Chance-Constraint-Based Heuristics for Production Planning in the Face of Stochastic Demand and Workload-Dependent Lead Times

Tarik Aouam and Reha Uzsoy

Abstract While the problem of planning production in the face of uncertain demand has been studied in various forms for decades, there is still no completely satisfactory solution approach. In this chapter we propose several heuristics based on chance-constrained models for a simple single stage single product system with workload-dependent lead times, which we compare to two-stage and multi-stage stochastic programing formulations. Exploratory computational experiments show promising performance for the heuristics, and raise a number of interesting issues that arise in comparing solutions obtained by the different approaches.

1 Introduction

In today's global supply chains, effective coordination of operations across space and time is vital to capital-intensive industries like semiconductor manufacturing with short product life cycles and rapidly changing market conditions. However, despite the fact that problems related to the planning of production and inventories have been the stock in trade of industrial engineering and operations research for the last five decades, a comprehensive solution to the problem as faced in industry is still unavailable [65]. Current research has followed the basic paradigms of deterministic mathematical programing and stochastic inventory models, resulting in highly compartmentalized streams of research that each focus on certain aspects of the

T. Aouam
School of Business Administration, Al Akhawayn University,
P.O. Box 104, 53000 Ifrane, Morocco
e-mail: t.aouam@aui.ma

R. Uzsoy (✉)
Edward P. Fitts Department of Industrial and Systems Engineering,
North Carolina State University, Campus Box 7906,
Raleigh, NC 27695-7906, USA
e-mail: ruzsoy@ncsu.edu

D. Armbruster and K. G. Kempf (eds.), *Decision Policies for Production Networks*,
DOI: 10.1007/978-0-85729-644-3_8, © Springer-Verlag London 2012

problem at the expense of others. In particular, the problem of planning production releases and allocating production capacity among different products has ignored the nonlinear congestion effects induced by capacitated resources subject to queueing, and has been treated in isolation from the problem of maintaining service levels in the face of stochastic demand.

This work is motivated by two basic limitations of the mathematical programing models used for production planning in both industrial practice and academia. The first of these is that the vast majority of these models fail to capture the nonlinear relationships between work in process inventories (WIP), cycle times, and work releases. Queueing models of production systems [17, 52] show that cycle times increase nonlinearly with resource utilization, which in turn is determined by the release plan produced by the planning system. Capital-intensive industries such as semiconductor manufacturing, with long, complex production processes, must run at high utilization to be profitable. Under these conditions small fluctuations in utilization may cause large changes in cycle times, rendering the effects of this dependence important to effective planning.

In addition to this nonlinear dependence, uncertain demand is a fact of life in most supply chains, requiring the deployment of safety stocks to ensure desired customer service levels. The production of these safety stocks, in turn, requires the release of additional work, affecting cycle times, and hence the work release and capacity allocation decisions made by planning models. It is thus notable that the planning of safety stocks [43, 107] has largely been addressed separately from capacity allocation, presumably due to the motivation for much inventory research arising from retail and distribution applications.

The large size and stochastic nature of industrial supply chain planning problems renders their exact solution computationally prohibitive. Thus industrial practice requires efficient approximations with reliable solution quality. However, when approximations are proposed, assessing the quality of their solutions is fraught with all the difficulties encountered in evaluating the quality of heuristic solutions for deterministic optimization problems [92]. There is thus a need to develop exact solution methods to provide insight into the structure of optimal solutions, as well as benchmarks against which different approximation methods can be compared and assessed.

This, in turn, presents additional complications. Problems of production planning and control in the face of stochastic demand admit several different formulations that often have quite different assumptions, advantages and drawbacks. Inventory and queueing models [107], for instance, tend to produce optimal solutions under steady-state conditions, but have difficulty in addressing transient solutions. Conventional mathematical programing models [103] solve a deterministic approximation to the actual stochastic problem, sometimes with inventory targets based on off-line analysis included as constraints. Stochastic dynamic programming models, including Markov decision processes [89], give state-based reactive decision rules that do not directly consider information about future demand that may be available. Stochastic programing [12] and chance-constrained models [87] make different assumptions about recourse actions that can result in subtle theoretical and practical difficulties.

A conclusive, unifying solution to these complex issues is clearly a long way in the future. Our objective in this chapter is more modest and exploratory in nature. We consider a simple single-stage single product production-inventory system subject to workload-dependent lead times and stochastic demand. We then develop a number of alternative formulations for this system, including two different chance-constrained models, a two-stage stochastic programing model, and a multi-stage stochastic programing approach. The multistage stochastic programing model is the only one of these that has potential to yield an exact solution, and that conditional upon the choice of scenarios; the other three are heuristics. We compare the solutions obtained from these different models by subjecting them to a simulation of uncertain demand realizations. Our exploratory computational experiments suggest that when parameters are appropriately chosen, heuristics based on chance constrained models may provide near-optimal solutions that are competitive with those from much larger stochastic programing models, although the stochastic programing models consider a very limited number of scenarios. Our results suggest a number of directions for future work on improving the heuristics, and further experimentation aimed at elucidating the strengths and limitations of the chance constraint-based heuristics.

2 Previous Related Work

A comprehensive review of the literature on production planning under uncertainty is clearly beyond the scope of this chapter. Instead, we briefly review the literature most relevant to this paper. Overviews of the production planning domain are given by de Kok and Fransoo [27], Voss and Woodruff [103] and Missbauer and Uzsoy [78].

Most deterministic production planning models establish optimal production, inventory and release levels over a given finite planning horizon to meet the total demand [16, 45, 50]. The planning horizon is divided into discrete periods during which production and demand rates are assumed to be constant; the capacity of the system is represented by the number of hours available on key resources in a planning period; and the production, inventory, WIP and demand associated with a period are treated as continuous quantities. These models allocate capacity to products to optimize a specified objective and satisfy aggregate constraints representing system capacity and dynamics. However, models of this type are subject to the utilization-lead time dependence discussed in Sect. 1. The estimates of cycle times used in planning models are referred to as *lead times*.

The most common approximation in both the research literature and industrial practice is to treat lead times as a fixed, exogenous quantity independent of resource load. The Material Requirements Planning (MRP) approach [82] uses fixed lead times in its backward scheduling step to determine job releases. Several authors have suggested ways of adapting MRP to uncertain demand. Meal [73] and Grubbstrom [39] derive component plans with safety stocks in the MRP records. Miller [75] proposes hedging of the master schedule to provide safety stocks within the system. However, all these approaches assume fixed exogenous lead times.

Another common approach to production planning under fixed lead times and deterministic demand is the use of linear(LP) and integer programing(IP) models, of which a wide variety exist [42, 56, 103]. These represent capacity as a fixed upper bound on the number of hours available at the resource in a period, and model input and output time lags between stages. However, these time lags are independent of workload.

Several authors have proposed enhanced models that address the dependency between lead times and resource utilization to some degree. Lautenschlager and Stadtler [69] suggest a model where the production in a given period becomes available over several future periods. Voss and Woodruff [103] propose a nonlinear model where the function linking lead time to workload is approximated as a piecewise linear function. Kekre et al. [63] and Ettl et al. [31] take a similar approach, adding a convex term representing the cost of carrying WIP as a function of workload to the objective function. Graves [37], Karmarkar [61], Missbauer [76], Anli et al. [2] and Asmundsson et al. [5, 4] use nonlinear clearing functions to model the dependency between workload and lead times. Several related models are proposed in the recent book by Hackman [41]. Pahl et al. [83] and Missbauer and Uzsoy [78] review production planning models with load-dependent lead times. We shall discuss clearing functions, which are used in the models in this chapter, more extensively in the next section.

Another approach to modeling the operational dynamics of the system has been the use of detailed simulation or scheduling models in the planning process. Dauzere-Peres and Lasserre [26] use a scheduling model to check whether the plans their IP model develops are feasible. Other approaches use simulation models in the same manner, e.g., Pritsker and Snyder [88]. The use of simulation or scheduling models captures the operational dynamics of the system correctly. However, this approach does not scale well, since simulation models of large systems are time-consuming to run and analyze. An innovative approach to integrating simulation and LP is that of Hung and Leachman [53]. Given initial lead-time estimates, an LP model for production planning is formulated and solved. The resulting plan is fed into a simulation model to estimate the lead-times the plan would impose on a real system. If these lead-times do not agree with those used in the LP, the LP is updated with the new lead-time estimates and resolved. This iteration is repeated until convergence. Similar models have been proposed by others [6, 18, 19, 66, 95]. However, the convergence of these methods is not well understood [55, 57]. The computational burden of the simulation runs required is also a significant disadvantage for large systems such as those encountered in semiconductor manufacturing.

Stochastic inventory models seek an optimal inventory policy (when to order, and how much to order) for individual items in the face of different environmental conditions (e.g. demand patterns, modes of shipment from suppliers) and constraints (e.g. supply restrictions, budget limitations, and desired customer service levels). Much of the work in this area [54, 59, 101, 102] is in the context of ordering from suppliers, modeling demand carefully but treating supply as known and unlimited, generally with a fixed lead time. Many subsequent papers have addressed variations of this basic problem [46, 47, 107]. However, the vast majority assume that a supplier can supply any amount of material within the specified lead time, i.e., has unlimited capacity.

Federgruen and Zipkin [32, 33] consider the capacitated inventory problem with uncertain demand and explore the optimality of "modified" base stock policies when the cost for the single period is convex in the base stock level. Tayur [99] extends this work by discussing the computation of the optimal base stock level. However, these models use simple capacity constraints that ignore the dependency between load and lead times. Ciarallo et al. [23] describe the structure of optimal policies for problems with uncertain production capacity and a time-stationary demand distribution. Anupindi et al. [3] provide bounds and heuristics for the problem with nonstationary demand and stochastic lead times, where the lead time distribution is stationary over time.

The idea of combining inventory and queueing models has attracted attention from many researchers [17, 52, 91]. Zipkin [106] develops a queueing framework to analyze supply chains facing a stationary demand distribution and where a (Q, r) policy is used to release units onto the shop floor. Ettl et al. [31] develop an optimization model combining queueing and inventory models to set base-stock levels for a multi-item batch production system facing non-stationary demands. Liu et al. [71] extend this approach.

One of the most popular frameworks for planning under uncertainty is stochastic programing [12, 58, 87]. Uncertainty is represented by using a number of discrete scenarios to represent possible future states, which allows stochastic linear programs to be modeled as large linear programing problems. Constraints are formulated requiring that an optimal solution be feasible for all scenarios, and the objective function is usually to minimize the expected value of the specified objective function. A number of authors have formulated production planning problems as multi-stage stochastic linear programs (M-SLPs) [48, 85], but the approach presents challenges.

A significant difficulty of M-SLPs is that the problem size tends to grow exponentially with the number of possible realizations (scenarios) of uncertain parameters, requiring solution methods that exploit their special structure. The scenario-based structure of M-SLPs makes decomposition methods attractive. Most decomposition methods exploit convexity of the recourse function to use outer linearization. Commonly used methods include Dantzig-Wolfe decomposition (inner linearization) and Benders decomposition (outer linearization), which decompose the large-scale problem into a master problem and several independent subproblems. Dantzig-Wolfe decomposition adds new columns to the master problem based on the suproblem solutions [25]. Benders decomposition, on the other hand, proceeds by adding new constraints (supporting hyperplanes known as optimality cuts) that are computed using dual solutions to the subproblems (e.g., Lasdon [68]).

Van Slyke and Wets [100] extended Benders' decomposition to solve two-stage stochastic linear programs (2-SLPs) via the L-Shaped Method. M-SLPs are much more challenging computationally than 2-SLPs. An extension of the L-shaped method to more than two stages, called nested decomposition, was first proposed by Louveaux [72] for multi-stage quadratic programs and by Birge [11] for multistage linear programs. The algorithm generates cuts for an ancestor scenario problem that has feasible completion in all descendant scenarios. As in the L-shaped method, nested decomposition achieves outer linearization by generating feasibility and optimality cuts until it converges to an optimal solution. A number of different strategies have been used to select the next subproblem for deterministic problems. Numerical experiments by Gassmann [35] found that the fast-forward-fast-back procedure of Wittrock [105] outperforms other strategies.

There have been recent attempts to model production planning problems using robust optimization approaches [7, 10]. Leung et al. [70] develop a robust optimization model to solve the aggregate production planning problem. Raa and el-Aghezzaf [90] use robust optimization to obtain a dynamic planning strategy for the stochastic lot-sizing problem.

Chance constrained programing dates back to the work of Charnes and Cooper [20, 21, 22]. A more recent overview of these methods is given by Prékopa [87]. In chance constrained programing, constraints can be violated with a specified probability, which is quite useful to model, for instance, service levels in supply chain problems [40]. Continuous probability distributions are often assumed on the uncertain parameters. This approach achieves a substantial decrease in the size of the model, and avoids the problem of defining the penalty function. However, it fails to capture the cost consequences of constraint violations, which can result in anomalous behavior [14].

Given that exact solutions to stochastic optimization problem are computationally challenging, a number of approaches to obtain solutions via decision rules have been proposed. These approaches classify decision variables according to whether they are implemented before (first stage decisions), or after (second stage decisions) an outcome of the random variable(s) is observed. However, in the decision rule-based approach, the second stage recourse decisions are determined by a rule that incorporates both the first stage decisions and the observed outcomes. A commonly encountered example of such a rule that is in fact optimal in form is the well-known base stock policy for inventory systems with unlimited capacity, deterministic replenishment lead time and linear holding and backorder costs. However, as pointed out by Garstka and Wets [34], the decision rule approach assumes a specific form for the optimal solution to the stochastic program. Since very few multistage stochastic programs yield a closed-form characterization of the optimal solution, solutions obtained assuming decision rules cannot be guaranteed to be optimal in the vast majority of cases.

A well-known family of decision rules are the Linear Decision Rules, where the second stage recourse decision is a linear function of the first stage decision variables and the observed outcomes. The pioneering Linear Decision Rule (LDR) was developed by Holt, Modigliani, Muth and Simon (HMMS) in the mid 1950s [49, 51]. Extensions to this rule have been proposed by several authors [8, 28, 36, 44, 84]. While the HMMS model and its variations incorporate demand uncertainty, these models treat capacity, specifically workforce levels, as a decision variable that can be varied continuously, which avoids the problem of workload-dependent lead times encountered under fixed capacity limits. In addition, the specific quadratic form of the objective function adopted allows the construction of a deterministic equivalent that simply replaces each random variable with its expectation. However, it is well known that this approach does not yield optimal solutions in general.

In summary, a variety of models have been proposed that address the issues of workload-dependent lead times and demand uncertainty separately at best, and in many cases do not address either. The LP and MRP approaches do not address workload-dependent lead times, and generally ignore stochastic demand. Most inventory models focus on modeling demand, with simple models of replenishment that do not consider workload-dependent lead times. The combined queueing-inventory models capture the interaction between workload-dependent lead times and inventory levels correctly, but assume specific inventory policies of the order up to type, and make different assumptions about the representation of a production unit. Stochastic programing approaches are hampered by their exponentially growing computational burden as the number of products and planning periods (stages) increase. Our heuristics, in contrast, consider non-stationary demand distributions to provide production plans over a finite planning horizon, taking available information about future demand into account. The work in this paper is an initial step in assessing the performance of this approach.

In the next section we present an overview of the clearing function concept that we use to develop a LP model that addresses the load dependent lead time and demand uncertainty aspects simultaneously for a single-product supply chain.

3 Clearing Function Basics

Clearing functions (CF) [37, 61, 78, 98], express the expected throughput of a capacitated resource over a given period of time as a function of some measure of WIP level at the resource over that period, which in turn, is determined by the average resource utilization over the period. We shall use the term "WIP" and the generic variable W to denote any reasonable measure of WIP level over a planning period.

To motivate the use of a nonlinear CF, it is helpful to begin with a single resource that can be modeled as a *G/G/1* queueing system in steady state. The expected number in system (i.e., expected WIP) for a single server is given by Medhi [74] as:

$$W = \frac{(c_a^2 + c_s^2)}{2}\frac{\rho^2}{(1-\rho)} + \rho$$

where c_a and c_s denote the coefficients of variation of service and interarrival times, respectively and ρ the utilization of the server. Setting $c = (c_a^2 + c_s^2)/2$ and rearranging (1) we obtain a quadratic in W whose positive root yields the desired ρ value. Solving for ρ with $c > 1$, we obtain

$$\rho = \frac{\sqrt{(W+1)^2 + 4W(c^2-1)} - (W+1)}{2(c^2-1)}$$

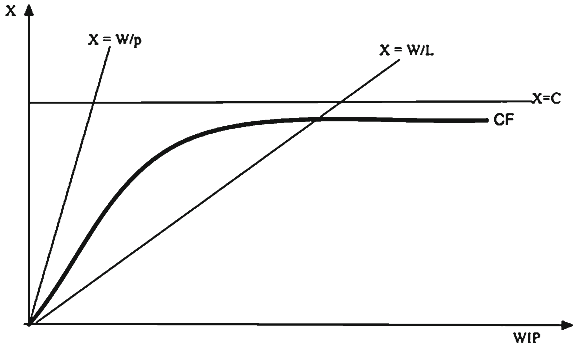

Fig. 1 Examples of CFs (Karmarkar [61])

which has the desired concave form. When $0 \leq c < 1$, the other root of the quadratic will always give positive values for ρ. When $c = 1$, the expression simplifies to yield $\rho = W/(1+W)$, again of the desired concave form. If we use utilization as a surrogate for output, we see that for a fixed c value, utilization, and hence throughput, increase with WIP but at a declining rate. Utilization, and hence output, is decreasing in c due to variability in service and arrival rates.

Figure 1, derived from Karmarkar [61], depicts several examples of CFs considered in the literature, where X denotes the expected throughput in a planning period. The horizontal line $X = C$ represents a fixed upper bound on output over the period, but without a lead-time constraint it implies that production can occur without any WIP in the system if work input and production are synchronized. This approach is implemented in most LP models but is supplemented with a fixed lead time as described above. The linear CF of Graves [37] is represented by the $X = W/L$ line, which implies a lead time of L periods that is maintained independently of the WIP level. If a fixed lead time is maintained up to a certain maximum output, we have $X = \min\{W/L, C\}$. When the parameters of the Graves CF are set such that the lead time is equal to the average processing time, with no queueing delays at all, we obtain the line $X = W/p$, where p denotes the average processing time. Assuming that lead time is equal to the average processing time up to a maximum output level gives the "Best Case" model $X = \min\{W/p, C\}$ of Hopp and Spearman [52]. The workload-independent fixed lead time in most LP models differs from the linear CF of Graves in that the former does not link output to WIP, while the latter does [81]. The CF always lies below the $X = W/p$ and $X = C$ lines. For most capacitated production resources subject to congestion, limited capacity leads to a saturating (concave) shape of the CF, for which Asmundsson et al. [4] and Selçuk et al. [96] provide analytical support.

Several authors discuss the relationship between throughput and WIP levels in the context of queueing analysis, focusing on the long-run steady-state expected throughput and WIP levels. Agnew [1] studies this behavior in the context of optimal control policies. Spearman [97] presents an analytic congestion model for a family of closed production systems that describes the relationship between throughput and WIP. Srinivasan et al. [98] derives the CF for a closed queueing network with a product form solution. Asmundsson et al. [4, 5] and Missbauer [77] study the problem of estimating CFs from experimental data, obtained either from industry or simulation models. Missbauer and Uzsoy [78] review the state of the art in this area.

An important advantage of CFs for our purposes is their ability to reflect different sources of variability in the production process. In queueing terms, this is accomplished by basing the CF on the effective processing time at the resources, which includes the effects of detractors such as uncertain yield, machine failures and setups, as discussed in Chap. 8 of Hopp and Spearman [52]. The manner in which these effects change the shape of the CF is described in Asmundsson et al. [4]. When the CFs are estimated from empirical data, the effects of the variability induced by detractors are present in the data to which the CF is fit, again capturing their effects.

Hence, given the current research on the derivation of CFs using both analytical and empirical approaches, in this chapter we shall proceed on the assumption that adequate methods of estimating CFs for different production systems will emerge from ongoing work. We focus on using CFs to develop production models that consider stochastic demand and the nonlinear relationship between workload and cycle in an integrated manner. We introduce our approach in the next section.

4 A Deterministic Model Based on Clearing Functions

In this section we develop a LP model for aggregate planning under the effects of congestion and demand uncertainty. We begin with a basic formulation prevalent in the literature, discuss its weaknesses, and use these to motivate our formulations, drawing heavily on the exposition in Bookbinder and Tan [15]. While there are clearly many formulations in the literature that capture additional aspects such as multiple stages, alternative production paths, etc., our focus is to find computationally tractable formulations that allow us to treat both the nonlinear dynamics of utilization and lead times and the stochastic nature of the demand as endogenous to the model. Hence to isolate these aspects of the problem for study, we focus on a single-stage single product system. The quantity of raw material released into the system in each time period is the key decision variable in our models. These releases are then converted into output according to different mechanisms defined by the models considered, which will be discussed as we proceed.

Consider the production planning problem for such a single stage production system producing a single product. The planning horizon is divided into T discrete periods of equal length. Demand in each period is assumed to be stochastic with known cumulative distribution function (CDF), and independent of demand in other periods. Service level requirements to be met are prespecified, and are thus treated as a constraint. We consider the simple objective of minimizing the sum of expected costs of holding finished goods inventory (FGI) and work in process (WIP) over the planning horizon. Following the literature, we do not consider the cost of stockouts in the objective function because we assume that the service level requirements are sufficiently high that the cost of stockouts is negligible. This assumption will be relaxed in our computational experiments. Clearly far more elaborate objective functions are possible, but our emphasis is on representation of production capacity and demand uncertainty.

To describe the models used in this paper we use three different classes of variables:

Decision or Control Variables: These variables represent the primary management decisions in a plan. In order to be implementable, a plan must specify either specific values for these variables, or specific rules by which they can be computed with the information available at the time a decision must be made.

State Variables: These variables define the behavior of the system, and their values are determined by the constraints determining the operational dynamics of the system and the values of the decision variables. These variables may be either deterministic or random.

Parameters: These are external inputs to the system and are prespecified in the model. We will assume these are always deterministic.

The notation used in the formulations is given below. We use a bold font, e.g., **X**, for a random variable and a normal font, e.g., *X*, for a deterministic variable.

R_t : Planned quantity of product released into the system during period *t*

X_t : Planned production quantity during period *t*

$\mathbf{I_t}$: Inventory on hand at the end of period *t*. The initial inventory on hand at the start of period 1 will be denoted by I_0.

h_t : Unit inventory holding cost for period *t*

C_t : Capacity, e.g., total number of machine hours available, in period *t*

α : Specified service level

$G_{[t,t+k]}$: CDF of cumulative demand from period *t* to period $t+k$

$\mathbf{D_t}$: Demand during period *t*. Throughout this paper we shall assume the demand in each period *t* to be normally distributed with known mean μ_{t} and standard deviation σ_t. In our experiments we will assume demands are independent by time periods. However, the models presented remain valid for correlated demands as long as the variance-covariance matrix is known, or can be estimated with reasonable accuracy.

L_t : Average lead time in period *t*. For simplicity of exposition in presenting the models in this section we shall assume these are integer multiples of the planning period length. Fractional L_t values can be accommodated in a straightforward manner.

4.1 Basic Formulation

Most chance-constrained production planning models in the literature are similar to that of Bookbinder and Tan [15] given below; a slightly different version is given in Johnson and Montgomery [56]. Our model incorporates the following constraints:

- *Releases*

Since the lead time is L_t in period t, whatever is released into the system in period t is converted to output and available for consumption in period $t + L_t$. Hence the relationship between release quantities and output is given by

$$R_t = X_{t+L_t}, \text{ for all } t = 1, \ldots, T - L_t$$

The primary decision variable is the amount of work R_t released in period t, which must be specified at the start of the planning horizon. Hence both releases and production are deterministic. Note that the X_t and R_t variables are redundant, and the formulation can be written with only one of these two sets of variables.

In this model, work that is released into the production system at time t is in WIP for L_t periods until it emerges as finished product. Most LP models do not explicitly represent this quantity, or assign it a cost in the objective function, but it can easily be estimated for any period t as the difference between the cumulative releases and output up to a given period t.

- *Inventory balance*

The finished goods inventory on hand at the end of period t, $\mathbf{I_t}$, is a random variable for which the relationship

$$\mathbf{I_t} = \mathbf{I_{t-1}} + X_t - \mathbf{D_t}$$

holds for each time period t. Taking the expectation and repetitive substitution yields

$$\begin{aligned} E[\mathbf{I_t}] = E[\mathbf{I_{t-1}}] + X_t - E[\mathbf{D_t}] &= I_0 + \sum_{i=1}^{t} X_i - \sum_{i=1}^{t} E[\mathbf{D}_i] \\ &= I_0 + \sum_{i=1}^{t} X_i - \sum_{i=1}^{t} \mu_i, \text{ for all } t = 1, \ldots, T. \end{aligned}$$

All terms in this expression are now deterministic.

- *Capacity*

$$X_t \leq C_t, \text{ for all } t = 1, \ldots, T.$$

Service level: This constraint requires that the service level, defined by the probability of $\mathbf{I_t} < 0$, i.e., a stockout occurring, be less than $(1 - \alpha)$, implying

$$P\{\mathbf{I_t} \geq 0\} \geq \alpha \Rightarrow P\left\{ I_0 + \sum_{i=1}^{t} X_i \geq \sum_{i=1}^{t} \mathbf{D_i} \right\} \geq \alpha, \text{ for all } t = 1, \ldots, T.$$

The service level measure fits the chance constraint approach well, since the latter allows constraints to be violated with a certain probability. However, it does not

Table 1 Basic formulation

$\min \sum_{t=1}^{T} h_t\{I_0 + \sum_{i=1}^{t} X_i - \sum_{i=1}^{t} \mu_i\}$	subject to	
$P\{I_0 + \sum_{i=1}^{t} X_i \geq \sum_{i=1}^{t} \mathbf{D_i}\} \geq \alpha$	for all $t = 1, \ldots, T$	(SERVICE LEVEL)
$X_t \leq C_t$	for all $t = 1, \ldots, T$	(CAPACITY)
$X_t \geq 0$	for all $t = 1, \ldots, T$	(NONNEGATIVITY)

capture the degree to which the constraint is violated. Hence a production plan that has stockouts in many periods, but falls short by a very small fraction of the demand in each period, will appear to have a poor service level. To this end, we use the fill rate, the fraction of total period demand met from inventory, as another performance measure in our computational experiments. The basic formulation is summarized in Table 1, using the production variables X_t.

While the basic formulation is intuitive, it suffers from the following disadvantages:

- It ignores the effects of loading on WIP and lead times in a capacitated system [4, 5, 37, 60] by considering lead times to be fixed exogenous values.
- It assumes that safety stock must be held in finished goods inventory, based on the demand for each individual period. This is adequate when the lead times of the production system, which correspond to the replenishment time of the finished goods inventory, do not exceed one period, as in the models of Bookbinder and Tan [15] and Johnson and Montgomery [56]. However, if lead times span multiple periods, this becomes problematic. It is well known in the inventory literature [24] that in the presence of nonzero lead times the optimal policy in many cases, and a good heuristic in many more, is to set the inventory position, the sum of on-hand inventory and outstanding orders, to the desired percentile of the demand over the lead time (e.g., [29]). Hence this formulation fails to recognize that WIP can serve some of the function of safety stock [38], and hence might hold more finished goods inventory than required to maintain a given service level. We shall assume a replenishment policy of this form, which is not optimal for the production system we consider, in developing our heuristics.

  In the production-inventory context of this paper, outstanding orders are represented by material that has been released into the production line but has not yet emerged as finished goods, i.e., WIP [38]. The inventory position, which will be an important quantity for our development in the rest of this paper, will be defined in more detail in the following section.
- The model makes all decisions for the entire planning horizon at the beginning of the horizon, before any of the demands become known, and does not provide a way to use information as it becomes available. In other words, there is no recourse action.

In the rest of this section we extend this formulation to address these issues.

4.2 Development of Integrated Model

An elegant way of capturing the effect of capacity loading on WIP and lead times in production planning models is the use of clearing functions (CFs) as discussed in Sect. 3. Recall that up to this point all variables are deterministic except the inventory levels $\mathbf{I_t}$. Hence, incorporating CFs in the model requires:

- Introduction of WIP balance equations. If W_t is defined to be the WIP at a given time t, then the WIP balance equations are given as $W_t = W_{t-1} + R_t - X_t$ for all periods $t = 1, \ldots, T$. We treat R_t as a deterministic decision variable that is specified at the start of the planning horizon by solution of the planning model, and cannot be modified as uncertainty is realized.
- Replacing the original capacity constraint with a set of linear inequalities that represent the outer linearization of the original CF [4, 5]. The set of inequalities representing the CF is given by $X_t \le a_k W_{t-1} + b_k$, for all periods $t = 1, \ldots, T$ and line segments $k = 1, \ldots, n$ used to outer linearize the CF.

The use of CFs to represent the capacity of the production system takes a more complex view of the relationship between the planned release quantity R_t in period t and the planned output X_t of the system in that period. The releases in a period determine the planned WIP level W_t at the end of the period, together with the linearized CF represented by the constraints above, determines the planned system output X_{t+1} in the next period The release variables R_t are defined such that releases are made at the end of period t, and hence cannot contribute to output during period t. This is necessary because in later models, our linear decision rule observes the realization of the random demand $\mathbf{D_t}$ in period t to determine the releases R_t at the end of period t. This definition, together with the definition of the CF and the WIP balance equations, is thus internally consistent.

In inventory theory an optimal or near-optimal policy, when there is no fixed ordering cost and shortage and holding costs are linear, is to maintain the inventory position, the sum of on-hand and on-order inventory, at a critical fractile of the demand over the replenishment lead time [24]. Hence if $\mathbf{IP_t}$ denotes the inventory position at the end of period t, we have $\mathbf{IP_t} = W_t + \mathbf{I_t}$, where W_t represents orders that have been released to production but not yet completed. This analogy with inventory models suggests a service level constraint requiring a probability α that $\mathbf{IP_t}$ is at least as great as the demand over the replenishment lead time [38]. Assuming this replenishment lead time, corresponding to the cycle time of the production system under study, is known to be L_t periods in period t, we have

$$P\left\{\mathbf{IP_t} \ge \sum_{i=t+1}^{t+L_t} \mathbf{D}_i\right\} \ge \alpha \Rightarrow P\left\{\mathbf{I_t} + W_t \ge \sum_{i=t+1}^{t+L_t} \mathbf{D}_i\right\} \ge \alpha\,.$$

The L_t parameters on the right hand sides of our chance constraints define the distribution of the lead time demand that will be used to set safety stock levels. Noting that

$$\mathbf{I_t} = I_0 + \sum_{i=1}^{t} X_i - \sum_{i=1}^{t} \mathbf{D}_i \text{ and}$$

$$W_t = W_0 + \sum_{i=1}^{t} R_i - \sum_{i=1}^{t} X_i$$

we obtain

$$\mathbf{IP}_t = \mathbf{I_t} + W_t = (I_0 + W_0) + \sum_{i=1}^{t} R_i - \sum_{i=1}^{t} \mathbf{D}_i.$$

The chance constraint is now of the form

$$P\{\mathbf{IP}_t \geq 0\} \geq \alpha \Rightarrow P\left\{I_0 + W_0 + \sum_{i=1}^{t} R_i - \sum_{i=t+1}^{t+L_t} \mathbf{D}_i \geq \sum_{i=1}^{t} \mathbf{D}_i\right\} \geq \alpha$$
$$\Rightarrow \; P\{I_0 + W_0 + \sum_{i=1}^{t} R_i \geq \sum_{i=1}^{t+L_t} \mathbf{D}_i\} \geq \alpha\,.$$

Following the approach of Charnes and Cooper [22] the deterministic equivalent of the service level constraint can be written as

$$I_0 + W_0 + \sum_{i=1}^{T} R_i \geq G^{-1}_{[1,t+L_t]}(\alpha), \text{ for all } t = 1, \ldots, T.$$

where $G_{[1,\,t]}(\cdot)$ denotes the cumulative distribution function (CDF) of the cumulative demand random from periods 1 to *t,*

Replacing the probabilistic service level constraint with its deterministic equivalent yields the Zero-Order Inventory Position (ZOIP) formulation shown in Table 2. This formulation embodies a service level constraint on inventory position and a zero order decision rule where all decision variables are specified irrevocably at the start of the planning horizon.

It is important to note that there are two different lead times at work in the ZOIP model. The first of these is the estimated replenishment lead time L_t used to establish the inventory position required to approximately achieve the desired service levels. The second lead time in question is that realized in the production system, the time required for work released into the system to become available as finished product. The workload-dependent nature of this realized lead time is explicitly represented by the clearing function, whose effectiveness for this purpose we have demonstrated in prior work [4, 5]. Ideally, the two lead times should be equal, with the replenishment lead time used for setting inventory targets matching that realized by the production

Table 2 ZOIP formulation

$\min \sum_{t=1}^{T} h_t\{I_0 + \sum_{i=1}^{t} X_i - \sum_{i=1}^{t} \mu_i + W_t\}$	subject to	
$W_t = W_{t-1} - R_t - X_t$	for all $t = 1, \ldots, T$	(WIP BALANCE)
$I_0 + W_0 + \sum_{i=1}^{T} R_i \geq G^{-1}_{[1,t+L_t]}(\alpha)$,	for all $t = 1, \ldots, T$	(SERVICE LEVEL)
$X_t \leq a_k W_{t-1} + b_k$	for all $t = 1, \ldots, T; k = 1, \ldots n$	(CAPACITY)
$X_t, R_t, W_t \geq 0$	for all $t = 1, \ldots, T$	

system in the face of the release schedule recommended by the model. In other words, in an ideal situation the L_t should be an output of the model. This would require us to estimate L_t using Little's Law as $(W_t + W_{t-1})/2X_t$ assuming the planning periods are long enough for the law to apply; for the shorter periods some transient version of Little's Law such as those discussed by Bertsimas and Mourtzinou [9], Whitt [104] and Riaño [95] would be required. Even the use of the classical stationary version of Little's Law yields a highly nonlinear constraint. Hence for the sake of tractability we treat the replenishment lead time L_t on the right hand side of the chance constraints as an exogenous parameter, which reduces the right hand sides to constants that can be precomputed easily. Our model thus captures workload-dependent lead times correctly in defining the relationship between releases R_t, planned WIP level W_{t-1}, and expected output X_t, but uses an exogenous parameter to approximate the distribution of the lead time demand, which will be used to set the safety stocks. Computational experiments indicate that the realized lead time may deviate somewhat from the exogenously assumed value used to establish the chance constraints when used in this manner, but results are still favourable over base stock type models that do not consider clearing functions [94].

A full resolution of this issue appears to be challenging, and must be left for future research. A promising approach is to use an iterative scheme, where we solve the ZOIP model using an initial set of lead time estimates to obtain a release plan, i.e., a set of R_t values. These R_t values are then used to compute the resulting state variables X_t, W_t, and I_t, from which a new set of L_t values can be estimated as $L_t = W_t/X_t$. These new L_t values are then substituted into the model and the process is repeated until convergence is, hopefully, achieved. Orcun et al. [79] have implemented this procedure with favourable results, but formal analysis of its convergence remains for future work.

Up to this point we have developed a formulation that combines the modeling of congestion and lead times in the production system with the explicit representation of random demand using chance constraints. We now move on to adding flexibility to the decision mechanism by utilizing information as it becomes available.

4.3 A Linear Decision Rule

So far our formulations have zero order static decision rules, where the values of all decision variables are determined at the beginning of the time horizon and there is no recourse action after the outcomes are observed. We now follow Charnes and Cooper [22] and propose a linear decision rule to introduce flexibility in the decision mechanism, recalling that this approach does not yield an optimal solution. Since the releases are the decision variables in the CF formulations, the decision rule is based on releases.

We use a simple rule closely following that described in Johnson and Montgomery [56] that allows the releases to be modified as uncertain demand is observed, rendering them random variables. We define auxiliary variables Y_t that represent the change in planned inventory position from period $t-1$ to period t, implying that $\mathbf{R_t} = Y_t + \mathbf{D_t}$. Thus Y_t represents the amount of work released in period t over and above that necessary to replenish the inventory position after that period's demand has been withdrawn; note that it may be negative, if demand is decreasing in a given time interval. This decision rule thus represents a base stock policy, and it is straightforward to show that $Y_t = IP_t - IP_{t-1}$ for specified values of $\mathbf{IP_t}$ and $\mathbf{IP_{t-1}}$. Our heuristic establishes chance constraints that set the planned inventory position at the end of period t, $\mathrm{IP_t}$, to a percentile of the lead time demand distribution as described below. The releases $\mathbf{R_t}$ are now random variables derived from the Y_t and the realized demand $\mathbf{D_t}$. Thus the WIP variables $\mathbf{W_t}$ are now also random. Since the production $\mathbf{X_t}$ in a given period depends on the realized WIP level $\mathbf{W_{t-1}}$ at the start of that period, $\mathbf{X_t}$ is also a random variable. We have the relation

$$\begin{aligned} E[W_t] &= E[\mathbf{W}_{t-1} + \mathbf{R}_t - \mathbf{X}_t] = E[\mathbf{W}_{t-1} + Y_t + \mathbf{D}_t - \mathbf{X}_t] \\ &= W_0 + \sum_{i=1}^{t} (Y_i + \mu_i - \mathbf{X}_i) \end{aligned}$$

Since the release quantities are now random variables, there exists a possibility that they may be negative. To prevent this, we use the chance constraint

$$P\{\mathbf{R_t} \geq 0\} \approx 1 \Rightarrow P\{Y_t + \mathbf{D_t} \geq 0\} \approx 1 \Rightarrow Y_t + D_t^{\min} \geq 0,$$

where $D_t^{\min}$ is a value of demand in period t such that the probability of demand falling below this level is deemed by management to be extremely small. This is clearly an approximation when demand follows a distribution with unbounded support, like the normal distribution we assume, and is unlikely to be binding except when there is a very sudden, large decline in demand from one period to another.

We again define the event of a stockout as the event that the total lead time demand exceeds the inventory position $\mathbf{IP_t} = \mathbf{W_t} + \mathbf{I_t}$, yielding the chance constraint

$$W_0 + I_0 + \sum_{i=1}^{t} Y_i \geq G^{-1}_{[t+1,t+L_t]}(\alpha).$$

Table 3 DYNIP formulation

$\min \sum_{t=1}^{T} h_t\{I_0 + \sum_{i=1}^{t} X_i - \sum_{i=1}^{t} \mu_i +$		
$W_0 + \sum_{i=1}^{t} (Y_i + \mu_i - X_i)\}$	subject to	
$W_0 + I_0 + \sum_{i=1}^{t} Y_i \geq G^{-1}_{[t+1,t+L_t]}(\alpha)$	for all $t = 1, \ldots, T$	(SERVICE LEVEL)
$X_t \leq a_k(W_0 + I_0 + \sum_{i=1}^{t-1} Y_i) + b_k$	for all $t = 1, \ldots, T; k = 1, \ldots n;$	(CAPACITY)
$Y_t + \mathrm{D}_t^{\min} \geq 0$	for all $t = 1, \ldots, T$	(REL. NON-NEG.)
$X_t, Y_t \geq 0$	for all $t = 1, \ldots, T$	

Since releases, WIP and production are all interrelated, all decision variables are now random variables except the Y_t, creating difficulties in establishing a tractable formulation. Hence for tractability in the solution procedure, we will assume that the production variables X_t and the auxiliary variables Y_t are determined at the start of the planning horizon, with the $\mathbf{W_t}, \mathbf{I_t},$ and $\mathbf{D_t}$ remaining as random variables. The assumption here is that when a production target X_t is in danger of not being met, the system will take extraordinary measures to meet it, such as running an extra shift or buying from an outside source. The cost of this is not captured in the models, but is, of course, considered in our computational experiments, where we assume the production system has no outside recourse when planned production levels cannot be achieved. This ensures that all models are treated similarly in the computational experiments. Incorporating this rule in the ZOIP formulation gives us our final Dynamic Inventory Position (DYNIP) formulation summarized in Table 3.

The models presented above have been analyzed by Ravindran et al. [93]. They compare the performance of the ZOIP and DYNIP models with a static base stock policy and find that DYNIP performs significantly better in terms of backorders. They also analyze the structure of optimal solutions to the model under the linear clearing function of Graves. These results indicate that the ZOIP model will overstock consistently, while DYNIP will not.

5 Stochastic Programing Models

For comparison with the chance constrained models, we develop two different stochastic programing models along with their implementation strategies. We first present a two-stage stochastic programing model. A multi-stage stochastic programing formulation is also presented along with static and dynamic implementation strategies.

5.1 The Two-Stage Model (2-SP)

As in the rest of the paper, we assume the primary source of uncertainty is the demand in each period, and consider the simple objective of minimizing the sum of expected WIP holding, FGI holding and backorder costs over the planning horizon of T periods. We assume that the demand evolves as a discrete time stochastic process with a finite probability space. This information structure can be interpreted as a scenario tree, where the nodes in stage t of the tree constitute the states of the world that can be distinguished by information available up to period t. The size of the scenario tree is clearly exponential in the number of periods T, and depends on the number of possible demand realizations considered at each stage.

The computational burden of any model based on scenario trees will rapidly become impractical. Therefore even for relatively small problem instances used to benchmark our heuristics, some means of reducing the size of the scenario tree must be devised. To this end, we shall follow Escudero et al. [30] and consider a two-stage formulation which consists of specifying a number of scenarios ξ composed of demand realizations for all periods. The first-stage problem involves deciding the production, release, and planned WIP levels for all periods, regardless of the state of the world. The second stage determines the FGI and backlog levels at the end of each period subject to the realized state. Thus the X_t, R_t, and W_t variables are only indexed by time periods (since they do not change with the realized state) while FGI variables I_t^{ξ} and backorder B_t^{ξ} at the end of period are indexed by the scenario ξ. The model can be stated as follows:

$$\text{Min} \sum_{t=1}^{T} h_t W_t + E_{\xi}[Q(X_t, \xi)]$$

subject to

$$W_t = W_{t-1} + R_t - X_t \quad \text{for all } t = 1, \ldots, T$$

$$X_t \leq f(W_t), \text{ for all } t = 1, \ldots, T$$

$$X_t, R_t, W_t \geq 0 \text{ for all } t = 1, \ldots, T$$

where $Q(X_t, \xi)$ denotes the recourse function which is defined as

$$Q(X_t, \xi) = \min \sum_{t=1}^{T} (h_t I_t^{\xi} + b_t B_t^{\xi})$$

subject to

$$I_t^{\xi} - B_t^{\xi} = I_{t-1}^{\xi} - B_{t-1}^{\xi} + X_t - D_t^{\xi}, \text{ for all } t = 1, \ldots, T$$

$$I_t^{\xi} - B_t^{\xi} \geq 0, \text{ for all } t = 1, \ldots, T$$

Unlike DYNIP, this model assumes no recourse for the R_t variables. In fact under the two-stage model the first stage decision variables X_t, R_t, and W_t are determined at the beginning of the planning horizon, while the second stage problem simply computes the realized FGIs and backorders after demands are realized. This model has the advantage that the size of the model grows linearly with the number of scenarios considered, and that it has complete recourse, in that all first-stage decisions are feasible for the second stage. The disadvantage is that it does not allow recourse action to be taken as demand is realized, placing it on a par with the ZOIP model in this regard.

In order to determine a 2-SP production planning strategy, one has to generate multiple scenarios, each consisting of demand realizations for periods $1, \ldots, T$. The 2-SP model is then solved and the optimal decisions $(R_t^*, X_t^*, W_t^*), t = 1, \ldots, T$ yield a production plan that is completely defined at the beginning of the planning horizon.

5.2 The Multi-Stage Model (M-SP)

A natural extension of the two-stage model is to allow recourse actions as demand is observed. This is accomplished by representing the demand process $\{D_t\}$ as a scenario tree. Each node n in the tree represents a demand realization in the corresponding period $t(n)$ with a probability q_n. The root node $(n=1)$ of the tree represents the current demand, i.e. D_1. Node *a(n)* is the direct ancestor of node n. The direct descendants of node n are called the children of node n. The subtree with root node n is denoted by $T(n)$. A path from the root node to a node n describes one realization of the stochastic process from the present (period 1) to period $t(n)$. The set of all the nodes on this path is denoted as $P(n)$. A full evolution of the demand process over the entire planning horizon, i.e., the path from the root node to a leaf node, is called a scenario.

The scenario tree representation of the demand process is an approximation of the actual demand distribution due to its use of a finite number of possible demand outcomes in each period. Also, generally the size of scenario tree increases exponentially with increasing time horizon. The cumulative demand, production, and releases for the partial realization of the demands represented by a path from the root node 1 to a node n in the tree are given by

$$D(1,n) = \sum\nolimits_{m \in P(n)} D_m$$

$$X(1,n) = \sum\nolimits_{m \in P(n)} X_m$$

$$R(1,n) = \sum\nolimits_{m \in P(n)} R_m$$

The stochastic programing formulation of the production planning problem with congestion is given by the following model:
(MSP):

$$
\begin{array}{lll}
\min & \sum_{n \in T(1)} q_n [h_{t(n)}(I_n + W_n) + b_{t(n)} B_n] & \\
\text{s.t.} & W_n = W_0 - X(1,n) + R(1,n) & \\
 & I_n = I_0 + X(1,n) - D(1,n) + B_n & \forall n \\
 & X_n \leq f(W_{a(n)}) & \forall n \\
 & R_n,\ X_n,\ I_n,\ W_n,\ B_n \geq 0 & \forall n
\end{array}
$$

The objective in the M-SP model is to minimize the expected cost over the planning horizon, which includes the present cost determined by the root node decisions and the expected future cost. In any given period t, the release, WIP, and production can be determined before the knowledge of demand, and are hence called first-stage decisions. On the other hand inventory and backorder are recourse decisions because they depend on the first-stage decisions as well as the realization of the uncertain parameter (demand). In our implementation, the constraints related to the clearing function are piecewise linearized as in Asmundsson et al. [4] for computational convenience.

The MSP model has considerable similarities to the Model Predictive Control approach deployed in the engineering disciplines. The similarities between control theoretic and mathematical programing approaches were noted early on by Kleindorfer et al. [67] and their application to supply chain management problems has been discussed by Kempf [64].

5.3 Implementation Strategies for the M-SP Model

Based on the multi-stage stochastic programming model (M-SP) we develop two production planning strategies to satisfy future demand over the planning horizon. The first strategy is a static strategy (MSP) and the second is a dynamic strategy (MSP-DYN).

5.3.1 A Static Strategy (MSP)

MSP is a static strategy, which specifies completely the release, production, and WIP decisions for all future periods at the beginning of the planning horizon. Once demands are realized, the FGI and backorders can be determined and the performance of the solution evaluated, in a manner similar to that used for ZOIP. The primary difference between ZOIP and MSP lies in the manner they model the uncertainty in the demand process. ZOIP assumes a known demand distribution in each period, and establishes constraints that may be violated with a prespecified probability. MSP, on the other hand, captures the uncertainty of demand through a limited number of demand values in each period. Another important difference between ZOIP and MSP is that ZOIP assumes no recourse action is possible as uncertain demand is revealed.

In order to determine the MSP strategy, i.e., the production planning decisions for all periods, we follow the procedure below. Note that all the steps are performed at the beginning of the planning horizon.

For $t = 1$, we construct a scenario tree $T(1)$, set the first period demand to the current demand and the initial inventories to some preset initial values, then solve the MSP model. We obtain the optimal production decisions for all the nodes in the tree, i.e., $(R_n, X_n, W_n)^*$. However we only save the root node decisions, which correspond to the decisions to be implemented in period 1, $(R_1, X_1, W_1)^*$, under the MSP strategy. Also, $(I_1, W_1)^*$ serve as initial inventories for the next period.

For $t = 2$, we construct a scenario tree $T(2)$ over the periods $2, \dots, T$, set the root node demand to μ_2 and solve the M-SP. Here μ_2 is the forecast of period 2 demand available to us in the beginning of the planning horizon.

The root node optimal decisions are recorded as $(R_2, X_2, W_2)^*$ and will be implemented in the second period under the MSP strategy. Repeat the same for $t = 3, \dots, T$. The optimal decisions $(R_t, X_t, W_t)^*,\ t = 1, \dots, T$ constitute the MSP production plan that is completely defined at the beginning of the planning horizon.

5.3.2 A Dynamic Strategy (MSP-DYN)

As pointed out in Powell et al. [86], a model is dynamic if “it incorporates explicitly the interaction of activities over time”. A model is applied dynamically if “the model is solved repeatedly as new information is received”. Under this definition, DYNIP is a dynamic model, while MSP-DYN presented below is a model applied dynamically.

In the MSP-DYN, the multi-stage SP model is applied dynamically over the planning horizon and only the decisions of the first period are implemented. As new information about demand becomes available the model is resolved and the release, production, and WIP decisions are made. Therefore, at the beginning of the planning horizon only period 1 decisions are known and future decisions will only be determined once the corresponding demand is realized. More specifically, we proceed as follows:

For the current period, $t = 1$, we construct a scenario tree $T(1)$, set the first period demand to the current demand and the initial inventories to some pre-set initial values, then solve the MSP model. We obtain the optimal production decisions for the root node decisions to be implemented in period 1, $(R_1(D_1), X_1(D_1), W_1(D_1))^* \cdot$ $(I_1, W_1)^*$ serve as initial inventories for the next period.

The current period is $t = 2$, the demand of period 2 is now realized and corresponds to the root node demand in a scenario tree to be constructed for periods $2, \dots, T$. The MSP model is solved and the root node decisions $(R_2(D_2), X_2(D_2),$ $W_2(D_2))^*$ are implemented. This process is repeated for $t = 3, \dots, T$. At the end of the planning horizon the values $(R_t(D_t), X_t(D_t), W_t(D_t))^*,$ for $t = 1, \dots, T$ constitute the MSP-DYN production plan.

6 Computational Experiments

In this section, we present a computational study where we compare the performance of the ZOIP, DYNIP, 2-SP, MSP, and MSP-DYN models considering various demand profiles and based on Fill Rate and Inventory Position. The former is a proxy for the level of customer service provided, while the latter serves as a proxy for the average inventory holding cost, considering both WIP and finished goods inventory levels. The models have been implemented in GAMS and solved using CPLEX 11.0. We begin by discussing the experimental design and then present the results and analysis.

Demand Profiles: Demand is forecasted over a horizon of three months, each consisting of four working weeks ($T = 12$ weeks). Demand in each period is independent and normally distributed. However, the means and variances of demand are allowed to vary across periods. Three possible levels of mean demand in a given month are considered: H (High $=$ 140), M (Medium $=$ 100), and L (Low $=$ 60). Based on these levels, seven demand profiles are constructed by considering different levels for each month (i.e., four week subinterval): LLL, MMM, HHH, LMH, HML, LHL, and HLH. For example, demand profile LMH represents an increasing monthly demand, where demand from week 1 to week 4 is 60, from week 5 to week 8 is 100, and from week 9 to week 12 is 140. These profiles show how the mean values of the demand distributions vary over the planning horizon. In all these profiles, we assume a constant coefficient of variation $\rho_t = \sigma_t/\mu_t = 0.25$ for the demand distributions in every period. This yields very small probability of negative demands; in the few cases in our experiments in which they arose, negative demands were set to zero.

In order to implement the stochastic programs 2-SP and M-SP, scenario trees based on the various demand profiles must be constructed. For the 2-SP model, three scenarios are considered, Low, Medium, and High, with demand in each period t is set to each of the values $\mu_t - \sigma_t$, μ_t, and $\mu_t + \sigma_t$, respectively. The probabilities of the three demand realizations are assumed to be 0.25, 0.5, and 0.25, respectively. This is clearly a limited representation of the demand uncertainty, and we shall return to this issue in our discussion of our computational results.

In the case of the M-SP model, successive stochastic programs (one in each period) have to be solved in order to obtain a production plan for the entire horizon. Therefore, in each period t a binary scenario tree starting from period t up to the end of the horizon is constructed. In each period we consider two possible demand realizations, Low Demand ($\mu_t - \sigma_t$) and High Demand ($\mu_t + \sigma_t$), with equal probabilities. Thus in any given period t, a M-SP is formulated and solved with a scenario tree containing $2^{T-t+1} - 1$ nodes and 2^{T-t} scenarios (number of leaf nodes).

The capacity of the production system is represented by a clearing function which captures the effect of congestion as discussed in Sect. 3. Following Karmarkar [61], we assume the form of the clearing function to be

$$f(W) = \frac{K_1 W}{K_2 + W},$$

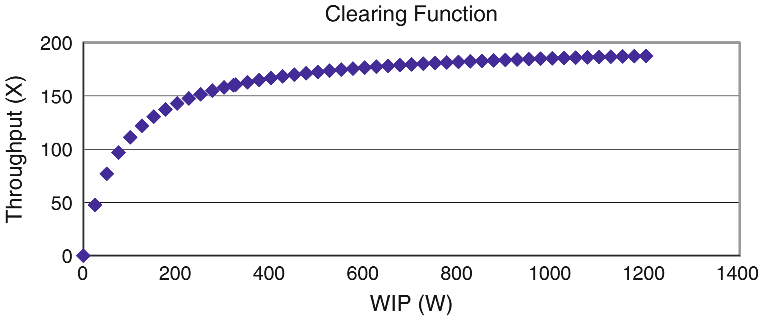

Fig. 2 Clearing function used in experiments

Table 4 Clearing function approximation

Segment	Intercept	Slope
1	0.0	0.5
2	136.0	0.069
3	154.8	0.036
4	161.8	0.023
5	180	0

where $K_1 = 200$ is the production capacity, and $K_2 = 80$ measures the curvature of the CF.

The resulting CF is shown in Fig. 2. Our piecewise linearization of this CF is given in Table 4.

There are clearly many specific issues involved in the estimation and piecewise linearization of CFs which are beyond the scope of this paper. These issues have been discussed extensively in Missbauer and Uzsoy [78]; specific approaches are illustrated in Asmundsson et al. [4], Missbauer [77], and Selcuk et al. [96], among others. Extensive experimentation in the course of this work has shown that the specific manner in which an appropriately fitted CF is piecewise linearized does not have much effect on the quality of the resulting production plans, although it does affect the estimates of the dual prices obtained for the associated constraints [62]. Since the primary purpose of this paper is to compare the solutions obtained from different formulations of the production planning problem with stochastic demand, all the models compared use the same piecewise linearized function. Hence the quality of the fit of the CF is not a factor in this study.

The values of L_t, i.e. the lead times in period $t = 1, .., T$ used in the formulations were chosen to be the same for all periods. This value, based on Little's Law, was chosen to be $L = W/\mu$, where μ is the average of all the demand means over the planning horizon and W the WIP value corresponding to a throughput of μ on the CF. This represents the behavior of a practitioner establishing a model based on historical data. The choice of values for I_0 and W_0 can be arbitrary, but the values we use are those recommended by Graves [38], setting $W_0 = L\mu$, and $I_0 = z_\alpha \sigma \sqrt{L}$.

To compare the performance of the production planning models, ZOIP, DYNIP, 2-SP, and M-SP (including the MSP and MSP-DYN strategies) we evaluate their optimal production plans in the face of simulated demand scenarios. For each demand profile, the evaluation procedure is as follows:

For ZOIP, 2-SP, and MSP

- Step 1 : Solve the four models for each demand profile and obtain the optimal values of the variables (R_t, X_t, W_t) for all periods to be specified at the beginning of the horizon before any actual demand has been observed, i.e. the first stage decision variables. These constitute the optimal production plan.
- Step 2 : Generate $N = 100$ demand scenarios from the normal distribution for each period and simulate the production plans for the models for each scenario. For each scenario a realization of inventories and backorders is obtained.
- Step 3 : Compute the performances for each model i.e., average and variance of backorders, fill rate, inventory position, and holding cost.

For DYNIP

- Step 1 : Solve the model for each demand profile and obtain the optimal values of the variable Y_t for all periods to be specified at the beginning of the horizon before any actual demand has been observed, i.e. the first stage decision variables.
- Step 2 : Generate $N = 100$ demand scenarios. For each scenario, once demand is realized in a given period, the corresponding (*R, X, W*) are determined and hence, the inventory and backlogs can be computed.
- Step 3 : Compute the performances for each model i.e., average and variance of backorders, fill rate, inventory position, and holding cost.

For MSP-DYN

- Step 1 : Generate 100 demand scenarios. For each scenario, once demand is realized in a given period *t*, solve a M-SP model for the periods $t, \ldots, T$ and implement the first period decisions, i.e., the (*R, X, W*) are determined as well as the ending inventory (*I*) and backlogs (*B*).
- Step 2 : Compute the performances for each model i.e., average and variance of backorders, fill rate, inventory position, and holding cost.

Since the chance constrained models ZOIP and DYNIP and the stochastic programing models 2SP, MSP and MSP-DYN use rather different modeling assumptions, care must be exercised when making comparisons. The chance constrained models assume a form for the demand distribution in each period, and do not consider shortage costs. However, it can be argued that an implicit judgement on the relative magnitude of holding and shortage costs is made in the specification of the required service level α, which also serves as the probability of constraint violation. The chance constrained models do not specify any particular recourse action when constraints are violated; in our computational experiments we assume any missed demands can be backlogged.

We thus consider three levels of the service level in our experiments: 90, 95 and 99.9%.

The stochastic programs, on the other hand, do not represent the demand distribution in a closed form. Instead, they use a discrete set of scenarios of outcomes to represent the uncertain nature of demand. Hence the effectiveness of these models is clearly linked to the number and degree of representativeness of the scenarios used to obtain the solutions. Another interesting issue is that stochastic programing models provide, by their nature, values for the decision variables corresponding to first stage decisions that must be made at the present time, as well as decision variables corresponding to each of the scenarios considered. However, since the scenarios considered in the model represent only a sample of possible realizations of the demand process, it is highly likely that in the future we will face a demand realization that does not match any of the scenarios used in obtaining them unless the stochastic program is solved on a rolling horizon basis. Since the size of the formulation to be solved for the stochastic programs is directly driven by the number of scenarios considered, this raises some interesting questions.

The performance of the stochastic programs (2-SP, MSP and MSP-DYN) is mainly affected by the magnitude of the backorder cost relative to the holding cost. We assume a unit production cost of $c = \$100$, and set the holding cost to $h = 0.2\,c$ and consider three levels for b the backorder cost: $0.5c$, c, and $4c$.

7 Results of Experiments

In order to facilitate a fair comparison between the different models, we have taken the approach of multiobjective optimization. The solution produced by any model represents a tradeoff between shortage and holding costs as that model perceives them, subject to the specific parameter settings used. The issue is further complicated by the different definitions of shortage that are possible. The chance constrained models require the specification of a maximum stockout probability. However, there is clearly a practical difference between a solution that stocks out by a large amount in one period, and one that stocks out by very small amounts in several.

We shall thus examine the issue in stages. We shall first consider the tradeoff between average inventory position, defined as the total finished goods and work in process inventory, and the fill rate, which is the fraction of demand in each period met from inventory. We shall then examine the difference between the planned and realized service levels in the chance constraint models, and also explore their sensitivity to errors in the estimation of the demand distributions used.

7.1 Inventory Position-Fill Rate Tradeoff

In order to examine the performance of the different models in terms of their tradeoff between inventory position and fill rate, we shall compute the scaled inventory

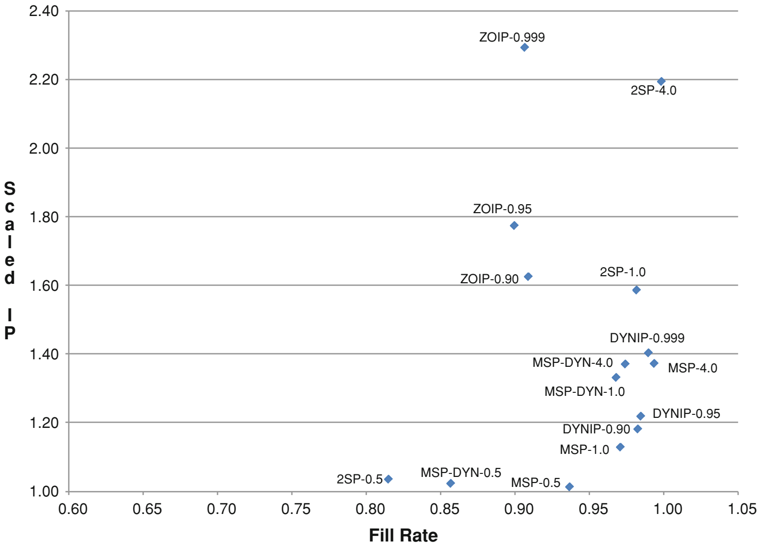

Fig. 3 Average performance of models over all demand configurations

position for each algorithm under each of our seven demand configurations. Let $IP(i,k)$ denote the average inventory position realized under model i under demand configuration k. Then we define the *Scaled* $IP(i,k) = IP(i,k)/\min_j\{IP(j,k)\}$. This quantity indicates the level of inventory position of a given model relative to the model with the lowest average inventory position obtained by any model for that demand configuration.

Figure 3 depicts the tradeoff between the models based on average performance across all demand configurations. The fill rate is plotted on the horizontal axis and the scaled inventory position on the vertical. Since we want fill rate to be high, and scaled IP to be low, the efficient frontier is to the bottom right of the plots.

Figure 3 yields a number of interesting insights. The ZOIP model is completely dominated, as we would expect. This is due to its complete lack of a recourse action, leaving it unable to react to the realized demand after it is observed. In particular, this leaves the model unable to react to demand that is lower than expected, causing it to overstock by a significant amount, as indicated by Ravindran et al. [93]. The two-stage stochastic program 2SP is also dominated. The efficient frontier is made up entirely of DYNIP and the static multistage model MSP, while the dynamic implementation of the M-SP, MSP-DYN, is also dominated.

Two salient features emerge from these results. The first and most encouraging from our perspective is the excellent performance of DYNIP. This model is highly competitive at service levels of 0.90 and 0.95, although it is dominated for a service level of 0.999. The relatively small difference in fill rate between the three service levels suggests that the model overstocks to some degree. There are two possible reasons for this behavior. One is that the assumptions of the chance constrained model are violated in the simulations we use, constituting an interesting direction for future work in understanding the sources of this behavior. Another possibility is that the lead time estimate used to set the safety stock levels is too high. The success of DYNIP over ZOIP is due to its incorporation of a dynamic recourse action—it can modify releases based on observed demand in the past, while ZOIP fixes all decisions at the start of the planning horizon; note that ZOIP and DYNIP use the same information about the demand process.

The comparison between DYNIP and MSP is more interesting. The results indicate that DYNIP obtains the same performance as MSP for a specific choice of service levels corresponding to a choice of parameters for MSP lying between $b = 100$ and $b = 400$. Given the very limited recourse action incorporated in DYNIP, this seems surprising at first sight; one would expect MSP to perform considerably better. However, we need to bear in mind that DYNIP is using a complete characterization of the demand distribution in each period, while MSP characterizes the demand uncertainty through the use of scenarios. Thus the number and choice of scenarios is critical for the MSP to obtain a good solution.

However, this is also where the size of the competing formulations needs to be taken into account. For a planning horizon of T periods, the DYNIP model requires $O(T)$ decision variables and constraints. Assuming two possible realizations for demand in each period as we do in this study, the scenario tree for MSP has $O(2^{T-1})$ nodes, implying that number of decision variables for what is a minimal amount of information on demand uncertainty. These results hold out the encouraging possibility that a minimal number of well-chosen scenarios may be sufficient for a stochastic program to make near-optimal decisions. However, the sheer size of the scenario trees required to model an industrial problem with multiple products, each with their own different demand processes, suggests that scaling conventional stochastic programing models up to solve industrial-sized problems poses substantial challenges.

Another interesting observation from Fig. 3 is the fact that the static MSP outperforms the dynamic version, MSP-DYN. The latter differs from the former in that the M-SP model is resolved at each period in the planning horizon, using the information from the realized demand in previous periods. Hence the recourse action taken at each period is to resolve the M-SP in the light of previously realized demand.

This result is particularly interesting since implementation on a rolling horizon or dynamic basis has been held up as a solution to the problem of uncertain demand in production planning for decades; the assumption is that only the decisions in the next period matter, and as long as we can revise decisions in the light of observed information we can obtain good results. However, some recent results suggest that our faith in this insight may be misplaced, at least under some circumstances. Orcun and Uzsoy [80] have shown that when the planning model does not accurately represent the behavior of the production system under study, rolling horizon implementations can result in undesirable oscillatory behavior similar to the nervousness discussed in the Material Requirements Planning (MRP) literature (e.g., [13]). What is striking in this case is that the extremely simple recourse action used in DYNIP yields just as good results as the far more sophisticated recourse action in MSP-DYN. This may well be due in part to the very limited demand information used in M-SP, as discussed above, which could potentially be remedied by including additional scenarios in the M-SP model. However, this would come at the cost of increasing the size of an already very large model. It is important to note that in the current experiments, the planning horizon T is fixed and does not recede into the future, which will cause ending effects to arise in decisions towards the end of the planning horizon. In particular, the limited planning horizon may cause the models to take decisions that are very good within the current horizon, but have very unfavorable consequences outside the current planning horizon. This issue clearly needs to be more carefully examined in future work.

7.2 Effect of Estimation Errors

In order to further explore the performance of DYNIP relative to MSP, we conducted two additional experiments in which the mean of the demand distribution used in the DYNIP models are perturbed by a random error uniformly distributed between 0 and 0.2 times the mean, representing a situation where demand is systematically overestimated. Our second case represents the case when demand is underestimated, represented by an error uniformly distributed between -0.2 and 0. The standard deviations are subjected to a random error uniformly distributed between -0.2 and 0.2. The purpose of this experiment is to examine the sensitivity of DYNIP to errors in the estimation of the demand distributions used.

The results of these experiments are shown in Fig. 4. The suffix "H" denotes the results for the case with overestimated demand, and "L" for the case with underestimated demand. The results for MSP are included for comparison. The results are quite intuitive. The impact of errors in demand estimation increases with the required service level. When $\alpha = 0.90$, the scaled IP varies between 1.12 and 1.21; for $\alpha = 0.95$, from 1.14 to 1.24; and for $\alpha = 0.999$, from 1.3 to 1.47. The changes in fill rate are all less than 1%. The MSP results are dominated except for MSP-4.0, which achieves a higher service level than DYNIP-0.999 and DYNIP-0.999-H with lower inventory position. These results together suggest that DYNIP is relatively robust to errors in demand estimation, while at the same time supporting the earlier evidence that it tends to overstock relative to the desired service level.

The tradeoff between fill rate and scaled IP for the individual demand configurations was also examined, although detailed results are not presented for brevity. Comparing the HHH, LLL and MMM results indicates that for LLL and MMM, DYNIP dominates MSP, while for HHH MSP enters the efficient frontier, obtaining slightly lower fill rates with substantially lower inventory position than MSP, although DYNIP-0.90 and DYNIP-0.95 remain on the efficient frontier. In all demand configurations except HML, DYNIP is represented in the efficient frontier; in that configuration MSP dominates all the DYNIP models, obtaining both higher fill rate and lower inventory position. Interestingly, the converse is true for the LHL configuration, where MSP is dominated by the DYNIP models.

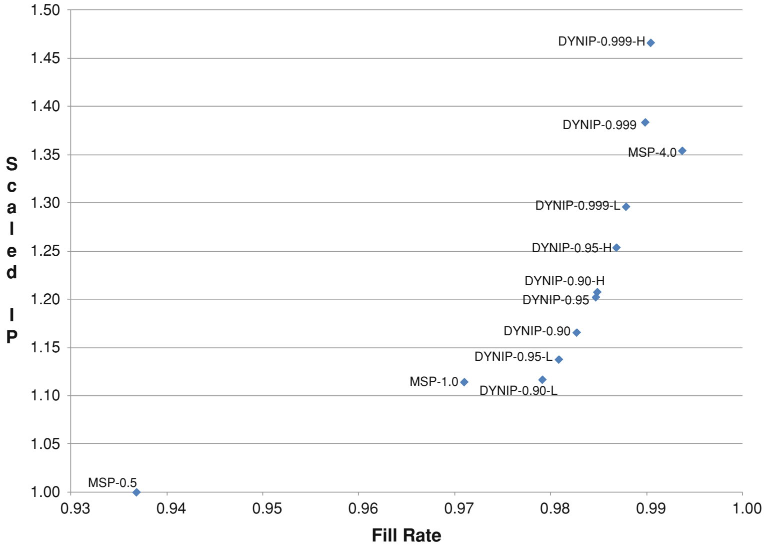

Fig. 4 Sensitivity of DYNIP to errors in demand estimation

Taken as a whole, these results suggest that DYNIP is at least a contender as a solution technique for the planning problem considered in this paper. While it exhibits some weaknesses in the face of high demand variability, its performance appears to be relatively robust to errors in the estimation of the demand distributions it uses, and it consistently achieves a position on the efficient frontier of the fill rate—inventory position tradeoff. It appears to have a tendency to overstock, which is likely due to the discrepancy between the assumptions of the model and the environment in which the simulations take place.

7.3 Service Level-Fill Rate Comparison

An interesting comparison that sheds some additional light on the behavior of the different models is to compare the average service levels and fill rates. The entries in Table 5 are computed by taking the average over all periods in each realization, and then taking the grand average of these over all realizations of a specific demand configuration. It is immediately apparent that the service levels realized by DYNIP are higher than the planned service levels, resulting in even higher fill rates. The reason for this behavior is very likely that the lead time being used to compute the inventory targets is higher than the average lead time that is realized in the simulations. Interestingly, even though the same lead time parameters are used in the ZOIP model, ZOIP's service level is markedly worse than that of DYNIP. ZOIP and DYNIP appear to perform better when the demand distribution is time-stationary

Table 5 Realized fill rates and service levels

		LLL	MMM	HHH	LMH	HML	LHL	HLH
2SP-0.5	SL	0.580	0.580	0.618	0.662	0.558	0.616	0.583
	FR	0.815	0.815	0.828	0.839	0.791	0.812	0.806
2SP-1.0	SL	0.935	0.944	0.948	0.954	0.967	0.966	0.948
	FR	0.978	0.979	0.979	0.978	0.991	0.990	0.978
2SP-4.0	SL	0.983	0.992	1.000	0.996	0.978	0.994	0.993
	FR	0.998	0.999	1.000	1.000	0.995	0.999	0.999
DYNIP-0.90	SL	1.000	0.992	0.958	0.950	0.833	0.983	0.892
	FR	1.000	1.000	0.993	0.987	0.931	0.999	0.969
DYNIP-0.95	SL	1.000	1.000	0.967	0.975	0.833	0.983	0.908
	FR	1.000	1.000	0.995	0.991	0.934	1.000	0.973
DYNIP-0.999	SL	1.000	1.000	0.983	0.983	0.875	1.000	0.925
	FR	1.000	1.000	0.998	0.999	0.952	1.000	0.980
MSP-0.5	SL	0.885	0.892	0.898	0.832	0.869	0.742	0.788
	FR	0.966	0.967	0.969	0.938	0.950	0.872	0.896
MSP-1.0	SL	0.917	0.926	0.951	0.883	0.903	0.884	0.884
	FR	0.979	0.980	0.985	0.965	0.971	0.957	0.960
MSP-4.0	SL	0.965	0.983	0.997	0.951	0.957	0.971	0.973
	FR	0.994	0.997	0.999	0.989	0.989	0.993	0.995
MSP-DYN-0.5	SL	0.846	0.846	0.713	0.667	0.775	0.633	0.658
	FR	0.939	0.937	0.879	0.837	0.897	0.728	0.782
MSP-DYN-1.0	SL	0.917	0.929	0.971	0.817	0.892	0.863	0.879
	FR	0.980	0.986	0.995	0.932	0.964	0.950	0.970
MSP-DYN-4.0	SL	0.917	0.929	0.983	0.842	0.887	0.896	0.921
	FR	0.980	0.986	0.996	0.937	0.965	0.970	0.986
ZOIP-0.90	SL	0.858	0.858	0.650	0.697	0.668	0.772	0.660
	FR	0.964	0.964	0.877	0.909	0.858	0.912	0.880
ZOIP-0.95	SL	0.801	0.737	0.655	0.713	0.682	0.795	0.666
	FR	0.925	0.910	0.881	0.915	0.863	0.921	0.883
ZOIP-0.999	SL	0.776	0.748	0.684	0.738	0.732	0.848	0.656
	FR	0.920	0.913	0.891	0.925	0.887	0.933	0.878

(demand configurations LLL, MMM, and HHH) than when it is not. In contrast, MSP maintains a consistent level of fill rate across all scenarios. The fact that the fill rate is consistently higher than the service level for the chance constrained models (ZOIP and DYNIP) suggests that even though stockouts occur, the amount of the stockout is quite modest in most cases.

8 Conclusions and Future Directions

Acknowledging at the outset the exploratory nature of this chapter, our results raise some interesting issues. Planning procedures with recourse (MSP, MSP-DYN and DYNIP) consistently outperform those without recourse (ZOIP and 2SP) as one would expect. However, the performance of DYNIP suggests that when appropriately parameterized it may be able to compete effectively, in terms of producing near-optimal solutions in reasonable CPU time, with far larger multi-stage stochastic programing models that employ a limited number of scenarios to capture demand uncertainty—at least under certain conditions. It also appears to be relatively robust to errors in estimation of the demand distribution used to construct the model. On the other hand, the MSP model appears to be able to produce good solutions with a minimal number of demand scenarios, considering only two possible values in each planning period. Even so, the MSP approach results in very large models relative to DYNIP. Finally, a dynamic, rolling horizon implementation of MSP yielded no apparent advantage over the static procedure. This finding is interesting in itself, since a rolling implementation is widely held to be the remedy for demand uncertainty.

Given the limited number of experiments carried out, these findings raise more questions than they answer, suggesting several directions for future work to clarify or confirm these findings. Clearly future work needs to focus on procedures with recourse, such as MSP and DYNIP. The reason why DYNIP appears to consistently overstock needs to be understood, and methods found to alleviate this issue if possible. It may be as simple as setting the lead time parameter used to compute the inventory targets more accurately, but it may also be related to the fact that the assumptions used in developing the model are violated in the experimental environment. If the latter is the case, careful mathematical analysis must be carried out to reveal the reason, and suggest an approach to correct the problem. The sensitivity of DYNIP to errors in estimating the demand distributions, and approaches for using it in the face of very limited demand information also need to be explored.

The MSP model used in this work highlights the primary issue with multistage stochastic programing when applied to production planning: the size of the scenario tree grows very rapidly, resulting in very large formulations even when a very limited number of different demand realizations are considered in each period. There needs to be a systematic investigation of how many scenarios need to be considered to provide a reasonably good solution (however that is to be defined, which is another complex issue), and possible solution methods that will allow a scaling up of these approaches to problems of industrial size.

Finally, as we have noted in our analysis, the chance constrained and stochastic programing models make quite different assumptions in formulating the models. The chance constrained models ignore shortage costs, and require a specified stockout probability. The stochastic programing models require a number of scenarios that describe the demand uncertainty and explicit holding and shortage costs. These different assumptions have been shown in the literature to lead to paradoxical behavior for the chance constrained models under certain circumstances, such as a negative value of the expected value of perfect information [14]. While the mathematical existence of such behavior is well documented, it may yet remain the case that chance constrained models, when appropriately formulated and parameterized, can provide effective heuristics for the problem of production planning under uncertain demand.

Acknowledgments The research of Reha Uzsoy was partially supported by the National Science Foundation under Grant No. CCM-1029706. The opinions in this paper are those of the authors, and do not necessarily reflect the position of NSF.

References

1. Agnew C (1976) Dynamic modeling and control of some congestion prone systems. Oper Res 24(3):400–419
2. Anli OM, Caramanis M, Paschalidis IC (2007) Tractable supply chain production planning modeling non-linear lead time and quality of service constraints. J Manuf Syst 26(2):116–134
3. Anupindi R, Morton TE, Pentico D (1996) The nonstationary stochastic lead-time inventory problem: near-myopic bounds, heuristics, and testing. Manag Sci 42(1):124–129
4. Asmundsson JM, Rardin RL, Turkseven CH, Uzsoy R (2009) Production planning models with resources subject to congestion. Naval Res Logist 56:142–157
5. Asmundsson JM, Rardin RL, Uzsoy R (2006) Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. IEEE Trans Semicond Manuf 19:95–111
6. Bang JY, Kim YD (2010) Hierarchical production planning for semiconductor wafer fabrication based on linear programming and discrete-event simulation. IEEE Trans Autom Sci Eng 7(2):326–336
7. Ben-Tal A, Nemirovski A (2000) Robust solutions of linear programming problems contaminated with uncertain data. Math Program 88(3):411–424
8. Bergstrom GL, Smith BE (1970) Multi-Item production planning—an extension of the Hmms rules. Manag Sci 16(10):B614–B629
9. Bertsimas D, Mourtzinou G (1997) Transient laws of non-stationary queueing systems and their applications. Queueing Syst 25:115–155
10. Bertsimas D, Thiele A (2006) A robust optimization approach to inventory theory. Oper Res 54(1):150–168
11. Birge JR (1985) Decomposition and partitioning methods for multistage stochastic linear programs. Oper Res 33(5):989–1007
12. Birge JR, Louveaux F (1997) Introduction to stochastic programming. Springer, New York
13. Blackburn JD, Kropp DH, Millen RA (1986) A comparison of strategies to dampen nervousness in Mrp systems. Manag Sci 32(4):412–439
14. Blau RA (1974) Stochastic programming and decision analysis: an apparent dilemma. Manag Sci 21(3):271–276
15. Bookbinder JH, Tan JY (1988) Strategies for the probabilistic lot sizing problem with service level constraints. Manag Sci 34(9):1096–1108
16. Buffa ES, Taubert WH (1972) Production-inventory systems; planning and control. R.D. Irwin, Homewood Ill
17. Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, Englewood Cliffs
18. Byrne MD, Bakir MA (1999) Production planning using a hybrid simulation-analytical approach. Int J Prod Econ 59:305–311
19. Byrne MD, Hossain MM (2005) Production planning: an improved hybrid approach. Int J Prod Econ 93–94:225–229
20. Charnes A, Cooper WW (1959) Chance-constrained programming. Manag Sci 6(1):73–79

21. Charnes A, Cooper WW (1963) Deterministic equivalents for optimizing and satisficing under chance constraints. Oper Res 11:18–39
22. Charnes A, Cooper WW, Symonds GH (1958) Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. Manag Sci 4(3):235–263
23. Ciarallo FW, Akella R, Morton TE (1994) A periodic review, production planning-model with uncertain capacity and uncertain demand—optimality of extended myopic policies. Manag Sci 40(3):320–332
24. Clark AJ, Scarf H (1960) Optimal policies for a multi-echelon inventory problem. Manag Sci 6(4):475–490
25. Dantzig GB, Wolfe P (1960) Decomposition principle for linear programs. Oper Res 8(1): 101–111
26. Dauzere-Peres S, Lasserre JB (1994) An integrated approach in production planning and scheduling. Springer, Berlin
27. de Kok AG, Fransoo JC (2003) Planning supply chain operations: definition and comparison of planning concepts. In: Kok AG, Graves SC (eds) Or handbook on supply chain management, Elsevier, Amsterdam, pp 597–675
28. Deckro RF, Hebert JE (1984) Goal programming approaches to solving linear decision rule based aggregate production planning-models. IIE Trans 16(4):308–315
29. Eppen G, Martin RK (1988) Determining safety stock in the presence of stochastic lead times. Manag Sci 34:1380–1390
30. Escudero LF, Kamesan PV, King AJ, Wets JB (1993) Production planning via scenario modelling. Ann Oper Res 43:311–335
31. Ettl M, Feigin G, Lin GY, Yao DD (2000) A supply chain network model with base-stock control and service requirements. Oper Res 48:216–232
32. Federgruen A, Zipkin P (1986) An inventory model with limited production capacity and uncertain demands I: the average cost criterion. Math Oper Res 11(2):193–207
33. Federgruen A, Zipkin P (1986) An inventory model with limited production capacity and uncertain demands II: the discounted cost criterion. Math Oper Res 11(2):208–215
34. Garstka SJ, Wets RJB (1974) On decision rules in stochastic programming. Math Program 7(2):117–143
35. Gassmann HI (1990) Mslips: a computer code for the multistage stochastic linear programming problem. Math Program 47:407–423
36. Goodman DA (1974) Goal programming approach to aggregate planning of production and work force. Manag Sci Ser B Appl 20(12):1569–1575
37. Graves SC (1986) A tactical planning model for a job shop. Oper Res 34:552–533
38. Graves SC (1988) Safety stocks in manufacturing systems. J Manuf Oper Manag 1:67–101
39. Grubbstrom RW (1998) A net present value approach to safety stocks in planned production. Int J Prod Econ 56(57):213–229
40. Gupta A, Maranas CD (2003) Managing demand ncertainty in supply chain planning. Comput Chem Eng 27(8–9):1219–1227
41. Hackman S (2008) Production economics. Springer, Berlin
42. Hackman ST, Leachman RC (1989) A general framework for modeling production. Manag Sci 35:478–495
43. Hadley G, Whitin TM (1963) Analysis of inventory systems. Prentice-Hall, Englewood Cliffs
44. Hanssmann F, Hess SW (1960) A linear programming approach to production and employment scheduling. Manag Technol 1(1):46–51
45. Hax AC, Candea D (1984) Production and inventory management. Prentice-Hall, Englewood Cliffs
46. Heyman DP, Sobel MJ (1982) Stochastic models in operations research. McGraw-Hill, New York
47. Heyman DP, Sobel MJ (1990) Stochastic models. Elsevier Science Publishing Co., New York

48. Higle JL, Kempf KG (2010) Production planning under supply and demand uncertainty: a stochastic programming approach: stochastic programming: the state of the art. G. infanger. Springer, Berlin
49. Holt CC, Modigliani F, Muth JF (1956) Derivation of a linear rule for production and employment. Manag Sci 2(2):159–177
50. Holt CC, Modigliani F, Muth JF, Simon HA (1960) Planning production, inventories and work force. Prentice Hall, Englewood Cliffs
51. Holt CC, Modigliani F, Simon HA (1955) A linear decision rule for production and employment scheduling. Manag Sci 2(1):1–30
52. Hopp WJ, Spearman ML (2001) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston
53. Hung YF, Leachman RC (1996) A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. IEEE Trans Semicond Manuf 9(2):257–269
54. Iglehart DL, Karlin S (1962) Optimal policy for dynamic inventory process with nonstationary stochastic demands. Stanford University Press, Stanford Calif, pp 127–147
55. Irdem DF, Kacar NB, Uzsoy R (2010) An exploratory analysis of two iterative linear programming-simulation approaches for production planning. IEEE Trans Semicond Manuf 23:442–455
56. Johnson LA, Montgomery DC (1974) Operations research in production planning, scheduling and inventory control. Wiley, New York
57. Kacar NB, Irdem DF, Uzsoy R (2010) An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. Research Report, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University
58. Kall P, Wallace SW (1994) Stochastic programming. Wiley, Chichester
59. Karlin S (1960) Dynamic inventory policy with varying stochastic demands. Manag Sci 6(3):231–258
60. Karmarkar US (1987) Lot sizes, lead times and in-process inventories. Manag Sci 33(3): 409–418
61. Karmarkar US (1989) Capacity loading and release planning with Work-in-Progress (WIP) and lead-times. J Manuf Oper Manag 2:105–123
62. Kefeli A, Uzsoy R, Fathi Y, Kay M (2011) Using a mathematical programming model to examine the marginal price of capacitated resources. Int J Prod Econ 131(1):383–391
63. Kekre S (1984) Some issues in job shop design. University of Rochester, Rochester NY
64. Kempf KG (2004) Control-oriented approaches to supply chain management in semiconductor manufacturing. In: Proceedings of the American control conference, Boston, MA, United States
65. Kempf KG, Keskinocak P, Uzsoy R (2010) Preface. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extended enterprise: a state of the art handbook, Springer, Amsterdam, pp 1–20
66. Kim B, Kim S (2001) Extended model for a hybrid production planning approach. Int J Prod Econ 73:165–173
67. Kleindorfer PR, Kriebel CH, Thompson GL, Kleindorfer GB (1975) Discrete optimal control of production plans. Manag Sci 22(3):261–273
68. Lasdon LS (1970) Optimization theory for large systems. Macmillan, New York
69. Lautenschläger M, Stadtler H (1998) Modelling lead times depending on capacity utilization. Research Report, Technische Universitat Darmstadt
70. Leung SCH, Wu Y (2004) A robust optimization model for stochastic aggregate production planning. Prod Planning Control 15(5):502–514
71. Liu L, Liu X, Yao DD (2004) Analysis and optimization of multi-stage inventory queues. Manag Sci 50:365–380

72. Louveaux F (1980) A solution method for multistage stochastc programs with recourse with application to an energy investment problem. Oper Res 28(4):889–902
73. Meal H (1979) Safety stocks in Mrp systems. Operations Research Center, Massachusetts Institute of Technology, Cambridge MA
74. Medhi J (1991) Stochastic models in queuing theory. Academic Press, Boston
75. Miller JG (1979) Hedging the master schedule. Dissagregation problems in manufacturing and service organizations. LP Ritzman, Martinus Nijhoff, Boston MA
76. Missbauer H (2002) Aggregate order release planning for time-varying demand. Int J Prod Res 40:688–718
77. Missbauer H (2011) Order release planning with clearing functions: a queueing-theoretical analysis of the clearing function concept. Int J Prod Econ 131(1):399–406
78. Missbauer H, Uzsoy R (2010) Optimization models for production planning. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extended enterprise: a state of the art handbook, Springer, New York, pp 437–508
79. Orcun S, Kempf KG, Uzsoy R (2009) An integrated production planning model with load-dependent lead times and safety stocks. Comput Chem Eng 32:2159–2136
80. Orcun S, Uzsoy R (2011) The effects of production planning on the dynamic behavior of a simple supply chain: an experimental study. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning in the extended enterprise: a state of the art handbook, Springer, Berlin, pp 43–80
81. Orcun S, Uzsoy R, Kempf KG (2006) Using system dynamics simulations to compare capacity models for production planning. Winter Simulation Conference, Monterey
82. Orlicky J (1975) Material requirements planning: the new way of life in production and inventory management. McGraw-Hill, New York
83. Pahl J, Voss S, Woodruff DL (2005) Production planning with load dependent lead times. 4OR Q J Oper Res 3:257–302
84. Parlar M (1985) A stochastic production planning model with a dynamic chance constraint. Eur J Oper Res 20(2):255–260
85. Peters RJ, Boskma K, Kupper HAE (1977) Stochastic programming in production planning: a case with non-simple recourse. Statistica Neerlandica 31:113–126
86. Powell WB, Jaillet P, Odoni A (1995) Stochastic and dynamic networks and routing. In: Ball M, Magnanti T, Monma C (eds) Handbooks in operations research and the management sciences. Amsterdam, Elsevier, pp 141–295
87. Prékopa A (1995) Stochastic programming. Kluwer Academic Publishers, Boston
88. Pritsker AAB, Snyder K (1997) Production scheduling using factor. In: Artiba A, Elmaghraby SE (eds) The planning and scheduling of production systems. Chapman and Hall
89. Puterman ML (2005) Markov decision processes: discrete stochastic dynamic programming. Wiley, New York
90. Raa B, Aghezzaf EH (2005) A robust dynamic planning strategy for lot-sizing problems with stochastic demands. J Intell Manuf 16(2):207–213
91. Rao SS, Gunasekaran A, Goyal SK, Martikainen T (1998) Waiting line model applications in manufacturing. Int J Prod Econ 54(1):1–28
92. Rardin RL, Uzsoy R (2001) Experimental evaluation of heuristic optimization algorithms: a tutorial. J Heuristics 7:261–304
93. Ravindran A, Kempf KG, Uzsoy R (2008) Dynamic base stock models for production-inventory systems with nonstationary demand. Research Report, Edward P. Fitts Department of Industrial and Systems Engineering, North carolina State University
94. Ravindran A, Kempf KG, Uzsoy R (2011) Production planning with load-dependent lead times and safety stocks. Int J Plan Sched 1(1–2):58–89
95. Riaño G (2003) Transient behavior of stochastic networks: application to production planning with load-dependent lead times. School of industrial and systems engineering. Georgia Institute of Technology, Atlanta GA
96. Selçuk B, Fransoo JC, de Kok AG (2007) Work in process clearing in supply chain operations planning. IIE Trans 40:206–220

97. Spearman ML (1991) An Analytic congestion model for closed production systems with Ifr processing times. Manag Sci 37(8):1015–1029
98. Srinivasan A, Carey M, Morton TE (1988) Resource pricing and aggregate scheduling in manufacturing systems. Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh
99. Tayur SR (1993) Computing the optimal policy for capacitated inventory models. Commun Stat Stoch Models 9(4):585–598
100. Van Slyke RM, Wets JB (1969) L-shaped linear programs with applications to optimal control and stochastic programming. SiAM J Appl Math 17(4):638–663
101. Veinott AF (1965) Optimal policy for a multi-product, dynamic, nonstationary inventory problem. Manag Sci 12(3):206–222
102. Veinott AF (1965) Optimal policy in a dynamic single product nonstationary inventory model with several demand classes. Oper Res 13(5):761–778
103. Voss S, Woodruff DL (2003) Introduction to computational optimization models for production planning in a supply chain. Springer, Berlin
104. Whitt W (1991) A Review of $L = \Lambda w$ and Extensions. Queueing Syst 9:235–268
105. Wittrock RJ (1983) Advances in a nested decomposition algorithm for solving staircase linear programs. Technical Report SOL-83-2, Systems Optimization Laboratory
106. Zipkin PH (1986) Models for design and control of stochastic, multi-item batch production systems. Oper Res 34(1):91–104
107. Zipkin PH (2000) Foundations of inventory management. Irwin, Burr Ridge IL