Simulation-Based Performance Assessment of Production Planning Models With Safety Stock and Forecast Evolution in Semiconductor Wafer Fabrication

Timm Ziarnetzky, Lars Mönch, and Reha Uzsoy, *Senior Member, IEEE*

***Abstract*—We study the performance of production planning models for wafer fabs in a rolling horizon setting under stochastic demand modeled using the Martingale Model of Forecast Evolution (MMFE). The models differ in how they incorporate lead times and safety stocks. The number of frozen periods is considered as an experimental factor to study the nervousness of the resulting plans. Our simulation experiments show that chance-constrained production planning models that integrate safety stock and production planning decisions outperform others. The formulations with workload-dependent lead times also outperform those with fixed lead times that are an integer multiple of the period length.**

***Index Terms*—Production planning, rolling horizon, forecast evolution, safety stocks.**

I. Introduction

THE SEMICONDUCTOR industry presents challenging production planning and control problems [1] whose importance for semiconductor supply chains has been increasing over time [2], [3]. In this paper we consider the problem of determining releases into a single wafer fab to optimize profit while maintaining acceptable customer service in the face of uncertain demand. Our models explicitly consider the need for safety stocks, which both practical experience [4] and theoretical insights [5], [6] show are crucial to achieving high service levels while maintaining profitability.

Research on inventory planning and production planning has evolved largely independently. Most of the literature on production planning [7], [8] has used deterministic mathematical programming models, most commonly linear programs, in which the demand in each period is represented by a point estimate. Most of these models represent cycle times, the time between work being released into the fab and its emergence as finished product that can be used to meet demand, as a fixed, exogenous parameter independent of resource utilization. However, queueing theory [9], [10], simulation models [11] and industrial observation [12] all indicate that the mean and variance of cycle times increase nonlinearly with resource utilization. Utilization, in turn, is determined by the release decisions made by the planning models. This suggests that cycle times should be treated as endogenous to the release planning problem. Mathematical programming models thus facilitate the incorporation of complex technological constraints, but ignore the pervasive stochastic aspects of both the production process and demand.

Inventory models [5], [6], on the other hand, have emphasized the stochastic nature of demand using simple models of capacity. It is well known that the amount of safety stock required to ensure a specified level of customer service depends on the distribution of the lead time demand, the total demand over the replenishment lead time. When an inventory is replenished from a production facility, the cycle time can be a large portion of the replenishment lead time. The need for safety stock requires the production of additional material, increasing resource utilization and thus the mean and variance of the lead time demand.

In this paper we propose approximate solutions to the problem of jointly planning releases and safety stocks by considering stochastic models of demand forecast evolution. Demand forecasts are an essential input to all production planning models, and are often correlated over both different products and over time because they are based on the same information such as market conditions [13], [14]. Demand forecasts evolve and decisions are revised over time as new information becomes available [15], [16], motivating the use of planning models in a rolling setting. Computational studies in simplified settings [16], [17] indicate that planning models exploiting statistical models of demand forecast evolution may outperform deterministic planning models based on point forecasts of demand [8].

In this paper we study production planning models for a wafer fab in a rolling horizon setting under stochastic demand modeled using the Martingale Model of Forecast Evolution (MMFE) [13] and different approaches to setting safety stocks [16], [18]. We also contribute to research on production planning models with workload-dependent lead times [19], [20] by assessing their performance in a rolling

Manuscript received May 26, 2019; revised October 25, 2019; accepted December 2, 2019. Date of publication December 10, 2019; date of current version February 3, 2020. *(Corresponding author: Lars Mönch.)*

T. Ziarnetzky and L. Mönch are with the Department of Mathematics and Computer Science, University of Hagen, 58097 Hagen, Germany (e-mail: timm.ziarnetzky@fernuni-hagen.de; lars.moench@fernuni-hagen.de).

R. Uzsoy is with the Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695 USA (e-mail: ruzsoy@ncsu.edu).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TSM.2019.2958526

horizon setting, as opposed to the static setting of most past work. It is not immediately evident that a planning model displaying promising performance in a static environment will maintain its advantage in a rolling setting [21]. The rolling setting also raises the issue of nervousness, the negative effects of repeated changes to planning decisions before they are actually implemented [22].

The next section describes the problem and discusses previous related work. The production planning models are introduced in Section III, and the rolling horizon framework used for the simulation experiments in Section IV. The simulation results are presented in Section V, and conclusions and future research directions in Section VI.

II. Problem Setting and Analysis

A. Problem Description

Since release planning models seek to release work into the fab to meet demand in the best possible manner, they must consider the *cycle time*, the delay between the release of work into the fab and its emerging as output. Average cycle times in wafer fabs can be of the order of 10–12 weeks, and so cannot realistically be neglected. Most release planning models use point estimates of the cycle time, referred to as *lead times*, which are treated as exogenous parameters [8], [23]. We also examine planning models that represent workload-dependent lead times with non-linear clearing functions (CFs) [19], [23].

We consider a planning window of T planning periods of equal length. Information on the state of the fab and estimated future demand are collected and demand forecasts updated at the beginning of each period. We refer to each point in time at which information is updated and a new plan is computed as a *planning epoch*. At each planning epoch s a new production plan is computed for the next T periods $s+\Delta, s+\Delta+1, \ldots, s+T-1$ where $0 \leq \Delta \leq T-1$ represents the number of frozen periods in which no changes to planning decisions computed at the previous epoch $s-\Delta-1$ are permitted to reduce nervousness. The planning window at any epoch s thus consists of Δ frozen periods and the subsequent $T-\Delta$ periods for which new release decisions can be computed using the updated demand information.

This paper is motivated by the following observations:

1. While there is some evidence from the literature that CF-based production planning models outperform those with fixed lead times if the CFs are parameterized correctly, our knowledge of the performance of CF-based models in a rolling horizon setting is limited.
2. The impact of different safety stock setting strategies on profit, service level, and nervousness in a rolling horizon environment under different planning models is not clear from the literature.
3. We examine the impact of the number of frozen periods Δ on the different performance measures.

B. Previous Related Work

The literature on integrated production and inventory planning is small. Orçun *et al.* [24] propose an iterative approach based on a chance-constrained linear programming (LP) model where chance constraints define the safety stock levels required for a given service level. The safety stock targets are updated iteratively based on the cycle times determined by a CF-based planning model, and convergence is demonstrated in numerical experiments. A similar iterative approach using a guaranteed service safety stock placement model [25] is described in [26].

A multi-stage stochastic programming model of a simplified semiconductor wafer fab that includes a stochastic model of demand evolution over time [13] is studied in [27]. Albey *et al.* [16] examine a production planning model that represents stochastic demand with the additive MMFE and uses chance constraints to capture target service levels. Similar chance-constrained approaches are considered in [28]–[30]. Ziarnetzky *et al.* [17] extend the chance-constrained approach in [16] to time-varying demand distributions using the multiplicative MMFE. Their procedure is tested in a rolling horizon setting using a simulation model of a scaled-down wafer fab [31]. The models in [16], [17] require exogenous allocation of capacity to products in order to implement the chance constraints. Albey *et al.* [32] propose a planning model under the additive MMFE model with endogenous capacity allocation that outperforms models with exogenous capacity allocation. These approaches set safety stocks using chance constraints derived from the demand distribution. Hence we shall refer to them as demand-driven safety stock approaches. Hung and Chang [18] use the cycle time distribution to compute safety stock for each product in each period. We refer to this approach as cycle time-driven safety stock approach.

Due to the evolving nature of both shop-floor state and demand information, most planning models are implemented in a rolling horizon setting. The design of rolling horizon approaches, which involves specifying the length T of the planning window and the number of frozen periods Δ, must consider the issue of nervousness [22], [33]–[36]. There is some evidence from the literature [37]–[41] that reducing the replanning frequency and freezing certain periods can reduce nervousness. However, few papers address nervousness issues in planning models with workload-dependent lead times. Lin and Uzsoy [42] compare chance-constrained production planning models with a deterministic CF-based model in a single-stage, single-item system. The chance-constrained models reduce nervousness and maintain high service levels in a rolling horizon environment. Lin and Uzsoy [43] use a LP model of a single-stage single-product production system in a rolling setting to examine the impact of penalizing planned changes in the objective function and freezing release decisions for some future periods. Ziarnetzky *et al.* [17] find that the impact of the value of T on profit and service level is limited under non-stationary demand due to the rolling horizon setting, while [44] shows that longer planning periods yield slightly improved performance by CF-based models under stationary demand. This paper addresses the issue of nervousness by incorporating safety stock through chance constraints with advance demand information in a rolling horizon setting. We also explore whether freezing a certain portion of the production plan reduces nervousness without unduly compromising profit. Both the additive and the multiplicative MMFE with

different covariance structures are used to represent a wide range of possible demand patterns.

III. Production Planning Formulations

A. Basic Formulation With Exogenous Lead Times

Consider a planning epoch s at which a new instance of the planning model must be solved for the next T periods. We assume that the release decisions for the following $0 \leq \Delta \leq T-1$ periods are frozen, and cannot be changed. Our first planning model considers exogenous lead times that are an integer multiple of the period length and remain constant over the planning window. We define the following notation:

Sets and Indices:

t: period index
g: product index
k: workcenter index
l: operation index
G: set of all products with $\Gamma := |G|$
K: set of all workcenters
$\Omega(g)$: set of all operations required by product g
$\bar{\Omega}(k)$: set of all operations performed at workcenter k

Decision Variables:

$X_t^{(g)}$: quantity of product g released to the first workcenter of its routing in period t
$Y_{tl}^{(g)}$: output of product g from operation l in period t
$Y_t^{(g)}$: output of product g from the last operation of its routing in period t
$W_t^{(g)}$: work in progress (WIP) inventory of product g at the end of period t
$I_t^{(g)}$: finished goods inventory (FGI) of product g at the end of period t
$B_t^{(g)}$: backlog of product g at the end of period t

Parameters:

$h_t^{(g)}$: unit FGI holding cost for product g in period t
$b_t^{(g)}$: unit backlogging cost for product g in period t
$\omega_t^{(g)}$: unit WIP cost for product g in period t
$D_t^{(g)}$: demand for product g during period t
C_k: capacity of workcenter k per period in units of time
$\alpha_l^{(g)}$: processing time of operation l of product g
$L_l^{(g)}$: estimated time elapsing from the release of the raw material of product g to the completion of the operation l of product g
$\hat{X}_t^{(g)}$: quantity of product g released to the first workcenter of its routing in period t, computed at epoch $s-1$.

The model can be stated as follows:

$$\min \quad \sum_{g\in G}\sum_{t=s}^{s+T-1}\left[\omega_t^{(g)}W_t^{(g)}+h_t^{(g)}I_t^{(g)}+b_t^{(g)}B_t^{(g)}\right] \tag{1}$$

$$\text{subject to} \quad Y_t^{(g)}+I_{t-1}^{(g)}-I_t^{(g)}+B_t^{(g)}-B_{t-1}^{(g)} = D_t^{(g)}, \quad g\in G, t=s,\ldots,s+T-1 \tag{2}$$

$$W_{t-1}^{(g)}+X_t^{(g)}-Y_t^{(g)} = W_t^{(g)}, g\in G, t=s,\ldots,s+T-1 \tag{3}$$

$$Y_{tl}^{(g)} = X_{t-\left\lfloor L_l^{(g)}\right\rfloor}^{(g)}, g\in G, t=s,\ldots,s+T-1, l\in\Omega(g) \tag{4}$$

$$\sum_{g\in G, l\in O(k)}\alpha_l^{(g)}Y_{tl}^{(g)} \leq C_k, t=s,\ldots,s+T-1, k\in K \tag{5}$$

$$X_t^{(g)} = \hat{X}_t^{(g)}, g\in G, t=s,\ldots,s+\Delta, \Delta\geq 0 \tag{6}$$

$$X_t^{(g)}, Y_{tl}^{(g)}, Y_t^{(g)}, W_t^{(g)}, I_t^{(g)}, B_t^{(g)} \geq 0, \quad g\in G, t=s,\ldots,s+T-1, l\in\Omega(g). \tag{7}$$

The objective function (1) minimizes the sum of WIP, FGI, and backlog costs over all products and periods. Constraints (2) ensure conservation of flow for FGI, while the WIP balance constraints (3) allow computing the WIP cost in the objective function. Initial WIP is considered as proposed in [7] where the capacity required to process the initial WIP at each of its operations can be netted out from available capacity in the planning periods using the exogenous lead time. Constraints (4) link the output of each operation of each product to the releases of that product assuming instantaneous material transfer between successive operations. The capacity constraints (5) ensure that the workcenter capacity cannot be exceeded. The frozen periods are modeled by (6), while (7) ensure the non-negativity of the decision variables.

The recursion $L_l^{(g)} := L_{l-1}^{(g)} + FF^{(g)}\alpha_l^{(g)},\ g\in G, l\in\Omega(g)$ is used to determine the value of $L_l^{(g)}$ with initial condition $L_0^{(g)} := 0$ and a product-specific flow factor $FF^{(g)}$ [1], which is determined from simulation runs with a given target bottleneck utilization. Since we obtain the lead time $L_l^{(g)}$ by rounding the average cycle time from the initial simulation runs down to an integer multiple of the period length, we refer to this model as the Simple Rounding Down (SRD) model.

B. Formulation With Workload-Dependent Lead Times

The non-linear relation between mean cycle time and resource utilization [1], [9] motivates the use of non-linear CFs to represent workload-dependent lead times in production planning models [23], [45], [46]. Following the literature, we use outer linearization of the CFs to obtain a LP model using the following additional notation:

Sets and Indices:

$C(k)$: index set of the linear segments used to approximate the CF for workcenter k
$K(l)$: set of workcenters where operation l can be performed

Decision Variables:

$X_{tl}^{(g)}$: quantity of product g starting operation l in period t
$W_{tl}^{(g)}$: WIP of product g at operation l at the end of period t
$Z_{ktl}^{(g)}$: fraction of output from workcenter k allocated to operation l of product g in period t

Parameters:

μ_k^n: intercept of segment n of the CF for workcenter k
β_k^n: slope of segment n of the CF for workcenter k.

The resulting planning model is obtained from the SRD model (1)-(7) by replacing constraints (3)-(5) with the following constraints that represent the behavior of the resources:

$$W_{t-1,l}^{(g)}+X_{tl}^{(g)}-Y_{tl}^{(g)} = W_{tl}^{(g)}, g\in G, t=s,\ldots,s+T-1, l\in\Omega(g) \tag{8}$$

$$\alpha_l^{(g)} Y_{tl}^{(g)} \le \mu_k^n Z_{ktl}^{(g)} + \beta_k^n \alpha_l^{(g)} \left(X_{tl}^{(g)} + W_{t-1,l}^{(g)} \right), \quad g \in G, t = s, \ldots, s+T-1, l \in \Omega(g), k \in K(l), n \in C(k) \tag{9}$$

$$\sum_{g \in G, l \in \bar{\Omega}(k)} Z_{ktl}^{(g)} = 1, t = s, \ldots, s+T-1, k \in K \tag{10}$$

$$X_{tl}^{(g)}, W_{tl}^{(g)}, Z_{ktl}^{(g)} \ge 0, g \in G, t = s, \ldots, s+T-1, k \in K, l \in \Omega(g). \tag{11}$$

The WIP balance constraints (8) are required for each operation, while (9) approximate the non-linear CFs with linear segments. The CFs (9) describe the output in a period as a function of the total amount of work available for processing in that period, which is given by the sum of the releases in that period and the WIP at the end of the previous period. Constraints (10) allocate the total output of the workcenter among the operations taking place there, following the Allocated CF (ACF) formulation [45], [46].

We now discuss the extension of these models to stochastic demand modeled using the MMFE.

C. Demand-Driven Safety Stock Setting

The demand-driven safety stock approach of Norouzi and Uzsoy [47] and Albey *et al.* [16] seeks to maintain the inventory position of product g at the start of period t at a specified target level $S_t^{(g)}$ that will ensure a specified maximum stockout probability [5]. However, the capacity constraints may not allow sufficient production to raise the inventory position to the target level in one period. The random variable

$$U_t^{(g)} := \max\left(0, S_t^{(g)} - \left(Y_t^{(g)} + I_{t-1}^{(g)} - B_{t-1}^{(g)}\right)\right), \tag{12}$$

defines the *shortfall*, the amount by which the target inventory level exceeds the actual inventory. The quantity $Y_t^{(g)} + I_{t-1}^{(g)} - B_{t-1}^{(g)}$ represents the net inventory of product g on hand at the end of period t. Following [16], we incorporate shortfall into chance constraints that seek to achieve a specified stock-out probability. Combining (2) and (12) gives

$$I_{t+1}^{(g)} - B_{t+1}^{(g)} + D_{t+1}^{(g)} + U_{t+1}^{(g)} \ge S_{t+1}^{(g)}, g \in G, t = s, \ldots, s+T-2, \tag{13}$$

where $S_t^{(g)*} := F^{-1}_{Q_t^{(g)}}(b_t^{(g)}/(h_t^{(g)} + b_t^{(g)}))$ denotes the cost-minimizing target inventory level for the uncapacitated case and $F_{Q_t^{(g)}}$is the distribution function of the random variable $Q_t^{(g)} := D_t^{(g)} + U_t^{(g)}$ [48]. The SRD and ACF models can then be extended to consider stochastic demand by adding the chance constraints (13) and the additional non-negativity constraint

$$U_{t+1}^{(g)} \ge 0, g \in G, t = s, \ldots, s+T-2. \tag{14}$$

To avoid increasing the shortfall without bound, shortfall must be penalized in the objective function (1), yielding:

$$\sum_{g \in G} \sum_{t=s}^{s+T-1} \left[\omega_t^{(g)} W_t^{(g)} + h_t^{(g)} I_t^{(g)} + b_t^{(g)} B_t^{(g)} + u_t^{(g)} U_t^{(g)} \right] \tag{15}$$

with a unit shortfall cost $u_t^{(g)}$ of product g in period t. Detailed justification of this approach is given in [16].

Additive MMFE: We first consider the situation when demand follows the additive MMFE [13], [47]. Let $D_{st}^{(g)}$, $s \le t \le s+H-1$, denote the demand forecasts for product g available at the end of epoch s where H denotes the length of the forecast window, and $D_t^{(g)} = D_{tt}^{(g)}$ the realized demand for product g in period t. The additive MMFE represents the forecast update computed at planning epoch s for period t in the current forecast horizon as the random variable:

$$\varepsilon_{st}^{(g)} := D_{st}^{(g)} - D_{s-1,t}^{(g)}, g \in G, t = s, \ldots, s+H-1, \tag{16}$$

where $\varepsilon_{st}^{(g)} \sim N(0, \sigma_{gt}^2)$. The corresponding forecast update vector is given by $\varepsilon_s := (\varepsilon_{ss}^{(1)}, \ldots, \varepsilon_{s,s+H-1}^{(1)}, \ldots, \varepsilon_{ss}^{(\Gamma)}, \ldots, \varepsilon_{s,s+H-1}^{(\Gamma)})^T$ and the realized demand in each period by

$$D_s^{(g)} = \mu_g + \sum_{j=0}^{H-1} \varepsilon_{s-H+j,s}^{(g)}, \tag{17}$$

where $\mu_g := E(D_s^{(g)})$. Güllü [49] shows that the unconditional covariance between $D_t^{(g)}$ and $D_{t+i}^{(g)}$ in epoch s is

$$\gamma_{\mathrm{i}}^{(g)(h)} := \mathrm{Cov}\left(D_t^{(g)}, D_{t+i}^{(h)}\right) = \sum_{j=0}^{H-1-i} \mathrm{Cov}\left(\varepsilon_{t,t+j}^{(g)}, \varepsilon_{t,t+i+j}^{(h)}\right). \tag{18}$$

We use $\gamma_{\mathrm{i}}^{(g)} := \gamma_{\mathrm{i}}^{(g)(g)}$ to simplify notation. When $i = 0$, (18) reduces to

$$\sigma_g^2 := \gamma_0^{(\mathrm{g})} = \mathrm{Cov}\left(D_t^{(g)}, D_t^{(g)}\right) = \sum_{k=0}^{H-1} \sigma_{gk}^2. \tag{19}$$

Norouzi and Uzsoy [47] show that the conditional covariance given the σ-field F_s describing the information available at the end of period s is

$$\gamma_{s,(t,t+i)}^{(g)} := \mathrm{Cov}\left(D_t^{(g)}, D_{t+i}^{(g)} | F_s\right) = \sum_{j=1}^{t-s} \mathrm{Cov}\left(\varepsilon_{s,t-j}^{(g)}, \varepsilon_{s,t-j+i}^{(g)}\right) \tag{20}$$

for $1+s \le t \le H+s-i$. The conditional covariance at epoch s provides advance demand information, reducing demand uncertainty over the forecast horizon. Albey *et al.* [16] use results from [50], [51] to show that

$$S_{s,t+1}^{(g)*} = \frac{1}{\theta} \ln\left(1 + b_{t+1}^{(g)}/h_{t+1}^{(g)}\right) - \beta + D_{st}^{(g)} + \frac{1}{2} \gamma_{s,(t,t+i)}^{(g)} \theta, \tag{21}$$

where $\beta := 0.583\sqrt{\gamma_0^{(g)} + 2\sum_{i=1}^{H-1} \gamma_i^{(g)}}$, $CR_t^{(g)}$ denotes the capacity allocated to product g in period t and $\theta := 2(CR_t^{(g)} - \mu_g)/(\gamma_0^{(g)} + 2\sum_{i=1}^{H-1} \gamma_i^{(g)})$. When no advance demand information is taken into account, i.e., forecast evolution is ignored, Norouzi [48] shows that (21) reduces to

$$S_{s,t+1}^{(g)*} = \frac{1}{\theta} \ln\left(1 + b_{t+1}^{(g)}/h_{t+1}^{(g)}\right) - \beta + CR_t^{(g)} \tag{22}$$

where (21) and (22) are exact for $b_t^{(g)} >> h_{\mathrm{t}}^{(\mathrm{g})}$.

Multiplicative MMFE: The multiplicative MMFE permits modeling of demand where the magnitude of the forecast

updates depends on that of the forecast [13], [47]. Assuming strictly positive demand forecasts, a logarithmic transformation yields a forecast update vector ε_s with components

$$\varepsilon_{st}^{(g)} := \ln D_{st}^{(g)} - \ln D_{s-1,t}^{(g)},\ g \in G,\ t = s, \ldots, s+H-1 \quad (23)$$

where $\varepsilon_{st}^{(g)} \sim N(-\sigma_{gt}^2/2, \sigma_{gt}^2)$ [17]. The demand is given by

$$D_s^{(g)} = \exp\left(\ln\mu_g + \sum_{j=0}^{H-1} \varepsilon_{s-H+j,s}^{(g)}\right). \quad (24)$$

In a multi-product setting the unconditional covariance is given by

$$\mathrm{Cov}\left(D_t^{(g)}, D_{t+i}^{(h)}\right) = \mu_g\mu_h\left(\exp\left(\sum_{j=0}^{H-1-i} \mathrm{Cov}\left(\varepsilon_{t,t+j}^{(g)}, \varepsilon_{t,t+i+j}^{(h)}\right)\right) - 1\right) \quad (25)$$

and the conditional covariance by

$$\begin{aligned}\gamma_{s,(t,t+i)}^{(g)} &:= \mathrm{Cov}\left(D_t^{(g)}, D_{t+i}^{(g)} | F_s\right) \\ &= D_{st}^{(g)} D_{s,t+i}^{(g)}\left(\exp\left(\sum_{j=1}^{t-s} \mathrm{Cov}\left(\varepsilon_{s,t-j}^{(g)}, \varepsilon_{s,t-j+i}^{(g)}\right)\right) - 1\right)\end{aligned} \quad (26)$$

for $1 + s \le t \le H + s - i$ [17]. When $b_t^{(g)} >> h_{\mathrm{t}}^{(\mathrm{g})}$ we obtain the cost-minimizing base stock level

$$\begin{aligned}S_{s,t+1}^{(g)*} = \frac{1}{\theta}\ln\left(1 + b_{t+1}^{(g)}/h_{t+1}^{(g)}\right) - \beta + \lambda_{st}^{(g)} D_{s,t+1}^{(g)} \\ - \theta\left(D_{s,t+1}^{(g)}\lambda_{st}^{(g)}\right)^2 \ln\lambda_{st}^{(g)},\end{aligned} \quad (27)$$

where $\lambda_{st}^{(g)} := D_{s,t+1}^{(g)} / \sqrt{(D_{s,t+1}^{(g)})^2 + \gamma_{s,(t+1,t+1)}^{(g)}}$. When forecast evolution is ignored, (27) reduces to

$$S_{s,t+1}^{(g)*} = \frac{1}{\theta}\ln\left(1 + b_{t+1}^{(g)}/h_{t+1}^{(g)}\right) - \beta + \lambda_g\mu_g - \theta\left(\mu_g\lambda_g\right)^2\ln\lambda_g \quad (28)$$

where $\lambda_g := \mu_g/\sqrt{\sigma_g^2 + \mu_g^2}$. The planning models with fixed and workload-dependent lead times and safety stock levels based on the chance constraints (21) or (27) are denoted by SRD-CC-U and ACF-CC-U. The notation U indicates the use of advance demand information. Those that neglect forecast evolution by using chance constraints (22) and (28) are denoted by SRD-CC-N and ACF-CC-N, where N indicates no use of advance demand information.

D. Cycle Time-Driven Safety Stock Setting

The approach of Hung and Chang [18] assumes that all lots will complete their processing in the same sequence as that in which they are released into the fab. Thus when lots are indexed by their release date into the fab, the last operation on the routing of lot $k-1$ of product g will always complete before that of lot k of product g. Let $\tau_k^{(g)}$ denote the release time of the k'th lot of product g to be released, and $C_k^{(g)}$ its cycle time. The probability of at least k lots being produced by the end of period t is then given by

$$P\left(O_t^{(g)} > k\right) = P\left(\tau_k^{(g)} + C_k^{(g)} \le t\right), \quad (29)$$

where $O_t^{(g)}$ denotes the number of lots of product g completed by the end of period t. Equation (29) allows us to convert cycle time uncertainty into output uncertainty. Let $\tilde{C}^{(g)} := t - \tau_k^{(g)}$, and assume that lot k of product g follows approximately the same cycle time distribution $F_{C^{(g)}(t)}$ as a lot of product g completed at the end of period t. Then the safety stock for product g at the end of period t is given by

$$SS_t^{(g)} = E\left(O_t^{(g)}\right) - k = \left(\tilde{C}^{(g)} - E\left(C^{(g)}(t)\right)\right)r^{(g)}, \quad (30)$$

where $r^{(g)}$ denotes the production rate for product g. We set $\tilde{C}^{(g)} = F^{-1}_{C_k^{(g)}}(\vartheta)$ where ϑ represents the probability that lot k of product g will be produced before the end of period t and $r^{(g)} \approx Y_t^{(g)}$ in (30). Thus the planned safety stock $E(O_t^{(g)}) - k$ is given by a safety lead time $\tilde{C}^{(g)} - E(C^{(g)}(t))$ multiplied by the planned production rate in (30). The quantity $\sum_{\tau=1}^{t} D_\tau^{(g)}$ gives the desired number of lots that should be completed by the end of period t, i.e., $k = \sum_{\tau=1}^{t} D_\tau^{(g)}$, when the demand is expressed in lots. Equation (2) yields $E(O_t^{(g)}) - k = I_g^{(t)} - B_g^{(t)}$ since $E(O_t^{(g)}) = Y_t^{(g)} + \sum_{\tau=1}^{t-1} D_\tau^{(g)} + I_{t-1}^{(g)} - B_{t-1}^{(g)}$. The capacity constraints may result in shortfall, yielding the chance constraint

$$\begin{aligned}I_t^{(g)} - B_t^{(g)} + U_t^{(g)} \ge S_t^{(g)*} = \left(F^{-1}_{C_k^{(g)}}\left(\vartheta^*\right) - L_t^{(g)}\right)Y_t^{(g)}, \\ g \in G, t = s, \ldots, s+T-1,\end{aligned} \quad (31)$$

where $L_t^{(g)}$ is the expected cycle time of a lot of product g with completion date t. Since the ratio $b_t^{(g)}/(h_t^{(g)} + b_t^{(g)})$ can be interpreted as a target service level [48], we set $\vartheta^* = b_t^{(g)}/(h_t^{(g)} + b_t^{(g)})$. The SRD and ACF models with the modified objective function (15) are extended by constraints (31) and

$$U_t^{(g)} \ge 0,\ g \in G, t = s, \ldots, s+T-1. \quad (32)$$

The resulting planning models with fixed and workload-dependent lead times and cycle time-driven safety stocks are referred to as the SRD-SS and the ACF-SS models, respectively. Note that the cycle time-driven safety stock setting approach does not consider demand uncertainty, focusing on output uncertainty expressed by the cycle time distribution.

Overall, we consider eight different planning models that differ in how lead times and safety stocks are incorporated as summarized in Table I. Recall that the notation U indicates the use of advance demand information whereas N indicates that such information is not used.

IV. Rolling Horizon Framework

A. Simulation Environment

A simulation infrastructure from previous work [17], [44] is used to assess the performance of the planning models in a rolling horizon setting. A simulation model of a wafer fab represents the execution level, while an in-memory data layer

TABLE I
CHARACTERISTICS OF THE PLANNING MODELS

Formulation	Lead time modeling		Safety stock modeling		
	fixed	CF	cycle time-driven	demand-driven	
				N	U
SRD	x	-	-	-	-
SRD-SS	x	-	x	-	-
SRD-CC-N	x	-	-	x	-
SRD-CC-U	x	-	-	-	x
ACF	-	x	-	-	-
ACF-SS	-	x	x	-	-
ACF-CC-N	-	x	-	x	-
ACF-CC-U	-	x	-	-	x

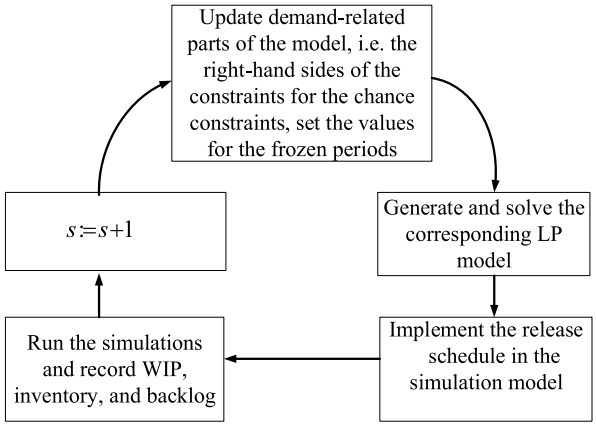

Fig. 1. Rolling horizon planning framework.

contains business objects such as lots and machines whose state is updated based on events at the execution level. The simulation engine stops at each planning epoch to collect WIP, FGI, and backlog information from the data layer.

This information is then combined with demand forecast updates for the different products to generate a new instance of the planning model based on the most current information. The production quantities for the present planning epoch are obtained by solving this instance of the planning model. The appropriate number of lots are then released into the fab at a uniform rate over the duration of the planning period. The end of the planning period represents the next planning epoch, when a new instance of the planning problem is constructed and solved. This procedure is repeated until the end of the simulation horizon is reached, as shown in Figure 1.

The MIMAC I testbed [52], [53] represents a large-scale wafer fab with 84 workcenters and over 200 machines, some with batch processing and sequence-dependent setup times. First-In-First-Out dispatching is used, machine breakdowns are exponentially distributed, and processing times are deterministic. Two products with reentrant process flows and instantaneous material transfer between successive operations are represented, each with more than 200 process steps. The number of steppers in the original MIMAC I model is reduced to ensure that the lithography area is the planned bottleneck when both products are produced in equal proportion. Each lot contains 48 wafers.

B. Calculation of Model Parameters

Empirical data are collected from the simulation model under each breakdown scenario for different bottleneck utilization (BNU) levels to fit the non-linear CFs. We observe the quantities $X_{kt} := \sum_{g\in G, l\in O(k)} \alpha_l^{(g)} X_{tl}^{(g)}$, $Y_{kt} := \sum_{g\in G, l\in O(k)} \alpha_l^{(g)} Y_{tl}^{(g)}$, and $W_{k,t-1} := \sum_{g\in G, l\in O(k)} \alpha_l^{(g)} W_{t-1,l}^{(g)}$ for each workcenter k and period t. Following [20], we partition the resource load axis into two intervals with an equal number of data points. We then fit separate linear functions to the data in each interval using linear regression. A third segment with a slope of zero and intercept equal to the maximum capacity of the workcenter is then added.

The value of θ used in the chance constraints (21), (22), (27) and (28) requires specifying a capacity allocation $CR_t^{(g)}$ among the different products g. To this end we first determine an upper bound on the total throughput of the fab in units of lots. This quantity is then allocated to the products in each period in the ratio of their expected workload on the bottleneck workcenter.

To determine the cycle time distribution $F_{C_k^{(g)}}$ of product g for the SRD-SS and ACF-SS models, simulation runs are performed for each of the demand and machine failure scenarios discussed in Section V-A. The χ^2 test confirms that the cycle times follow a lognormal distribution with parameters μ and σ, yielding $F^{-1}_{C_k^{(g)}}(\vartheta^*) = \exp(\sigma\Phi^{-1}(F^{-1}_{C_k^{(g)}}(\vartheta^*)) + \mu)$ which we use in the right-hand side of constraints (31); here Φ denotes the CDF of the standard normal distribution.

C. Demand Generation Scheme

Modeling demand using the additive and multiplicative MMFE requires specifying the mean demand μ_g for each product and the variance covariance (VCV) matrix

$$\Sigma = E\left(\varepsilon_s \varepsilon_s^T\right) - E(\varepsilon_s)E(\varepsilon)^T \in IR^{\Gamma(H+1)\times\Gamma(H+1)} \qquad (33)$$

of the update vectors. Two update scenarios are considered, representing late and early uncertainty resolution. The former is characterized by $\sigma_{g0} \le \sigma_{g1} \le \cdots \le \sigma_{gH}$, and the latter by $\sigma_{g0} \ge \cdots \ge \sigma_{g,H-1} \ge \sigma_{gH}$. Cholesky decomposition of the VCV matrix is used to generate realizations of the update vectors following Scheuer and Stoller [54]. If the matrix Σ obtained in this manner is not positive definite, the nearest positive definite VCV matrix is determined by solving an approximation problem [17].

V. SIMULATION EXPERIMENTS

A. Design of Experiments

We expect that the planned BNU level will significantly impact the performance of the planning models. Therefore, long simulation runs are performed prior to the main experiments to determine mean demand values μ_g yielding the desired target BNU levels of 70% and 90%. Several initial WIP distributions are taken at random from these runs. One of

TABLE II
SPECIFICATION OF THE σ_{gt} VALUES

MMFE		**Additive (relative to μ_g)**			
CV		**0.10**		**0.25**	
Product		**1**	**2**	**1**	**2**
Period	1	0.0080	0.0160	0.0199	0.0400
	2	0.0088	0.0200	0.0220	0.0500
	3	0.0160	0.0240	0.0400	0.0600
	4	0.0249	0.0282	0.0600	0.0705
	5	0.0329	0.0292	0.0800	0.0730
	6	0.0400	0.0440	0.1000	0.1100
	7	0.0800	0.0720	0.2000	0.1800
		Multiplicative			
Period	1	0.0079	0.0160	0.0196	0.0394
	2	0.0088	0.0200	0.0217	0.0492
	3	0.0160	0.0239	0.0394	0.0591
	4	0.0239	0.0281	0.0591	0.0694
	5	0.0319	0.0291	0.0788	0.0719
	6	0.0399	0.0439	0.0985	0.1083
	7	0.0798	0.0718	0.1970	0.1773

these is then randomly selected to initialize the WIP values for the planning instance at the start of the first planning epoch.

A period length of one week and a fixed planning window length of $T = 7$ weeks are used because previous work [17] finds that the value of T affects profit, service level, and stability only slightly in a rolling horizon setting if T is sufficiently large. Cycle times of up to three periods motivate the extension of the planning window by three periods following [7] to avoid end-of-horizon effects. Demand in each of the additional periods is set to the average demand over the last three periods of the planning window and equal release rates are enforced for each of the additional periods.

Demand following the selected MMFE is considered for a forecast window of $H = T$ weeks. Demand uncertainty is varied by considering coefficient of variation values of $CV_g = 0.10$ and 0.25 for the demand of each product g. Using (19) and (25) the sum of the variance of the update quantities $\sum_{t=1}^{H} \sigma_{gt}^2 = \mu_g^2 CV_g^2$ and $\sum_{t=1}^{H} \sigma_{gt}^2 = \ln(CV_g^2+1)$ are computed for the additive and multiplicative MMFE, respectively. σ_{gt} values for early uncertainty resolution are given in Table II. Instances with late resolution are obtained by reversing the order of the diagonal and off-diagonal elements of the VCV matrix.

The correlation matrix $(\rho_{ij}^{(g)(h)}) \in IR^{\Gamma(H+1)\times\Gamma(H+1)}$ is defined as follows for the VCV matrix under positive correlation:

$$\rho_{ij}^{(g)(h)} := \begin{cases} 1.0, & \text{if } g = h, i = j \\ 0.5, & \text{otherwise,} \end{cases} \tag{34}$$

while for negative correlation we use

$$\rho_{ij}^{(g)(h)} := \begin{cases} 1.0, & \text{if } g = h, i = j \\ -0.5, & \text{otherwise.} \end{cases} \tag{35}$$

We also expect that the performance of the planning models will be affected by the variability of the production process, which is governed by the machine failures. The Mean Time to Failure (MTTF) and Mean Time to Repair (MTTR) values for the workcenters in the MIMAC I model are used for our short failure scenarios, assuming exponential distributions. Long failure scenarios are obtained by doubling the MTTR and MTTF values, yielding the same average availability.

TABLE III
DESIGN OF EXPERIMENTS

Factor	**Level**	**Count**
Planning model	SRD, SRD-SS, SRD-CC-N, SRD-CC-U, ACF, ACF-SS, ACF-CC-N, ACF-CC-U	8
Number of frozen periods	$\Delta \in \{0,1\}$	2
Demand generation scheme	Additive (ADD), Multiplicative (MULT)	2
BNU level	Low (70%), High (90%)	2
CV	0.10, 0.25	2
Correlation	Positive, Negative	2
Uncertainty resolution	Early, Late	2
Machine failures	Short (S), Long (L)	2
Independent demand realizations		3
Independent simulation replications		5
Total simulation runs		15,360

A simulation horizon s^* of one year with plan revision after each period is used. The number of frozen periods is $\Delta \in \{0, 1\}$. Five independent simulation runs are performed for each of the three independent demand instances characterized by their BNU and CV levels. The experimental design is summarized in Table III.

We consider the expected profit over the simulation horizon given by the mean difference between the realized revenue and the realized WIP, FGI, and backlog costs over all independent replications, as a performance measure.

We also examine two measures of service level. The weighted average of the individual product service levels based on the product mix is used. To consider nervousness, the plan stability is measured as:

$$\Psi := \frac{1}{T(s^*-1)\Gamma} \sum_{g\in G} \sum_{s=2}^{s^*} \sum_{t=s}^{s+T-1} 2^{s-t-1} \left| X_{st}^{(g)} - X_{s-1,t}^{(g)} \right|, \tag{36}$$

where $X_{st}^{(g)}$ is the planned release quantity of product g in period t computed in planning epoch s. The weights $1/2^{t-s+1}$ ensure that deviations in later periods are weighted less than those in earlier periods. The unit revenue is 450, and the unit backlogging, FGI, and WIP costs per period are 90, 10, and 60, respectively. A unit shortfall cost of half of the unit backlogging cost [17] is used. The cost ratio $b_t^{(g)}/(b_t^{(g)} + h_t^{(g)}) = 0.9$ implies a target service level of 90% in the absence of capacity constraints.

TABLE IV
AVERAGE PROFIT FOR PLANNING MODELS OVER ALL EXPERIMENTAL CONDITIONS

Frozen	BNU	70%		90%	
	CV	0.10	0.25	0.10	0.25
$\Delta = 0$	ACF	1.071	1.058	1.046	1.052
	ACF-SS	1.115	1.079	1.083	1.074
	ACF-CC-N	1.162	1.128	1.114	1.101
	ACF-CC-U	1.162	1.129	1.112	1.116
	SRD	1.000	1.000	1.000	1.000
	SRD-SS	1.045	1.034	1.032	1.002
	SRD-CC-N	1.160	1.126	1.092	1.066
	SRD-CC-U	1.162	1.133	1.097	1.091
$\Delta = 1$	ACF	0.994	0.982	0.950	0.951
	ACF-SS	1.040	1.003	0.988	0.974
	ACF-CC-N	1.089	1.056	1.018	0.998
	ACF-CC-U	1.089	1.058	1.020	1.018
	SRD	0.919	0.921	0.902	0.898
	SRD-SS	0.968	0.958	0.936	0.899
	SRD-CC-N	1.087	1.054	0.997	0.965
	SRD-CC-U	1.090	1.063	1.001	0.992

The simulation infrastructure is implemented in the C++ programming language, while the LP models are solved using ILOG CPLEX libraries. The AutoSched AP simulation software is used for the simulation experiments. A computer with 3.6 GHz Intel Core i7-4790 CPU with 16GB RAM requires up to 1.4 and 54 seconds to solve one instance of the SRD and ACF models, respectively. Additional safety stock constraints increase the computation time by an amount ranging from from 0.1 to 11.5 seconds depending on the planning model. The number of decision variables and constraints encountered in the experiments for the SRD-type models are $O(\Gamma\Omega T)$ and $O(T(|K|+\Gamma\Omega))$, respectively where $|K|$ is the number of work-centers and Ω the maximum number of operations over G. The ACF-type models have $O(\Gamma\Omega T|K|)$ decision variables and $O(\Gamma\Omega T|K|\kappa)$ constraints where κ is the number of segments of the linearized CFs. Thus the computational burden of the planning formulations does not constitute a significant obstacle to practical use of these models.

B. Simulation Results

Due to space limitations and the very large amount of data generated by the experiments, we shall structure our discussion of the results in this paper in a heuristic manner, based on what in statistical terms would be called mean treatment effects. A rigorous statistical analysis supporting these conclusions is presented in the supplementary material, with much additional material given in [55].

Overall Performance of Planning Models: Table IV examines the relative performance of the planning models in terms of expected profit. All entries represent the average profit over all simulation runs with the specified levels of utilization and CV, relative to the SRD model with no freezing.

It is immediately apparent from Table IV that without the chance constraints the SRD model is consistently outperformed by the ACF approach. The advantage of the ACF model increases with utilization and demand CV, being highest at BNU = 90% and CV = 0.25. This advantage of ACF is maintained both when freezing is present, and when it is not. The presence of freezing results in lower profit for all planning models, as one would expect based on the literature. Incorporation of the SS approach for safety stock planning gives a modest improvement for both ACF and SRD models, which diminishes at high utilization and CV. This is not surprising, since the SS approach is focused on production uncertainty without considering demand uncertainty at all. Hence, it is unlikely to perform well in an environment with substantial demand uncertainty such as that examined here.

Including the chance constraints in the ACF-CC and SRD-CC models yields significant improvement over the corresponding deterministic models (ACF and SRD). At BNU = 70%, the addition of the chance constraints essentially equalizes the performance of the ACF and SRD models under both freezing and non-freezing conditions. At BNU = 90%, the ACF-CC models have a small advantage over the SRD-CC models. At BNU = 70%, there appears to be no advantage to the forecast updating under no freezing for the ACF models, although a slight improvement is seen between SRD-CC-N and SRD-CC-U. At BNU = 90%, there is no improvement between ACF-CC-N and ACF-CC-U, and a slight improvement from SRD-CC-N to SRD-CC-U. The same patterns hold under freezing. Thus, when averaged over all demand models, the chance constraints yield significant benefit for both SRD and ACF models, with ACF consistently, but slightly, outperforming SRD at high BNU and CV. Taken as a whole, these results suggest that while including the chance constraints adds considerable value to the planning models, the advance demand information obtained from explicit representation of forecast evolution is of less value. At low BNU, this is to be expected, as there is sufficient excess capacity to allow unpredicted demand variation from one period to the next to be accommodated easily.

We now examine the results for BNU = 90% more closely, at the level of specific demand models, to assess whether forecast evolution is beneficial in any of these cases. These results are summarized in Tables V and VI for the non-frozen and frozen cases, respectively. Since our interest is in whether explicit use of the advance demand information provided by forecast evolution is beneficial, we compare the CC-N and CC-U versions of the SRD and ACF models. Table V presents a fairly consistent picture. Under the additive MMFE and low CV, forecast updating yields no significant benefit for the ACF models, but a slight improvement for the SRD models. Under the additive MMFE and high CV, the CC-U models perform better for both ACF and SRD. As would be expected, late uncertainty resolution leads to lower profits for all models without greatly affecting their relative performance. Interestingly, under the multiplicative MMFE there is no advantage to forecast evolution when CV = 0.1; the performance of the CC-U and CC-N models are essentially identical. However, at CV = 0.25, forecast updating results in additional benefit for both SRD and ACF, although the improvement is considerably smaller under the

TABLE V
AVERAGE PROFIT FOR PLANNING MODELS WITH BNU = 90%, NON-FROZEN

			ACF-CC		SRD-CC	
			N	U	N	U
			CV = 0.10			
ADD	neg.	early	1.025	1.024	1.000	1.014
		late	1.016	1.013	1.000	1.006
	pos.	early	1.030	1.028	1.000	1.013
		late	1.015	1.020	1.000	1.001
MULT	neg.	early	1.020	1.015	1.000	1.003
		late	1.019	1.020	1.000	0.998
	pos.	early	1.021	1.018	1.000	1.007
		late	1.018	1.014	1.000	1.000
			CV = 0.25			
ADD	neg.	early	1.058	1.084	1.000	1.055
		late	1.020	1.041	1.000	1.049
	pos.	early	1.056	1.071	1.000	1.040
		late	0.985	1.021	1.000	1.027
MULT	neg.	early	1.042	1.048	1.000	1.018
		late	1.044	1.050	1.000	1.005
	pos.	early	1.024	1.024	1.000	0.994
		late	1.032	1.035	1.000	1.004

TABLE VI
AVERAGE PROFIT FOR PLANNING MODELS WITH BNU = 90%, FROZEN

			ACF-CC		SRD-CC	
			N	U	N	U
			CV = 0.10			
ADD	neg.	early	1.026	1.030	1.000	1.014
		late	1.015	1.017	1.000	1.006
	pos.	early	1.032	1.030	1.000	1.012
		late	1.014	1.023	1.000	0.999
MULT	neg.	early	1.019	1.018	1.000	1.002
		late	1.020	1.028	1.000	0.995
	pos.	early	1.020	1.021	1.000	1.006
		late	1.018	1.019	1.000	0.998
			CV = 0.25			
ADD	neg.	early	1.066	1.102	1.000	1.067
		late	1.019	1.049	1.000	1.058
	pos.	early	1.061	1.080	1.000	1.046
		late	0.980	1.023	1.000	1.028
MULT	neg.	early	1.043	1.054	1.000	1.019
		late	1.052	1.068	1.000	1.008
	pos.	early	1.022	1.026	1.000	0.994
		late	1.033	1.042	1.000	1.005

multiplicative MMFE. Under freezing, the results in Table VI exhibit essentially the same pattern.

Taken as a whole, the results suggest that the SS procedure for safety stock placement enhances profit performance only slightly since it ignores demand uncertainty. It is worth pointing out that based on previous experiments [20], the variability in the production process exhibited in the MIMAC-I data remains somewhat low compared to that encountered in industrial facilities.

The shortfall-based chance constraints in the CC models increase expected profit relative to the deterministic SRD model by up to 16% at low BNU and up to 11% under high BNU in the absence of freezing. Freezing the first period in the planning window reduces this improvement to about 9%. Explicit modeling of forecast evolution in the chance constraints yields very slight (although statistically significant) benefits at high utilization and high CV, and essentially none at lower utilization.

The limited benefit of explicitly capturing forecast updating in the chance constraints is initially somewhat surprising. While one might expect this effect to be secondary in magnitude relative to that of adding the chance constraints to the deterministic formulations, previous work on single-stage systems under the additive MMFE model by Albey *et al.* [16] found considerably more benefit from including forecast evolution. A number of possible causes suggest themselves. The first is the choice of planning period length; the average cycle time of the fab studied here is above two planning periods, while the forecast updating window is $H = 7$ periods, and the planning window of $T = 7$ periods comprises two full cycle times. Hence there are on average only two opportunities for revising demand forecasts before a lot is completed, limiting the scope for adjustment of releases and reallocation of machine capacity to reprioritize WIP. The reduced profit observed with freezing suggests that some replanning of WIP is indeed taking place.

Another reason why advance demand information is of limited benefit may be the choice of utilization levels in our experiments. At the relatively low utilization level of BNU = 70%, there is sufficient excess capacity that any updates in forecasts can be accommodated easily. At BNU = 90%, on the other hand, there is far less excess capacity available, limiting the models' ability to take advantage of the advance demand information provided by forecast evolution; the resources are almost fully utilized in any case. We return to this question in the discussion of the service levels below.

Service Levels Across All Experimental Conditions: We now examine the performance of the different planning models in terms of two service level measures. The α-service level denotes the average probability of a stockout in any period, while the β-service level denotes the fill rate, the average fraction of demand met from inventory in each period. These results are summarized in Table VII and Table VIII for non-frozen and frozen cases, respectively.

None of the models succeeds in achieving the uncapacitated α service level of 0.90 implied by the ratio of holding and backorder costs. The deterministic versions of both models perform very poorly, as is to be expected, while the best the chance constrained models can do is 0.81 under low BNU and CV and no freezing. However, the fill rates are consistently high under no freezing even with high BNU and CV, implying that although stockouts are frequent they are small in magnitude relative to the demand. This reflects satisfactory performance; a low fill rate would reflect a great deal of unmet demand, while a very high α service level would require excessive inventory. The ACF models are again consistently superior to the SRD models under no freezing.

TABLE VII
AVERAGE SERVICE LEVELS OVER ALL EXPERIMENTAL CONDITIONS, NON-FROZEN

	α-SL	β-SL	α-SL	β-SL
CV	**10**		**25**	
BNU	**70%**			
ACF	0.466	0.966	0.570	0.973
ACF-SS	0.573	0.971	0.642	0.977
ACF-CC-N	0.806	0.978	0.849	0.988
ACF-CC-U	0.813	0.979	0.861	0.989
SRD	0.186	0.939	0.277	0.955
SRD-SS	0.254	0.946	0.341	0.960
SRD-CC-N	0.658	0.971	0.695	0.981
SRD-CC-U	0.670	0.972	0.733	0.983
Overall	0.553	0.965	0.621	0.976
BNU	**90%**			
ACF	0.499	0.970	0.546	0.972
ACF-SS	0.633	0.975	0.645	0.977
ACF-CC-N	0.738	0.980	0.720	0.981
ACF-CC-U	0.736	0.980	0.748	0.983
SRD	0.211	0.960	0.310	0.962
SRD-SS	0.259	0.964	0.263	0.960
SRD-CC-N	0.423	0.974	0.474	0.973
SRD-CC-U	0.442	0.975	0.520	0.977
Overall	0.493	0.972	0.528	0.973

TABLE VIII
AVERAGE SERVICE LEVELS OVER ALL EXPERIMENTAL CONDITIONS, FROZEN

	α-SL	β-SL	α-SL	β-SL
CV	**10**		**25**	
BNU	**70%**			
ACF	0.330	0.692	0.399	0.688
ACF-SS	0.409	0.702	0.444	0.707
ACF-CC-N	0.572	0.691	0.593	0.699
ACF-CC-U	0.577	0.687	0.599	0.687
SRD	0.130	0.682	0.191	0.684
SRD-SS	0.173	0.671	0.244	0.676
SRD-CC-N	0.479	0.670	0.495	0.679
SRD-CC-U	0.464	0.681	0.510	0.690
Overall	0.392	0.684	0.435	0.689
BNU	**90%**			
ACF	0.356	0.698	0.388	0.667
ACF-SS	0.423	0.681	0.445	0.686
ACF-CC-N	0.511	0.694	0.506	0.699
ACF-CC-U	0.511	0.680	0.536	0.686
SRD	0.149	0.674	0.217	0.674
SRD-SS	0.184	0.660	0.184	0.686
SRD-CC-N	0.296	0.672	0.331	0.695
SRD-CC-U	0.313	0.682	0.366	0.693
Overall	0.343	0.680	0.372	0.686

The marked degradation in both service level measures under freezing is somewhat surprising. Since we only freeze the first of the seven periods in the planning window, this suggests that there is significant adaptation of the previous production plan from one epoch to the next in the no freezing case. This is again related to the choice of planning period length. Material released in the current period will exit the fab in either the next period or the subsequent one; if insufficient material is released it is unlikely that the deficient material can be made up under freezing. Overall, however, these results suggest that the chance constrained models provide satisfactory service levels under the experimental conditions we have explored, in addition to their strong profit performance.

TABLE IX
STABILITY OF DIFFERENT PLANNING MODELS

Frozen	**BNU**	**70%**		**90%**	
	CV	**10**	**25**	**10**	**25**
$\Delta = 0$	**ACF**	3.50	3.74	3.26	3.59
	ACF-CC-N	3.11	3.34	2.64	3.37
	ACF-CC-U	3.11	3.30	2.67	3.35
	ACF-SS	3.39	3.60	2.76	3.39
	SRD	1.52	2.35	3.57	5.40
	SRD-CC-N	1.74	2.33	1.84	3.59
	SRD-CC-U	1.74	2.26	1.85	3.76
	SRD-SS	2.20	2.56	1.97	2.56
	Overall	2.54	2.93	2.57	3.63
$\Delta = 1$	**ACF**	2.80	3.04	2.56	2.91
	ACF-CC-N	2.42	2.63	1.96	2.67
	ACF-CC-U	2.42	2.60	1.99	2.66
	ACF-SS	2.71	2.93	2.07	2.70
	SRD	0.82	1.65	2.89	4.69
	SRD-CC-N	1.05	1.64	1.16	2.88
	SRD-CC-U	1.07	1.57	1.16	3.08
	SRD-SS	1.51	1.87	1.27	1.87
	Overall	1.85	2.24	1.88	2.93

Stability of Planning Models: Table IX summarizes the stability performance of the planning models using the metric (36). At low utilization, the SRD models are much more stable than the ACF models. This is because SRD does not consider congestion, and thus can plan releases to chase demand as long as the aggregate capacity of the bottleneck resources is not exceeded. The concave clearing functions used in the ACF models require those models to release large amounts of material to increase output by even a small amount, resulting in release patterns that amplify the variations in demand. The addition of the chance constraints gives some improvement in stability, presumably due to the additional material in the line that reduces the need for additional releases.

Statistical Analysis: A comprehensive statistical analysis of the results has been carried out and is reported in [55]. The electronic supplement to this paper reports these results for the case with no freezing, using the Friedman test suggested by Conover [56]. The results indicate that many of the differences observed in the tables above are in fact statistically significant, even in some cases where the absolute magnitude is of little practical significance.

VI. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This paper has assessed several production planning models that differ in how cycle times and safety stock are modeled in a rolling horizon setting using a simulation model of a large-scale wafer fab. Demand is modeled using the MMFE, implying correlated forecast updates across both products and time are assumed for both time-stationary and non-stationary

demand. The simulation results demonstrate that it is helpful to consider safety stock in production planning models. Advance demand information can lead to improved system performance under certain experimental conditions, especially for chance-constrained models under high utilization and variance.

Models representing load-dependent lead time behavior using CFs outperform those based on exogenous, fixed lead times that are an integer multiple of the period length. Frozen periods in the planning models improved planning stability but reduced the expected profit.

There are several directions for future research. First of all, considering more products is an interesting future research topic since a large number of different products exist in most real-world wafer fabs. This would eventually require the use of multi-dimensional CFs [57]. Comparison with planning models using fractional exogenous lead time estimates [7], [58] is also of interest. Studying the interaction of production planning and shop-floor scheduling is another avenue of future research; a surprisingly small body of literature exists in this area. The interaction requires setting due dates of the individual lots that are released based on the instructions from production planning. Moreover, it seems extremely interesting to model order acceptance decisions in production planning formulations for wafer fabs. Only initial work is available for this research direction [59], [60]. The manner in which safety stocks are incorporated is also open for additional research. The approaches suggested use a number of approximations, whose accuracy we have examined in several cases, but for which alternative approaches may exist. The fact that the chance constrained approaches focus on demand uncertainty, while the cycle time-driven safety stock approach considers only production uncertainty raises the question of whether they can be integrated in a mutually reinforcing manner. Finally, although the chance constrained models presented here result in substantial improvements over deterministic models, it would be of great interest to develop benchmarks of the best possible performance that could be hoped for under different experimental conditions.

REFERENCES

[1] L. Mönch, J. W. Fowler, and S. J. Mason, "Production planning and control for semiconductor wafer fabrication facilities: Modeling, analysis, and systems," in *Operations Research/Computer Science Interfaces*, vol. 52. New York, NY, USA: Springer, 2013.

[2] C.-F. Chien *et al.*, "Modeling and analysis of semiconductor manufacturing in a shrinking world: Challenges and successes," *Eur. J. Ind. Eng.*, vol. 5, no. 3, pp. 254–271, 2011.

[3] L. Mönch, R. Uzsoy, and J. W. Fowler, "A survey of semiconductor supply chain models Part III: Master planning, production planning, and demand fulfillment," *Int. J. Prod. Res.*, vol. 56, no. 13, pp. 4565–4584, 2018.

[4] R. Uzsoy, J. W. Fowler, and L. Mönch, "A survey of semiconductor supply chain models Part II: Demand planning, inventory management, and capacity planning," *Int. J. Prod. Res.*, vol. 56, no. 13, pp. 4546–4564, 2018.

[5] S. Axsäter, *Inventory Control*. New York, NY, USA: Springer, 2010.

[6] P. H. Zipkin, *Foundations of Inventory Management*. Singapore: McGraw-Hill, 2000.

[7] R. Leachman, "Semiconductor production planning," in *Handbook of Applied Optimization*, P. Pardalos and M. Resende, Eds. New York, NY, USA: Oxford Univ. Press, 2001, pp. 746–762.

[8] S. Voß and D. Woodruff, *Introduction to Computational Optimization Models for Production Planning in a Supply Chain*, 2nd ed. New York, NY, USA: Springer, 2006.

[9] W. Hopp and M. L. Spearman, *Factory Physics*, 3rd ed. Long Grove, IL, USA: Waveland Press, 2011.

[10] G. L. Curry and R. M. Feldman, *Manufacturing Systems Modelling and Analysis*. Berlin, Germany: Springer, 2009.

[11] L. F. Atherton and R. W. Atherton, *Wafer Fabrication: Factory Performance and Analysis*. Boston, MA, USA: Kluwer, 1995.

[12] K. Wu, "An examination of variability and its basic properties for a factory," *IEEE Trans. Semicond. Manuf.*, vol. 18, no. 1, pp. 214–221, Feb. 2005.

[13] D. Heath and P. Jackson, "Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems," *IIE Trans.*, vol. 26, no. 3, pp. 17–30, 1994.

[14] L. Chen and H. Lee, "Information sharing and order variability control under a generalized demand model," *Manag. Sci.*, vol. 55, no. 5, pp. 781–797, 2009.

[15] A. S. Grove, *High-Output Management*. New York, NY, USA: Random House, 1983.

[16] E. Albey, A. Norouzi, K. Kempf, and R. Uzsoy, "Demand modeling with forecast evolution: An application to production planning," *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 3, pp. 374–384, Aug. 2015.

[17] T. Ziarnetzky, L. Mönch, and R. Uzsoy, "Rolling horizon, multi-product production planning with chance constraints and forecast evolution for wafer fabs," *Int. J. Prod. Res.*, vol. 56, no. 18, pp. 6112–6134, 2018.

[18] Y.-F. Hung and C.-B. Chang, "Determining safety stock for production planning in uncertain manufacturing," *Int. J. Prod. Econ.*, vol. 58, no. 2, pp. 199–208, 1999.

[19] N. B. Kacar, D. F. Irdem, and R. Uzsoy, "An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms," *IEEE Trans. Semicond. Manuf.*, vol. 25, no. 1, pp. 104–117, Feb. 2012.

[20] N. B. Kacar, L. Mönch, and R. Uzsoy, "Planning wafer starts using nonlinear clearing functions: A large-scale experiment," *IEEE Trans. Semicond. Manuf.*, vol. 26, no. 4, pp. 602–612, Nov. 2013.

[21] T. Ponsignon and L. Mönch, "Simulation-based performance assessment of master planning approaches in semiconductor manufacturing," *OMEGA*, vol. 46, pp. 21–35, Jul. 2014.

[22] F. Sahin, A. Narayanan, and E. P. Robinson, "Rolling horizon planning in supply chains: Review, implications and directions for future research," *Int. J. Prod. Res.*, vol. 51, no. 18, pp. 5413–5436, 2013.

[23] H. Missbauer and R. Uzsoy, "Optimization models of production planning problems," in *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook*, vol. 1, K. Kempf, P. Keskinocak, and R. Uzsoy, Eds. Berlin, Germany: Springer, 2011, pp. 437–508.

[24] S. Orçun, R. Uzsoy, and K. G. Kempf, "An integrated production planning model with load-dependent lead-times and safety stocks," *Comput. Chem. Eng.*, vol. 33, no. 12, pp. 2159–2163, 2009.

[25] A. Eruguz, E. Sahin, Z. Jemai, and Y. Dallery, "A comprehensive survey of guaranteed service models for multi-echelon inventory optimization," *Int. J. Prod. Econ.*, vol. 172, pp. 110–125, Feb. 2016.

[26] F. Tian, S. P. Willems, and K. Kempf, "An iterative approach to item-level tactical production andinventory planning," *Int. J. Prod. Econ.*, vol. 133, no. 1, pp. 439–450, 2011.

[27] J. Higle and K. G. Kempf, "Production planning under supply and demand uncertainty: A stochastic programming approach," in *Stochastic Programming: The State of the Art*, G. Infanger, Ed. Berlin, Germany: Springer, 2010, pp. 297–315.

[28] T. Aouam and R. Uzsoy, "Chance-constraint-based heuristics for production planning in the face of stochastic demand and workload-dependent lead times," in *Decision Policies for Production Networks*, K. Kempf and D. Armbruster, Eds. Boston, MA, USA: Springer, 2012, pp. 173–208.

[29] T. Aouam and R. Uzsoy, "Zero-order production planning models with stochastic demand and workload-dependent lead times," *Int. J. Prod. Res.*, vol. 53, no. 6, pp. 1661–1679, 2015.

[30] A. Ravindran, K. Kempf, and R. Uzsoy, "Production planning with load-dependent lead times and safety stocks for a single product," *Int. J. Plan. Scheduling*, vol. 1, nos. 1–2, pp. 58–86, 2011.

[31] D. Kayton, T. Teyner, C. Schwartz, and R. Uzsoy, "Focusing maintenance improvement efforts in a wafer fabrication facility operating under the theory of constraints," *Prod. Invent. Manag.*, vol. 38, no. 4, pp. 51–57, 1997.

[32] E. Albey, R. Uzsoy, and K. Kempf, "A chance constraint based multi-item production planning model using simulation optimization," in *Proc. Win. Simulat. Conf.*, 2016, pp. 2719–2729.

[33] J. D. Blackburn, D. H. Kropp, and R. Millen, "MRP system nervousness: Causes and cures," *Eng. Costs Prod. Econ.*, vol. 9, nos. 1–3, pp. 141–146, 1985.
[34] V. Sridharan and R. LaForge, "The impact of safety stock on schedule instability, cost and service," *J. Oper. Manag.*, vol. 8, no. 4, pp. 327–347, 1989.
[35] C. Ho and P. Carter, "An investigation of alternative dampening procedures to cope with MRP system nervousness," *Int. J. Prod. Res.*, vol. 34, no. 1, pp. 137–156, 1996.
[36] A. Kimms, "Stability measures for rolling schedules with applications to capacity expansion planning, master production scheduling and lot sizing," *OMEGA*, vol. 26, no. 3, pp. 355–366, 1998.
[37] N.-P. Lin and L. Krajewski, "A model for master production scheduling in uncertain environments," *Decis. Sci.*, vol. 23, no. 4, pp. 839–861, 1992.
[38] R. Venkataraman and J. Nathan, "Effect of forecast errors on rolling horizon master production schedule cost performance for various replanning intervals," *Prod. Plan. Control*, vol. 10, no. 7, pp. 682–689, 1999.
[39] X. Zhao, J. Xie, and Q. Jiang, "Lot-sizing rule and freezing the master production schedule under capacity constraint and deterministic demand," *Prod. Oper. Management*, vol. 10, no. 1, pp. 45–67, 2001.
[40] O. Tang and R. W. Grubbström, "Planning and replanning the master production schedule under demand uncertainty," *Int. J. Prod. Econ.*, vol. 78, no. 3, pp. 323–334, 2002.
[41] E. P. Robinson, F. Sahin, and L.-L. Gao, "Master production schedule time interval strategies in make-to-order supply chains," *Int. J. Prod. Res.*, vol. 46, no. 7, pp. 1933–1954, 2007.
[42] P.-C. Lin and R. Uzsoy, "Chance-constrained formulations in rolling horizon production planning: An experimental study," *Int. J. Prod. Res.*, vol. 54, no. 13, pp. 3927–3942, 2016.
[43] P. Lin and R. Uzsoy, "Estimating the costs of planned changes implied by freezing production plans," in *Heuristics, Metaheuristics and Approximate Methods in Planning and Scheduling*, G. Rabadi, Ed. Cham, Switzerland: Springer, 2016, pp. 17–44.
[44] T. Ziarnetzky, N. Kacar, L. Mönch, and R. Uzsoy, "Simulation-based performance assessment of production planning formulations for semiconductor wafer fabrication," in *Proc. Win. Simulat. Conf.*, 2015, pp. 2884–2895.
[45] J. Asmundsson, R. L. Rardin, and R. Uzsoy, "Tractable nonlinear production planning models for semiconductor wafer fabrication facilities," *IEEE Trans. Semicond. Manuf.*, vol. 19, no. 1, pp. 95–111, Feb. 2006.
[46] J. Asmundsson, R. L. Rardin, C. H. Turkseven, and R. Uzsoy, "Production planning with resources subject to congestion," *Naval Res. Logist.*, vol. 56, no. 2, pp. 142–157, 2009.
[47] A. Norouzi and R. Uzsoy, "Modeling the evolution of dependency between demands, with application to inventory planning," *IIE Trans.*, vol. 46, no. 1, pp. 55–66, 2014.
[48] A. Norouzi, "The effects of forecast evolution on production planning with resources subject to congestion," Ph.D. dissertation, E. P. Fitts Dept. Ind. Syst. Eng., North Carolina State Univ., Raleigh, NC, USA, 2012.
[49] R. Güllü, "A two-echelon allocation model and the value of information under correlated forecasts and demands," *Eur. J. Oper. Res.*, vol. 99, no. 2, pp. 386–400, 1997.
[50] P. Glasserman, "Bounds and asymptotics for planning critical safety stocks," *Oper. Res.*, vol. 45, no. 2, pp. 244–257, 1997.
[51] L. B. Toktay and L. M. Wein, "Analysis of a forecasting-production system with stationary demand," *Manag. Sci.*, vol. 47, no. 9, pp. 1268–1281, 2001.
[52] J. W. Fowler and J. Robinson, *Measurement and Improvement of Manufacturing Capacity (MIMAC) Final Report*, SEMATECH, Austin, TX, USA, 1995.
[53] MASM. (1997). *Data Sets*. [Online]. Available: http://p2schedgen.fernuni-hagen.de/index.php?id=242
[54] E. M. Scheuer and D. S. Stoller, "On the generation of normal random vectors," *Technometrics*, vol. 4, no. 2, pp. 278–281, 1962.
[55] (2019). *Detailed Simulation Results*. [Online]. Available: http://p2schedgen.fernuni-hagen.de/index.php?i-d=242
[56] W. J. Conover, *Practical Nonparametric Statistics*. New York, NY, USA: Wiley, 1980.
[57] E. Albey, Ü. Bilge, and R. Uzsoy, "Multi-dimensional clearing functions for aggregate capacity modelling in multi-stage production systems," *Int. J. Prod. Res.*, vol. 55, no. 14, pp. 4164–4179, 2017.
[58] N. B. Kacar, L. Mönch, and R. Uzsoy, "Modeling cycle times in production planning models for wafer fabrication," *IEEE Trans. Semicond. Manuf.*, vol. 29, no. 2, pp. 153–167, May 2016.
[59] N. Brahimi, T. Aouam, and E.-H. Aghezzaf, "Integrating order acceptance decisions with flexible due dates in a production planning model with load-dependent lead times," *Int. J. Prod. Res.*, vol. 53, no. 12, pp. 3810–3822, 2015.
[60] T. Aouam, K. Geryl, K. Kumar, and N. Brahimi, "Production planning with order acceptance and demand uncertainty," *Comput. Oper. Res.*, vol. 91, pp. 145–159, Mar. 2018.