34** 83-90.

Rubio, R., L. M. Wein. 1996. Setting Base Stock Levels using Product-Form Queueing Networks. *Management Sci.* **42** 259-268.

Siegmund, D. 1979. Corrected Diffusion Approximations for Certain Random Walk Problems. *Adv. Appl. Prob.* **11** 701-719.

——. 1985. *Sequential Analysis: Tests and Confidence Intervals.* Springer, New York.

Toktay, L. B. 1998. Analysis of a Production-Inventory System under a Stationary Demand Process and Forecast Updates. Unpublished Ph.D. dissertation, Operations Research Center, MIT, Cambridge, MA.

Veatch M., L. M. Wein. 1994. Optimal Control of a Two-Station Tandem Production/Inventory System. *Oper. Res.* **42** 337-350.

Veinott, A.F. 1965. Optimal Policy for a Multi-Product, Dynamic, Nonstationary Inventory Problem. *Management Sci.* **12** 206-222.