Mathematical Methods of Operations Research manuscript No.
(will be inserted by the editor)

The Stochastic Guaranteed Service Model with Recourse for Multi-Echelon Warehouse Management

Jörg Rambau · **Konrad Schade**

Received: date / Accepted: date

Abstract The Guaranteed Service Model (GSM) computes optimal order-points in multi-echelon inventory control under the assumptions that delivery times can be guaranteed and the demand is bounded. Our new Stochastic Guaranteed Service Model (SGSM) with Recourse covers also scenarios that violate these assumptions. Simulation experiments on real-world data of a large German car manufacturer show that policies based on the SGSM dominate GSM-policies.

Keywords Multi-echelon inventory control · guaranteed service model · stochastic programming · integer linear programming · real-world application

Mathematics Subject Classification (2000) MSC 90B05 · MSC 90C10

1 Introduction

We investigate the multi-echelon inventory control problem of a large German automobile manufacturer. The core of inventory control is to balance cost with service quality. Two main classes of mathematical models have been established in the literature: *Stochastic service models (SSM)* explicitly take into account the distribution in lead times and demands and account for all actions that have to be made to fulfill demands. *Guaranteed Service Models (GSM)* assume some *operational flexibility* outside the model that is not accounted for, and the model itself works with bounded

Supported by a grant of "Elitenetzwerk Bayern".

Jörg Rambau
Universität Bayreuth
Tel.: +49 921 55-7350
Fax: +49 921 55-7352
E-mail: joerg.rambau@uni-bayreuth.de

Konrad Schade
Volkswagen AG
E-mail: konradschade@googlemail.com

demands and bounded lead times instead. The GSM paradigm is motivated by the fact in most companies individuals have the capability to react to unforeseen events successfully in many ways, and no model can possibly capture all these reactions faithfully.

Research on the GSM was initiated by Simpson [17]. In [5] and [12] the investigations were extended to tree structured networks and acyclic networks, respectively. The first application of the GSM to spare parts distribution networks was carried out in [11] for the spare parts distribution system of a large German car manufacturer. Earlier investigations in other application contexts can be found, e.g., in [9, 10, 13].

The GSM is kind of an indirect model whose decision variables are service times that each warehouse in the system (called a *node*) guarantees to its successor nodes. Thus, in a sense, it computes decisions in the space of event times. These guaranteed service times have to be transformed into decisions on the time at which a replenishment of a certain quantity has to be ordered at each node. Whenever one of the wide-spread (but in general suboptimal) base-stock policies (basically: order up to an stock level S when the stock falls below s) is used, this is easy: Via a bounded-demand assumption, the guaranteed service times can be transformed into minimum stock level requirements at the nodes (see, e.g., [5] and [12]).

In [2] base-stock policies are theoretically justified by proving their optimality in elementary special cases. In many environments, (s,S) policies or the like are prescribed despite their suboptimality because the mechanism is easy to understand. Then, only the parameters s and S can be chosen. In such environments, the GSM can be used to determine values for s and S.

Finally, in [12] it has been shown, how the GSM including a determination of order points can (approximately) be solved by a mixed integer linear program (MILP). This gives the modeler the opportunity to easily add business constraints in the space of event-times to the GSM without affecting its solvability too much. This model was used in [11] for the spare parts distribution network that is also considered in this paper.

The main drawback of the GSM is that it cannot keep control of the usage of operational flexibility. The problem is that employees can use operational flexibility (even at no cost) but not beyond a certain capacity. Operational flexibility is used in the GSM to guarantee bounded lead times and bounded demands. That is, deliveries can be expedited and/or outsourced in emergency cases. Assuming a known joint distribution for the lead times and demands, one can keep control of the amount of operational flexibility by prescribing a *target service level* at the nodes: for example, if we want that during at least 90 % of the time we can deliver on time and enough without using operational flexibility, then we can set up the GSM with the 90 % quantile of the lead time/demand distribution as the bounded lead time and demands. Such target service levels are meaningful at the nodes delivering to end customers (demand nodes). However, prescribed target service levels at internal nodes are hard to justify.

The questions that remain are therefore: How should one decide on the internal target service levels? Should there be at all individual target service levels for the nodes? Do individual quantile-based lead time and demand bounds in a GSM really guarantee that the target service level is achieved in the demand nodes? In a sense, the core task of inventory control (balance cost with service quality) is shifted to the formal choice of the target service levels whose side-effects and correct real-world interpretations are not always straight-forward.

One way to avoid the weaknesses of the GSM is to use a model from the SSM category instead (see, e.g., [16] for the METRIC system, [3] for a survey, and [4] for a special version of a stochastic service model). However, in SSM adding further restrictions, e.g., imposed by the business processes of a company, can render a method impractical. This is because many algorithms are based on specialized dynamic programming algorithms that may fail to apply to a system with additional restrictions. In contrast to this, adding restrictions to the ILP model of the GSM to a certain extent does not affect the ILP solution procedure.

To keep the advantages and overcome some of the weaknesses of the GSM, we have introduced the *stochastic guaranteed service model with recourse (SGSM)* in [14] and applied the first basic version of it to the inventory control problem in a multi-echelon warehouse system of a spare part distributor. The SGSM adds a lead time and demand sampling component and a recourse component to the GSM. (See [1] for background on stochastic programming with recourse.) This way, each lead time and demand scenario that is covered by the sampling component is accounted for inside the model; in the basic version, generic operational flexibility allowing for smaller safety stock levels leads to additional recourse costs. The SGSM does not need any prescribed service level requirements; it yields service levels as a result of the computation.[1] However, estimates for the recourse costs have to be given for all scenarios where lead times and demands can only be handled with operational flexibility. Since we need not prescribe the service level requirements, the core task of inventory control – balancing cost and service quality – is done inside the model.

In this paper, we go beyond the conference presentation [14] in the following aspects (among others):

- We introduce the new SGSM with a non-trivial complete recourse consisting of a transportation option besides the penalty cost for unsatisfied demand, i.e., requested parts that cannot be delivered in time.
- We solve the SGSM by a combination of sample average approximation with state-of-the art scenario reduction techniques. This way, a better coverage of unlikely but expensive scenarios is achieved without increasing the computation times in the MILP solver. Our new asymmetric distance function for the asymmetric scenario reduction takes into account the influence of the scenario reduction on the result of the optimization. To the best of our knowledge, this is new.
- We present a more comprehensive documentation of extended computational results, including a new comparison to one representative [4] of the class of stochastic service models that could be implemented to cope with our test data.

In simulations on real data we observe that the new asymmetric scenario reduction technique is able to improve the approximation quality of the SGSM by a large margin. Moreover, the SGSM decreases the inventory holding cost and the recourse costs at the same time compared to the GSM.

[1] In fact, service level requirements can be prescribed also in the SGSM by using the original GSM demand constraints for the nodes for which this is desired, see Section 2.

It would be interesting from a theoretical point of view to also check performances on artificial randomized data. For this work, we focused on the practical impact in real-world applications, for which randomized data is rarely representative. We emphasize that, for this reason, our simulation experiments are completely independent of the assumptions of the tested models – it rather represents our partner's process as closely as possible.

In the following section we introduce the modeling of the GSM and the SGSM before we theoretically compare them in section 3. We develop methods for scenario generation and scenario reduction in section 4. After the description of the simulation method and some computational results in section 5 we end with our conclusions.

2 Modeling

In this section we first give an introduction to the GSM. We use the ILP modeling approach as in [12]. Then we present the SGSM in two different ways. First, in 2.2 we introduce the SGSM as a two stage stochastic mixed-integer linear program with simple recourse. Second, in 2.3 we show an extension where the recourse action of the locations supplying the end customers are modeled as a transportation problem. We present all models only for diverging networks, i.e., networks in which all nodes have unique predecessors.

2.1 The Guaranteed-Service-Model

In order to make the paper self-contained, we repeat in this section the known MILP formulation for the GSM as can be found in the original work in [12]. Since in spare-part systems, there are large expensive parts stored at small stock levels, the order points are required to be integral. Notational conventions are taken from [14].

Parameters of the model GSM are:

G	directed graph describing the warehouse network
N	number of warehouses
$N(G)$	set of nodes i in G
$A(G)$	set of arcs ij in G
$D(G)$	set of leaves in G (warehouses delivering to end-customers)
h_i	inventory holding cost in location i; positive
L_i	lead time to location i
$\bar{s}_i^{\text{out}}$	given service time for a leaf $i \in D(G)$
$\Phi_i(x_i)$	upper bound for the demand in $i \in N(G)$ during the time period x_i

Variables of the model GSM are the following for nodes $i \in N(G)$:

s_i^{in}	service times guaranteed by the predecessors of i

s_i^{out} service times guaranteed by i for its successors
x_i time period that i needs to bridge with its inventory
(i.e., the time between order and delivery
of replenishments from the predecessors of i)
y_i order-point in i

The model GSM now reads as follows:

$$\min \sum_{i \in N(G)} h_i y_i \tag{1}$$

$$s_i^{\text{out}} \leq \bar{s}_i^{\text{out}} \qquad \forall i \in D(G) \tag{2}$$

$$s_i^{\text{in}} \geq s_j^{\text{out}} \qquad \forall ji \in A(G) \tag{3}$$

$$x_i \geq s_i^{\text{in}} - s_i^{\text{out}} + L_i \qquad \forall i \in N(G) \tag{4}$$

$$y_i \geq \Phi_i(x_i) \qquad \forall i \in N(G) \tag{5}$$

$$s_i^{\text{in}},\ s_i^{\text{out}}, x_i,\ y_i \geq 0 \qquad \forall i \in N(G) \tag{6}$$

$$y_i \in \mathbb{Z} \qquad \forall i \in N(G) \tag{7}$$

With the well-known multiple-choice modeling of piecewise linear functions, this non-linear model can approximately be transformed into an MILP (see [12]).

Note that the GSM is a one-stage model although it deals with a multi-stage decision problem. In a sense, the GSM computes decisions in the space of event times: how long does it take (at most) after an order at node i has been placed by node j until the delivery of node i arrives at node j. These one-stage decisions in the space of event times imply the time x_i that has to be bridged with inventory until we can be sure to receive a replenishment. The necessary safety stocks y_i at all nodes i can then be computed from x_i using the maximal demand quantity $\Phi_i(x_i)$ during time x_i. These decisions can be considered stationary over time. They are transformed by a base-stock policy into the multi-stage sequence of decisions for every node, namely, how much to order given the current stock level.

2.2 The Stochastic Guaranteed-Service-Model with Simple Recourse

In this section, we present the simple-recourse version of the SGSM, following [14]. The short-comings of the GSM are addressed by turning the deterministic GSM into a stochastic model with recourse. Again, the service times are the decision variables.

We first fix our notation for the stochastic data. A *lead time/demand distribution* consists of the following:

- A finite sample set $\Omega = \{1, 2, \ldots, |\Omega|\}$ of states ω of the world, encoded by natural numbers.
- A positive probability for each ω, denoted by $p^{\omega} > 0$, inducing a probability measure P_{Ω} on 2^{Ω} via $P(A) = \sum_{\omega \in A} p^{\omega}$.

- For each $\omega \in \Omega$, a random vector of measurements ξ^ω that has the following components:
$$\xi^\omega = \left(L_i^\omega, \Psi_i^\omega\right)_{i \in N(G)}, \tag{8}$$
where $L_i^\omega \leq 0$ is the lead time in node i in state ω and Ψ_i^ω is a function that assigns to each time interval x a demand $\Psi_i^\omega(x)$, which denotes the demand presented to node i in state-of-the-world ω.

We call ξ^ω the *lead time/demand scenario*, or *scenario* for short, of state $\omega \in \Omega$. The probability that $\xi^\omega = \xi$ is equal to $\sum_{\omega \in \Omega : \xi^\omega = \xi} p^\omega$. The induced probability measure on 2^Ξ is denoted by P_Ξ. The set of all scenarios with positive probabilities is denoted by Ξ. The complete lead time/demand rate distribution is denoted by (Ξ, p).

We are given a finite lead time/demand distribution (Ξ, p). Whenever a node misses its "guaranteed" service time, then a recourse action has to be taken that delivers the part faster (expediting). Whenever in some state ω of the world a node cannot deliver any of the demanded parts, a recourse action has to be taken that gets the missing parts from somewhere else (outsourcing). In contrast to the GSM, where such actions can be taken tacitly at no cost, in the SGSM operational flexibility is made explicit in the model such that we can assign a cost to it. The remaining approximation of the SGSM is that expediting and outsourcing can always be done by paying a surplus, which leads to a model with simple complete recourse. In the following, we review the formalization of the SGSM from [14].

Scenario and recourse parameters of the SGSM:

Ω	index set of scenarios
p^ω	probability of scenario $\omega \in \Omega$
t_i	cost to compensate for one time unit of late delivery, non-negative
c_i	cost to compensate for one piece of unmet demand, non-negative
L_i^ω	actual lead time to i in scenario ω
$\Psi_i^\omega(x)$	actual demand in i, during time period x in scenario ω

Additional recourse variables of the SGSM:

r_i^ω	recourse variable for missed deadlines in scenario ω; "how many time units should be compensated at a cost of t_i per unit?"
q_i^ω	recourse variable for missing pieces in scenario ω; "how many pieces should be compensated at a cost of t_i per unit?"

The SGSM is modeled by the following MILP:

$$\min \sum_{i \in N(G)} \left(h_i y_i + \sum_{\omega \in \omega} p^\omega \left(t_i r_i^\omega + c_i q_i^\omega\right)\right) \tag{9}$$

$$s_i^{\text{out}} \leq \bar{s}_i^{\text{out}} \qquad \forall i \in D(G) \tag{10}$$

$$s_i^{\text{in}} \geq s_j^{\text{out}} \qquad \forall ji \in A(G) \tag{11}$$

$$x_i + r_i^{\omega} \geq s_i^{\text{in}} - s_i^{\text{out}} + L_i^{\omega} \qquad \forall i \in N(G),\ \forall \omega \in \Omega \tag{12}$$

$$y_i + q_i^{\omega} \geq \Psi_i^{\omega}(x_i) \qquad \forall i \in N(G),\ \forall \omega \in \Omega \tag{13}$$

$$x_i,\ s_i^{\text{in}},\ s_i^{\text{out}},\ r_i^{\omega},\ q_i^{\omega} \geq 0 \qquad \forall i \in N(G),\ \forall \omega \in \Omega \tag{14}$$

$$y_i,\ q_i^{\omega} \in \mathbb{Z} \qquad \forall i \in N(G),\ \forall \omega \in \Omega \tag{15}$$

Here, constraint (12) makes sure that whenever a guaranteed service time cannot be met, we have to buy r_i^{ω} time units of expediting services. Restriction (11) makes sure that no node can expect to receive a piece faster than its predecessor guarantees. Condition (10) indicates that each node delivering to end customers fulfills the service time requirements of the end customer and finally in (13) the reorder point y_i plus the outsourced quantities need to be higher than the demand during time x_i at node i, as given by $\Psi_i^{\omega}(x_i)$.

Remark 1 If one wants to prescribe a service level requirement in a node, one has to replace constraints (12) and (13) by the corresponding GSM constraint with the quantile-based demand function Φ_i and without recourse variables instead of the scenario-based demand function Ψ_i^{ω} with recourse variables.

Again, a linearization of $\Psi_i^{\omega}(x_i)$ can be carried out by the multiple-choice modeling of piecewise linear functions.

Note that the SGSM is now a two-stage model although we are dealing with a multi-stage decision problem. We are still working in the space of event times and obtain safety stock levels y_i for all nodes i. All decisions coming out of the SGSM are considered stationary. However, now we can detect extremal lead time and demand scenarios for which we need to apply a (simple but costly) recourse action.

In the variant of the SGSM with simple recourse, the second stage is equivalent to paying a penalty for missing parts and missed deadlines. The model in the following section contains a non-simple second stage.

2.3 Extension with External Suppliers and Lost Sales

The model with simple recourse from the previous section can be extended by modeling an explicit recourse process. We assume that unmet customer demands are lost. However, internal orders are backlogged. The locations that deliver parts to the end customers can order parts from external suppliers to prevent lost sales.

The external suppliers deliver the parts directly to the end customers such that there is no delay in the delivery. The costs of an order from an external supplier depends on the distance between the ordering location and the supplier. Of course, the suppliers do not have unlimited stock such that capacity constraints have to be taken into account. To concentrate on these recourse actions we assume that the lead times in the system are fixed. An extension with lead time uncertainties would be straight-forward.

We need some more notation to model the new situation

J	set of external suppliers

C_j capacity of the external supplier j

q_{ji}^{ω} recourse variable for the number of parts that location i orders at supplier j

c_{ji} costs for location i to order one part from supplier j

This leads us to the following model:

$$\min \sum_{i \in N(G)} \big(h_i y_i + \sum_{\omega \in \Omega} p^{\omega} \sum_{j \in J} c_{ji} q_{ji}^{\omega} \big) \tag{16}$$

$$s_i^{\text{out}} \le \bar{s}_i^{\text{out}} \qquad \forall i \in D(G) \tag{17}$$

$$s_i^{\text{in}} \ge s_j^{\text{out}} \qquad \forall ji \in A(G) \tag{18}$$

$$x_i \ge s_i^{\text{in}} - s_i^{\text{out}} + L_i \qquad \forall i \in N(G) \tag{19}$$

$$y_i + \sum_{j \in J} q_{ji}^{\omega} \ge \Psi_i^{\omega}(x_i) \qquad \forall i \in D(G),\ \forall \omega \in \Omega \tag{20}$$

$$\sum_{i \in D(G)} q_{ji}^{\omega} \le C_j \qquad \forall j \in J,\ \forall \omega \in \Omega \tag{21}$$

$$y_i \ge \Psi_i^{\omega}(x_i) \qquad \forall i \in N(G) \backslash D(G),\ \forall \omega \in \Omega \tag{22}$$

$$x_i,\ s_i^{\text{in}},\ s_i^{\text{out}},\ q_{ji}^{\omega} \ge 0 \qquad \forall i \in N(G),\ \forall j \in J,\ \forall \omega \in \Omega \tag{23}$$

$$y_i,\ q_{ji}^{\omega} \in \mathbb{Z} \qquad \forall i \in N(G),\ \forall \omega \in \Omega \tag{24}$$

The constraints (20) and (22) are forcing the reorder points y_i to be higher than the demand during the time that has to be covered. For locations with end customer demand there is the possibility to procure parts from external suppliers. As the suppliers do not have infinity capacity, constraint (21) must hold.

So far, this model does not have complete recourse. Therefore, we introduce another recourse variable. As before, we enable for every location the possibility to pay a penalty for a lost sale if it cannot deliver the ordered parts. For instance one can provide the customer with a replacement vehicle until the spare part can be delivered and the customer's car is fixed. The corresponding penalty recourse variable is denoted by q_i^{ω}, as in the first model, and the penalty costs are denoted by c_i again.

We obtain a two stage stochastic model with complete recourse:

$$\min \sum_{i \in N(G)} \Big(h_i y_i + \sum_{\omega \in \Omega} p^{\omega} \big(c_i q_i^{\omega} + \sum_{j \in J} c_{ji} q_{ji}^{\omega} \big) \Big) \tag{25}$$

$$s_i^{\text{out}} \le \bar{s}_i^{\text{out}} \qquad \forall i \in D(G) \tag{26}$$

$$s_i^{\text{in}} \ge s_j^{\text{out}} \qquad \forall ji \in A(G) \tag{27}$$

$$x_i \ge s_i^{\text{in}} - s_i^{\text{out}} + L_i \qquad \forall i \in N(G) \tag{28}$$

$$y_i + q_i^{\omega} + \sum_{j \in J} q_{ji}^{\omega} \ge \Psi_i^{\omega}(x_i) \qquad \forall i \in D(G),\ \forall \omega \in \Omega \tag{29}$$

$$\sum_{i \in D(G)} q_{ji}^{\omega} \le C_j \qquad \forall j \in J,\ \forall \omega \in \Omega \tag{30}$$

$$y_i + q_i^{\omega} \ge \Psi_i^{\omega}(x_i) \qquad \forall i \in N(G) \backslash D(G),\ \forall \omega \in \Omega \tag{31}$$

$$x_i,\ s_i^{\text{in}},\ s_i^{\text{out}},\ q_{ji}^{\omega} \ge 0 \qquad \forall i \in N(G),\ \forall j \in J,\ \forall \omega \in \Omega \tag{32}$$

$$y_i,\ q_{ji}^{\omega} \in \mathbb{Z} \qquad \forall i \in N(G),\ \forall \omega \in \Omega \tag{33}$$

Note, that by using the penalty recourse variables we force complete recourse but account for failure by some cost. The computational results in Section 5.4 suggest that the SGSM policies with the tested penalty values dominate GSM-policies in terms of both inventory and recourse cost, not only total cost. This means, the resulting SGSM policy, internally using those successful penalty values, will perform better than the corresponding GSM policies also for *any other* penalty values.

3 Theoretical Comparison of the SGSM with the GSM

One reason for the superior performance of policies based on the SGSM seems to be that it generates structurally different optimal solutions. In this section we give some theoretical evidence for this intuition.

We restrict ourselves to linear demand scenarios. Thus, the stochastic data simplifies to the following: For each $\omega \in \Omega$, the random vector of measurements ξ^{ω} now has the following components:

$$\xi^{\omega} = \left(L_i^{\omega}, \alpha_i^{\omega}\right)_{i \in N(G)}, \tag{34}$$

where $L_i^{\omega} \leq 0$ is the lead time in node i in state ω and $\alpha_i^{\omega} > 0$ is now the constant demand rate in node i in state ω.

Note that we assume that all demand rates are positive. We do this in order to avoid special cases with limited relevance. We call ξ^{ω} the *lead time/demand rate scenario*, or *scenario* for short, of state $\omega \in \Omega$. As before, the probability that $\xi^{\omega} = \xi$ is equal to $\sum_{\omega \in \Omega : \xi^{\omega} = \xi} p^{\omega}$. The set of all scenarios with positive probabilities is again denoted by Ξ. The complete lead time/demand rate distribution is again denoted by (Ξ, p).

3.1 Lead-Times/Demand-Rates Induced by a Distribution and a Target Service Level

In order to prepare for a formalization of these considerations, we elaborate in more detail on the chance-constraint interpretation of the GSM restrictions.

Note first that the conclusions of the GSM (optimality of the inventory decisions in the real world) are formally only correct if, in the real world, at each node all lead times and demand rates are bounded by tight choices of L_i and α_i of the GSM. In that case, a GSM solution will service all customers in time with minimal inventory.

Since this assumption is not realistic over long time-periods, consider the following chance-constraint interpretation: If at some node $i \in N(G)$ the lead time of demand rate exceeds the choice of the upper bounds L_i and α_i chosen for the GSM, then the guaranteed service times computed by the GSM are nevertheless achieved by unspecified actions outside the model (operational flexibility). We require, however, that a prescribed fraction of the demand is handled on time inside the model. If this fraction were set to zero, one would simply shift all demands outside the model, making the GSM pointless. The positive fraction of demands required to be handled inside the model is the target service level mentioned in the introduction. It can in principle be prescribed for each node. In the following we restrict ourselves to the case of a global target service level to reduce notational clutter.

In the chance-constraint interpretation, the GSM arises as a (possibly approximative) deterministic equivalent of the following chance-constraint stochastic program:

$$\min \sum_{i \in N(G)} h_i y_i \tag{35}$$

$$s_i^{\text{out}} \leq \bar{s}_i^{\text{out}} \qquad \forall i \in D(G) \tag{36}$$

$$s_i^{\text{in}} \geq s_j^{\text{out}} \qquad \forall ji \in A(G) \tag{37}$$

$$P_\Omega \begin{bmatrix} x_i \geq s_i^{\text{in}} - s_i^{\text{out}} + L_i^\omega & \forall i \in N(G) \\ y_i \geq \alpha_i^\omega x_i & \forall i \in N(G) \end{bmatrix} \geq n^{\text{target}} \tag{38}$$

$$s_i^{\text{in}},\ s_i^{\text{out}}, x_i,\ y_i \geq 0 \qquad \forall i \in N(G) \tag{39}$$

$$y_i \in \mathbb{Z} \qquad \forall i \in N(G) \tag{40}$$

Since joint chance constraints are difficult to handle in general, one hopes that for a suitable choice of L_i and α_i the GSM is (close to) a deterministic equivalent of such a chance-constraint stochastic program.

In the presence of a finite lead time/demand rate distribution we obtain: A solution to a GSM with lead times L_i and demand rates α_i satisfies all demands on time in scenario ξ^ω without operational flexibility if the GSM solution is also feasible in a GSM with lead times L_i^ω and demand rates α_i^ω. This is the case if $L_i^\omega \leq L_i$ and $\alpha_i^\omega \leq \alpha_i$. Thus, in order to achieve a target service level of $n^{\text{target}} \in (0,1]$, the GSM must choose minimal L_i and α_i such that we have

$$P_\Omega\big(\{\omega \in \Omega \mid L^\omega \leq L_i, \alpha_i^\omega \leq \alpha_i \ \forall i \in N(G)\}\big) \geq n^{\text{target}}. \tag{41}$$

Even for finite general lead time/demand rate distributions L_i and α_i may not be uniquely determined by the distribution and the target service level. The GSM may even produce distinct results for distinct choices.

Example 1 Consider a single-demand-node network with $h = 1$ and $\bar{s}^{\text{out}} = 0$. Let the scenario set be $\Xi = \big((L^1 = 1, \alpha^1 = 2), (L^2 = 3, \alpha^2 = 1)\big)$ with probability $\frac{1}{2}$ for both. In order to achieve a target service level of 0.5 tightly, we can either use $L = 1$, $\alpha = 2$ or $L = 3$ and $\alpha = 1$. The former leads to optimal GSM costs of 2 via $x = 1$ and $y = 2$, while the latter achieves optimal GSM costs of 3 via $x = 3$ and $y = 3$.

We can, however, find unique lead time and demand rate bounds if we assume the following *total-order property* of the lead time/demand rate distribution:

$$L_i^1 \leq L_i^2 \leq \cdots \leq L_i^{|\Omega|} \text{ and } \alpha_i^1 \leq \alpha_i^2 \leq \cdots \leq \alpha_i^{|\Omega|} \quad \forall i \in N(G). \tag{42}$$

In this case, we can identify a *critical scenario* ω^*, defined by

$$\omega^* := \min\big\{\omega \in \Omega \ \big|\ \sum_{\omega' \leq \omega} p^{\omega'} \geq n^{\text{target}}\big\}. \tag{43}$$

It has the following property: A feasible solution to the GSM with lead times $L_i^{\omega^*}$ and demand rates $\alpha_i^{\omega^*}$ is feasible for all scenarios ξ^ω with $\omega \le \omega^*$. Thus, with these choices any feasible solution to the GSM achieves at least the desired target service level. In an optimal GSM solution, for all $i \in N(G)$ constraints (4) and (5) of the GSM are binding. Consequently, the choices of $L_i^{\omega^*}$ and $\alpha_i^{\omega^*}$ are unique. In this situation, we can derive induced lead times and demand rates from a finite lead time/demand rate distribution and a target service level:

Definition 1 (Induced Lead-Times and Demand Rates) Let (Ξ, p) be a finite lead time/demand rate distribution with the total-order property (42). Moreover, let n^{target} be a positive target service level that determines a critical scenario $\xi^* = \xi^{\omega^*} \in \Xi$. Then the *lead time* and the *demand rate in node $i \in N(G)$ induced by (Ξ, p) and* n^{target} are $L^* = L_i^{\omega^*}$ and $\alpha^* = \alpha_i^{\omega^*}$, respectively. Given an SGSM with input data (Ξ, p), we call the GSM with identical marginal holding costs and induced lead times and demand rates the *GSM induced by the SGSM*.

Remark 2 In the case of node-individual target service levels, the chance constraints are not joint over all nodes, but lead time and demand rate restrictions still need to hold jointly. (Even prescribing separate target service levels for those is conceivable, which results in purely individual chance constraints.) It is then sufficient that scenarios can be reordered at each node individually such that lead times and demand rates are monotonically increasing simultaneously. We then have to keep track of a critical scenario for each node. The interpretation is that at each node individually at least the target service level fraction of the demand is handled inside the model. The fraction of the total demand over all nodes handled inside the model can be much smaller then. Prescribing more and more target service levels means the need to properly guess more and more parameter values. The SGSM can be seen as an optimization model for this task. Note that a total order property (global or local) was only introduced for a consistent chance-constraint interpretation of the GSM. The SGSM does not need it.

3.2 Solution Spaces of SGSMs and Induced GSMs

Now that we have related the SGSM input data with the GSM input data via the target service level, we can perform a tighter comparison between the two.

In the following, let (Ξ, p) be a finite lead time/demand distribution with the total-order property. For any target service level n^{target} we denote the critical scenario by $\xi^* = \xi^{\omega^*}$. Recall that $\omega^* \in \Omega$ is minimal such that $\sum_{\omega \le \omega^*} p^\omega = n^* \ge n^{\text{target}}$. Here, n^* is the *actual service level* of the GSM with induced lead times and demand rates. Moreover, in node $i \in N(G)$ let the critical lead time be $L_i^* = L_i^{\omega^*}$, and let the critical demand rate be $\alpha_i^* = \alpha_i^{\omega^*}$.

The following result tells us, given a finite lead time/demand rate distribution, that for any target service level in the GSM, the recourse costs of the SGSM can always be adjusted in such a way that the decisions of the GSM are optimal first-stage decisions in the SGSM.

Theorem 1 *Let (Ξ, p) be a finite lead time/demand distribution with the total-order property and positive demands. Let n^{target} be a target service level. Moreover, let $\big((s^{in})^{\text{GSM}}, (s^{in})^{\text{GSM}}, x^{\text{GSM}}, y^{\text{GSM}}\big)$ be optimal for the GSM with induced lead times $L_i^* \geq 0$ and demand rates $\alpha_i^* > 0$. Then there are marginal expediting costs t_i and marginal outsourcing costs c_i such that for the corresponding SGSM the following solution induced by the GSM is optimal:*

$$
\begin{aligned}
(s_i^{in})^{\text{SGSM}} &= (s_i^{in})^{\text{GSM}} && \forall i \in N(G), && (44)\\
(s_i^{out})^{\text{SGSM}} &= (s_i^{out})^{\text{GSM}} && \forall i \in N(G), && (45)\\
x_i^{\text{SGSM}} &= x_i^{\text{GSM}} && \forall i \in N(G), && (46)\\
y_i^{\text{SGSM}} &= y_i^{\text{GSM}} && \forall i \in N(G), && (47)\\
(r_i^{\omega})^{\text{SGSM}} &= \max\big(0, L_i^{\omega} - x_i^{\text{SGSM}} + (s_i^{in})^{\text{SGSM}} - (s_i^{out})^{\text{SGSM}}\big) && \forall i \in N(G), && (48)\\
(q_i^{\omega})^{\text{SGSM}} &= \max\big(0, \alpha_i^{\omega} x_i^{\text{SGSM}} - y_i^{\text{SGSM}}\big) && \forall i \in N(G). && (49)
\end{aligned}
$$

The proof is an extension of the proof for the equivalence of independent chance-constraint programs and two-stage stochastic programs with simple recourse (see, e.g., [1, Section 3.2]) via optimality conditions. The classic result cannot be applied directly because our technology matrix depends on the stochastic demand rates. The details of the rather technical proof are given in appendix A.

Next, we show that there are SGSM solutions that can not be found by solving an induced GSM.

Theorem 2 *There is a single-node warehouse network G, holding costs h, a lead time/demand-distribution with total-order property, and marginal expediting and outsourcing costs t and c, respectively, with the following property: There is no target service level n^{target} such that the induced GSM has an optimal solution that is optimal for the first stage of the SGSM.*

Proof The network consists of a single node with holding cost $h = 1$. Consider the following data:

$$
\begin{aligned}
&\Xi = \{\xi^1 = (L^1 = 1, \alpha^1 = 1), \xi^2 = (L^2 = 2, \alpha^2 = 2), \xi^3 = (L^3 = 3, \alpha^3 = 3)\},\\
&\qquad p(\xi^1) = p(\xi^2) = p(\xi^3) = \frac{1}{3}, \qquad t = 3, \qquad c = 2, \qquad \bar{s}^{\text{out}} = 0. \qquad (50)
\end{aligned}
$$

Depending on an actual service level of $\frac{1}{3}$, $\frac{2}{3}$, or 1, the corresponding induced GSM has the following unique optimal solutions with the indicated GSM costs:

$$
\begin{aligned}
&(s^{\text{in}} = 0, s^{\text{in}} = 0, x = 1, y = 1) \mapsto 1,\\
&(s^{\text{in}} = 0, s^{\text{in}} = 0, x = 2, y = 4) \mapsto 4,\\
&(s^{\text{in}} = 0, s^{\text{in}} = 0, x = 3, y = 9) \mapsto 9. \qquad (51)
\end{aligned}
$$

The corresponding SGSM solutions with fixed first-stage solutions, the optimal respective recourse values, and their costs read as follows:

$$(s^{\text{in}} = 0, s^{\text{in}} = 0, x = 1, y = 1, r^1 = 0, q^1 = 0, r^2 = 1, q^1 = 1, r^3 = 2, q^3 = 2) \mapsto 6,$$

$$(s^{\text{in}} = 0, s^{\text{in}} = 0, x = 2, y = 4, r^1 = 0, q^1 = 0, r^2 = 0, q^1 = 0, r^3 = 1, q^3 = 2) \mapsto \frac{19}{3},$$
$$(s^{\text{in}} = 0, s^{\text{in}} = 0, x = 3, y = 9, r^1 = 0, q^1 = 0, r^2 = 0, q^1 = 0, r^3 = 0, q^3 = 0) \mapsto 9. \tag{52}$$

However, the SGSM has the following solution with a smaller cost:

$$(s^{\text{in}} = 0, s^{\text{in}} = 0, x = 1, y = 2, r^1 = 0, q^1 = 0, r^2 = 1, q^1 = 0, r^3 = 2, q^3 = 1) \mapsto \frac{17}{3}. \tag{53}$$

□

The previous example shows that the first-stage part of an optimal SGSM solution need not be extremal in the space of first-stage variables restricted by one scenario only. Since in the example the second scenario equals the average scenario, using the average scenario does not help the GSM to find the optimal first stage either.

In particular, the set of policies that can be generated by SGSM instances is a strict superset of the set of policies that can be generated by induced GSM instances.

4 Scenario Generation and Reduction

For an appropriate finite approximation of a realistic distribution of lead times and demands we need enough scenarios to represent every relevant situation at least once. In subsection 4.1 we review some basics about Sample-Average-Approximation (SAA) Methods for general finite approximations of probability distributions. The extensive form of the deterministic equivalent problem grows with an increasing number of scenarios. Therefore, we employ scenario reduction as described in Subsection 4.2.

4.1 Scenario Generation: The SAA-Method

To approximate the distributions of the stochastic parameters we generate random numbers according to the assumed distribution. These random numbers build the scenarios in the discrete distribution approximating the real distribution of the stochastic parameters. All samples are assigned probabilities proportional to the number of times they were generated. Sampling techniques like this are quite common in stochastic programming. See for example [1].

The idea of SAA is to estimate the optimal value function $\mathscr{Q}(x)$ of the second stage of a two-stage stochastic program $\min_{x \in X} \big(c^T x + \mathscr{Q}(x)\big)$ with first-stage variables x, second stage variables y, and random parameters ξ. This function is originally defined as

$$\mathscr{Q}(x) = \mathbb{E}_\xi \big[Q(x,\xi)\big] = \mathbb{E}_\xi \big[\min_{y^\xi \in Y^\xi} q^T y^\xi\big]. \tag{54}$$

Consider a set Ξ of independent, identically distributed samples $\xi \in \Xi$ of the random parameters. Then the following is an unbiased estimator for $\mathscr{Q}(x)$:

$$\hat{\mathscr{Q}}(x) = \frac{1}{|\Xi|} \sum_{\xi \in \Xi} Q(x,\xi), \tag{55}$$

and the problem

$$\min_{x\in X}\left(c^T x + \hat{\mathscr{Q}}(x)\right), \tag{56}$$

which can be solved conceptually much easier than the original problem, yields an unbiased estimator for the solution of the original problem. Further information about SAA can be found in [15] for example.

4.2 Scenario Reduction: The Fast Forward Selection

The goal of scenario reduction is to approximate a discrete distribution with many scenarios by another discrete distribution with significantly fewer scenarios. There are several methods to achieve this goal, usually based on a metric on the space of all possible scenarios (see [6–8]).

Let us now sketch the principle of scenario reduction, since we have to make some choices. The approach to reduce the number of scenarios is based on a distance between two scenarios denoted by $d(\xi^1, \xi^2)$, a quantity that we have to define. Let Ξ' be any subset of Ξ. Then the distance $d(\xi, \Xi')$ of a scenario $\xi \in \Xi$ to the subset Ξ' is given by $d(\xi, \Xi') := \min_{\xi' \in \Xi'} d(\xi, \xi')$.

The discrete distribution with scenario set Ξ' that approximates Ξ best with respect to the distance d can be obtained in the following way: Add the probability of all $\xi \in \Xi \setminus \Xi'$ to the probability of the scenario $\xi'(\xi) \in \Xi'$ that is the closest scenario in Ξ' to ξ with respect to d. In a sense, the scenarios not in Ξ' are replaced by duplicates of the closest scenarios in Ξ'. The quality of this approximation can be quantified by the distance $d(\Xi, \Xi')$ between Ξ and Ξ', which is defined as $d(\Xi, \Xi') = \sum_{\xi \in \Xi} p^{\xi} d(\xi, \Xi')$. The smaller the distance, the better the approximation.

The goal of scenario reduction in general is to find a particular subset Ξ^* of a given cardinality such that $d(\Xi, \Xi^*)$ is minimal.

The exact scenario reduction problem given any distance function d on scenarios and a cardinality k can be formulated as the following k-median problem:

$$\min_{\Xi'}\left\{d(\Xi, \Xi') | \Xi' \subseteq \Xi, |\Xi'| = k\right\}. \tag{57}$$

There are several exact and heuristic approaches to this NP-hard combinatorial optimization problem. For scenario reduction, there are special methods available, like the *fast forward selection*, one of the heuristics introduced in [6–8].

The fast forward heuristic works as follows (see [6–8] for details). It uses the fact that, by enumeration in time quadratic in the number of scenarios, we can find the scenario that yields the best single-scenario approximation. The idea is now as follows. We sequentially extend the current approximation by one new scenario at a time such that in each step Ξ has minimal distance to the current scenario set union the new one.

Assume we already have constructed a subset Ξ_{m-1} consisting of $m-1$ scenarios, $1 \le m \le k$. Consider the updated distance function

$$\bar{d}(\xi, \xi_m) = \min\left\{d(\xi, \xi_m), d(\xi, \Xi_{m-1})\right\}. \tag{58}$$

By construction, we have for all $\xi \in \Xi$

$$\bar{d}(\xi, \xi_m) = d(\xi, \Xi_{m-1} \cup \{\xi_m\}) . \tag{59}$$

Thus, the best possible one-element extension of a reduced scenario set with respect to d is equal to the best possible one-element scenario approximation with respect to $\bar{d}$.

Summarized, the fast forward selection works as follows:

```
begin
  $\Xi^0 = \emptyset$
  $\bar{d} = d$
  for $m = 1, \ldots, k$ do
    Choose $\xi_m^* \in \operatorname{argmin}_{\xi \in \Xi \setminus \Xi_{m-1}} \bar{d}(\Xi, \{\xi_m\})$
    $\Xi_m = \Xi_{m-1} \cup \xi_m^*$
    $update(\bar{d}, \xi_m^*)$
  $\Xi^* = \Xi_k$
  for $\xi \in \Xi$ do
    Choose $\xi^*(\xi) \in \operatorname{argmin}_{\xi^* \in \Xi^*} d(\xi, \xi^*)$
  for $\xi^* \in \Xi^*$ do
    $p_{\xi^*}^* = \frac{1}{|\Xi|} + \sum_{\xi \in \Xi \setminus \Xi^* : \xi^* = \xi^*(\xi)} \frac{1}{|\Xi|}$
  return $\Xi^*$ and $p^*$
end
```

where $update(\bar{d}, \xi_m^*)$ is the following function:

```
begin
  for $\xi \in \Xi$ do
    for $\xi' \in \Xi$ do
      $\bar{d}(\xi, \xi') = \min\left\{\bar{d}(\xi, \xi'), \bar{d}(\xi, \xi_m^*)\right\}$
end
```

The approximation of the lead times and demand distributions is split into two parts. First, a finite number of samples $\xi \in \Xi$ is generated according to the assumed distribution. These samples build a first discrete approximation where every scenario instance occurs with equal probability $p^\xi = 1/|\Xi|$. Second, the resulting discrete distribution is fed into the fast-forward scenario reduction, i.e., it is approximated by a discrete distribution over a subset of scenarios of prescribed cardinality, which have, in general, non-uniform probabilities.

4.3 Symmetric and Asymmetric Distances of Lead-Time/Demand Scenarios

In our computational tests we use two different kinds of distances between two scenarios. The distances are defined by first defining component-wise distances for the lead times and demands of each node, separately. These component-wise distances can be

- transformed into a distance between complete scenarios by computing the Euclidean norm of the components' distances or
- used for a component-wise scenario reduction whenever the lead time/demand distribution is the product of the components' distributions (this is what we did in the simulations in section 5).

The first distance we will refer to as the *symmetric distance*. For the lead time component we define

$$d(L_i^1, L_i^2) = |L_i^1 - L_i^2|. \tag{60}$$

Since the demand component of a scenario in a node consists of different demand rates α for every time interval, we have to compare piecewise linear functions. We assume an equidistant discretization of time (e.g., in weeks) and a piecewise linear demand (e.g., constant demand rates in each week). Let the time intervals of linearity be numbered with increasing time by $r \in R$ (e.g., week numbers).

We will use a discount parameter Δ that depends on the length of the linear pieces of the demand. The motivation is as follows: Since any forecasting error has consequences for the complete remaining time, its influence is smaller if the error occurs later. In our experiments we used $\Delta = 2$ for a time discretization in months. The order $\Psi_i^1(x_i) > \Psi_i^2(x_i)$ depends on the x_i, since $\Psi_i^\omega(x_i)$ is piecewise linear. We separately compare the demand rates $\alpha^{\omega,r}$ on the different domains of linearity and assign a weigth to each difference. The higher r, the further in the future the demand rate is realized, and differences for high values of r get a lower weight in the total distance calculation than the ones at the beginning of the forecast period. This way, scenarios that only differ in later time intervals are considered closer than scenarios that differ in earlier time intervals. We formally define the distance between demand Ψ_i^1 and Ψ_i^2 of two scenarios 1 and 2 as

$$d(\Psi_i^1, \Psi_i^2) = \sum_{r=0}^{R-1} \left| \frac{\alpha_i^{1,r} - \alpha_i^{2,r}}{\Delta^r} \right|, \tag{61}$$

where $\alpha_i^{\omega,r}$ denotes the demand rate during time interval with index r (e.g., during the rth week) in scenario ξ^ω.

There is another option that leads to asymmetric distances. The idea is to anticipate that the approximation is constructed for the use in a stochastic optimization problem. Thus, we would like to find the approximation that yields the least change in the result of the optimization. To decide which scenario is more important for optimization, we need some information about the costs that occur in case of stockholding and in case of stock out. We have this information given as parameter h_i, costs for holding one piece in stock, and c_i costs for having a stockout of one piece.

This way, we can define the *asymmetric distance between the lead time components of two scenarios* as

$$d(L_i^1, L_i^2) = |L_i^1 - L_i^2| \frac{c_i}{h_i} \qquad \text{if } L_i^1 > L_i^2, \tag{62}$$

$$d(L_i^1, L_i^2) = |L_i^1 - L_i^2| \frac{h_i}{c_i} \qquad \text{otherwise.} \tag{63}$$

The definition of asymmetric distance between the demand components of two scenarios is based on the same idea. We define the following *asymmetric distance between the demand components of scenarios*:

$$d(\Psi_i^1,\Psi_i^2) = \sum_{r=0}^{R-1} \left| \frac{\alpha_i^{1,r} - \alpha_i^{2,r}}{\Delta^r} \right| \frac{c_i}{h_i} \qquad \text{if } \alpha_i^{1,0} > \alpha_i^{2,0}, \tag{64}$$

$$d(\Psi_i^2,\Psi_i^1) = \sum_{r=0}^{R-1} \left| \frac{\alpha_i^{1,r} - \alpha_i^{2,r}}{\Delta^r} \right| \frac{h_i}{c_i} \qquad \text{otherwise.} \tag{65}$$

The distance between complete scenarios ξ^1 and ξ^2, in both the symmetric and asymmetric case, can now be defined as

$$d(\xi^1,\xi^2) = \sqrt{\sum_{i\in N(G)} \left(d(L_i^1,L_i^2)^2 + d(\Psi_i^1,\Psi_i^2)^2 \right)}. \tag{66}$$

Alternatively, if the components are stochastically independent, we can perform scenario reduction component-wise. This way, a complete scenario in the reduced scenario set is a combination of scenario components from the reduced sets of scenario components.

The asymmetric reduction does not necessarily approximate the distribution itself as faithfully as the reduction technique based on symmetric distances. We get a bias in our approximation that depends on the fraction of h_i and c_i. It will be shown in the next section that this biased reduction is indeed a better approximation to the solution of the optimization problems because it takes into account the cost of a decision in a certain scenario. To the best of our knowledge, this is not a standard method in the Stochastic Programming literature. Although we have not yet any further evidence beyond the problem studied in this paper, we conjecture that objective-aware scenario reduction might be worth a try in other contexts, too.

5 Simulation Experiments on Real Data

We performed comprehensive computational tests on real-world data from our partner. All computational results report costs incurred by a method in a discrete-time simulation.

5.1 Discretization of Time

The SGSM can only take finite discrete distributions of lead times and demands. Moreover, all scenarios of the demand distributions must be represented by piecewise linear approximations in order to obtain an MILP formulation for the SGSM.

Our partner forecasts the demand for one month. The data includes the expected total demand in the current month, the expected total demand in the coming month and so on. Thus, a straight-forward approach would be to approximate the demand linearly during one month. However: If we simply assume linearity of the demand during one month, then the rough discretization of time into months leads to demand scenarios with too little variation over time.

We can, of course, choose a finer discretization of time in weeks or days. The finer the discretization is, the more realistic the demand function becomes. In order to get a feeling for this influence, we generated random numbers representing the demand over one month or one week. Figure 1 shows an example of differences in the scenarios for discretization in months and in weeks.

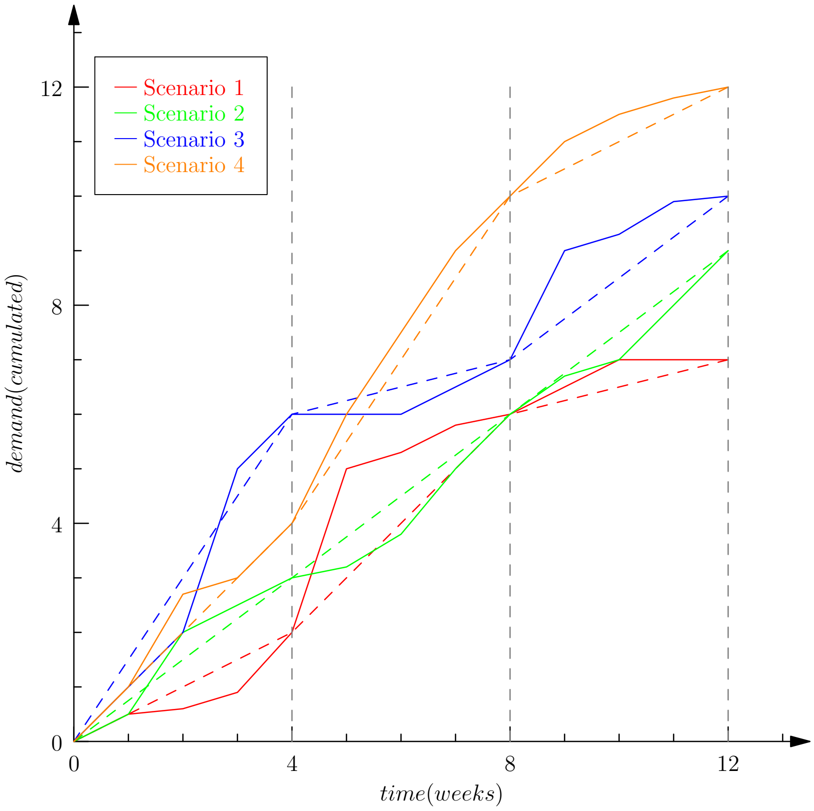

Fig. 1 Different demand scenarios with discretization of time in month (dashed lines) and in weeks (solid lines)

A problem arises if the discretization of time becomes too small. The shorter the linear pieces in the demand functions, the more variables and constraints in the resulting MILP. This is the reason why the results in section 5.4 are all based on discretization in months or weeks. In our simulations, the discretization in days did not lead to substantial savings compared to the one in weeks. (Our simulation itself does not linearize the demand.)

Besides the time discretization, the number of scenarios included in the model is the other quantity that is critical for the mere size and therefore to the computing time of the SGSM. Therefore, we checked the effectivity of the scenario reduction methods in our tests.

5.2 Setup of the Simulation Experiments

We implemented a discrete-time simulation system with independent uniformly distributed random lead times and Poisson distributed demands at all nodes in order to compare the approximated expected long-term dynamic costs incurred by various (s,S)-policies. The policies only differ in the method to compute s for each node.

We investigated two things:

1. How do the parameter settings for scenario generation and scenario reduction influence the long-term cost in the simulation?
2. How does the SGSM with good parameter settings compare with competing algorithms to optimize safety stock?

For the solution of the SGSM we used Sample Average Approximation (SAA) with subsequent scenario reduction according to the old and new techniques Section 4.1 with discount parameters $\Delta = 2$ and $\Delta = 1.25$ for a time discretization in months and weeks, respectively. In order to assess the best parameter settings of our scenario reduction, we computed the simulation results for various sizes of sampled and reduced scenario sets (see Section 4.2). The GSM was parameterized by the target service level: we investigated GSM with service levels that are frequently required at our partner's: 90 % and 96 %. These are denoted by GSM(90%) and GSM(96%). The decentralized policies DEZ(90%) and DEZ(96%) for the service levels of 90% and 96%, respectively, served as a benchmark for policies ignoring the network effects: In these models each location tries to reach the given service level target independently at the smallest cost.

All results reported in Section 5.4 refer to the average long-term cost returned from the respective ten simulation runs for a method under consideration. Since none of the methods is based on an exact model of the dynamic development of the system, each method produces individual systematic errors in the prediction of the long-term costs as soon as the environment does not satisfy all the respective assumptions. Therefore, we chose to make all cost comparisons in a common simulation environment rather than in one of the safety-stock models. We chose a simulation environment that matches the real situation at our partner's as closely as possible. This way we could assess best the real-world impact of the different model approximations and the different computational approximations made in the various methods to compute safety stocks.

All calculations were carried out on a standard PC (CPU: *Intel(R) Core(TM) 2 Quad CPU Q9559 @ 2.83 GHz*, Mem: *8GB RAM*) using ubuntu 4.4.3.

5.3 Test Data

The data used in the simulation is a real-world data set from our partner. The warehouse network is a star-shaped two-echelon spare parts distribution system. It has one master warehouse (no. 0) and seven warehouses (nos. 1–7) for end customer service. The model SGSM is not restricted to this special structure; it can be applied to any acyclic network structure by straightforward modifications.

For our tests we used the same benchmark data set (inventory costs and demand intensities for 1127 spare parts) as in [14]. We used cost coefficients from cost estimates of our partner for inventory cost and the piece-based so-called "non-sales" cost, that determine the recourse cost coefficients. Since these cost coefficients depend on the part, there are too many of them to be listed here.

We used a simulation horizon of 25 months. The end-customer demands were Poisson distributed. The intensities were estimated from historical data. The lead times were uniformly distributed in an interval from 80 % through 120 % around the expected value. The scenario sets were generated as products of independently sampled lead-time and demand scenarios over all nodes. Each policy was confronted with identical sets of lead time and demand samples. Inventory was always controlled by an (s,S)-policy. The values for s were chosen by the models under consideration. Demands had to be fulfilled (backorders with higher priority) whenever possible. Depending on the experiment, unmet demand was backlogged or considered lost. The order quantity was taken from our partner data. Our simulation did not allow for faster delivery than the service times computed by the models.

Because of the limited complexity of the network topology, all instances could be solved in less than an hour for the benchmark assortment of 1127 parts by the MILP solver `gurobi 3.0` (set to an optimality gap of five percent). More accuracy made computations slower but did not gain much with respect to the long-term cost of the resulting policy.

5.4 Computational Results

We first tested the SGSM with many different cardinalities of generated and reduced scenario sets between 1 and 1000, the range in which computations times were viable.[2] Table 1 shows a small selection of the results that show the effectiveness of the reduction methods. The results presented in this table are the average long-term costs of ten simulation runs. The safety stocks were determined by the SGSM, where "$n \to k$" indicates that at each node n lead time and n demand scenarios were generated and reduced to k lead time and k demand scenarios. The lead times and demands were identical in all the simulations.

We can see an enormous reduction in the total costs by applying the reduction techniques introduced in section 4. In the case of generating only three scenarios we observe a very high variability in the costs over the ten simulation runs. During ten simulations, the minimal total costs were 16 379, and the maximal total costs were 33 546. Applying the symmetric/asymmetric reduction technique the minimal total

[2] In practice, before implementing a safety-stock computation based on the SGSM, it seems advisable to test an increasing number of scenarios (generated and reduced) until simulation results do not significantly change anymore or until the computation times become prohibitive. We know of no method that would allow an a-priori estimation of the number of scenarios necessary for an SGSM instance to guarantee a prescribed optimality gap compared to the SGSM instance with the underlying (non-finite) distributions.

Table 1 Results of the SGSM with different scenario reduction techniques

Reduction	Inventory Cost	Recourse Cost	Total Costs
SGSM(3 → 3), months, **no**	1 273	20 958	22 183
SGSM(50 → 3), months, **sym**	1 368	5 799	7 168
SGSM(50 → 3), months, **asym**	1 487	1 877	3 364
SGSM(50 → 50), months, **no**	1 660	1 528	3 188

costs were 6 622/3 246 and the maximal total costs were 7 606/3 546, respectively. The costs occurring in the single simulation runs are listed in appendix B.

These results show that applying scenario reduction leads to a much lower variability in the costs because also scenarios with small probability are taken into account. We can see that the results for the asymmetric reduction are quite close to those where all the fifty generated scenarios are included in the model.

Table 2 includes the service levels in the different locations during the first of the ten simulation runs.

Table 2 Comparison of service levels (%)

Warehouse	**no** 3 → 3	**symmetric** 50 → 3	**asymmetric** 50 → 3 asym	**no** 50 → 50
0	75.4	85.4	73.1	88.9
1	92.4	94.4	95.6	96.8
2	92.5	94.1	95.3	96.3
3	92.1	94.2	95.2	97.0
4	92.0	94.1	95.1	96.2
5	93.0	94.8	96.6	97.5
6	92.0	94.0	96.1	96.3
7	92.9	95.0	95.9	97.0

The service levels in table 2 show the differences between the symmetric and the (new) asymmetric reduction technique. The asymmetric technique takes into account that for many parts the quotient h_i/c_i is greater for the leaf warehouses than for the master warehouse. Therefore, for the symmetric technique we get a higher service level at the master warehouse (no. 0), but lower service levels at the warehouses (nos. 1–7).

Simulating the situation modeled in the SGSM with simple recourse leads to the results listed in table 3.

This table includes the average costs over ten simulation runs of the different approaches. Here we calculated the order points s using all the different methods and ran the simulation ten times with different lead times and demands. For all different approaches the lead times and demands in the simulations were identical.

The results for the decentralized method are worse than the results when the order points are calculated by the GSM. Using one of the listed SGSM approaches leads to a cost reduction of 30% and more. Again, the asymmetric scenario reduction dominates the symmetric one. Another important aspect to notice is that the results using a discretization of time in weeks are remarkably better than results using a discretiza-

Table 3 Results of simulation with uniformly distributed lead times and Poisson distributed demands

Model	Inventory Cost	Recourse Cost	Total Cost
(1) **DEZ(90%)**	2512	2012	4525
(2) **DEZ(96%)**	2987	1019	4006
(3) **GSM(90%)**	2497	1832	4329
(4) **GSM(96%)**	2983	963	3946
(5) **SGSM**(50), months	1555	1474	3029
(6) **SGSM**(200 → 50), months, sym	1561	1498	3058
(7) **SGSM**(200 → 50), months, asym	1690	1282	2972
(8) **SGSM**(200 → 1), months	1466	1410	2876
(9) **SGSM**(200 → 50), weeks, sym	1867	893	2761
(10) **SGSM**(200 → 50), weeks, asym	1884	808	2692

tion in month. Results for each of the ten simulation runs for Model (4) and (10) can be found in appendix B.

In Method (8) a special heuristic is applied (different from the fast forward reduction) that tries to find a critical scenario for the lead time and the demand for every location. This shows that much of the problem's structure can be encoded into a single scenario. This heuristic works properly for the discretization in months and may be extended to finer discretization. This is work in progress.

The resulting service levels for the different methods in the first simulations are shown in table 4.

Table 4 Comparison of service levels (%)

Warehouse	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0	94.1	96.0	94.2	96.3	78.9	89.7	90.0	71.3	92.4	89.8
1	97.6	98.5	97.7	98.4	96.7	96.8	96.9	96.8	97.2	97.3
2	97.1	98.0	97.1	97.9	96.4	96.6	96.7	96.6	96.9	96.9
3	97.4	98.3	97.5	98.3	96.7	97.0	97.0	97.0	97.6	97.8
4	97.4	98.1	97.4	98.1	96.5	96.5	96.6	96.4	96.8	96.8
5	98.4	99.1	98.4	99.1	97.4	97.5	97.5	97.2	97.8	97.7
6	97.1	98.0	97.1	98.1	96.5	96.7	96.7	96.1	96.7	96.7
7	97.4	98.3	97.4	98.2	97.2	97.3	97.5	97.5	97.6	97.7

The differences in the service levels of the symmetric and the asymmetric reduction are no longer substantial. The reason is that now the number of scenarios in the set Ξ^* is much higher; thus, both approaches lead to a good approximation of the distribution and its impact on resulting service levels.

As table 3 shows, the differences in the resulting costs are still remarkable. This is due to more scenarios in the more relevant parts of the distribution in the asymmetric reduction (high lead times and demands if h_i/c_i is low and vice versa).

The results of simulations of the SGSM with external suppliers from which missing parts can be ordered and lost sales (introduced in subsection 2.3) are listed in Table 5:

The simulation works different to the one applied in Tables 1–4. Here the demand that cannot be delivered immediately from the warehouses (nos. 1–7) to the end cus-

Table 5 Results of simulation with external suppliers

Method	Inventory costs	Recourse Costs	Total Costs
(1) **DEZ(90%)**	2 295	1 314	3 609
(2) **DEZ(96%)**	2 472	1 131	3 602
(3) **GSM(90%)**	2 268	1 311	3 580
(4) **GSM(96%)**	2 451	1 121	3 573
(5) **SGSM**(100 → 100), weeks	2 295	793	3 088
(6) **SGSM**(200 → 50), weeks, sym	2 272	868	3 140
(7) **SGSM**(200 → 50), weeks, asym	2 230	689	2 920
(8) **SGSM**(300 → 75), weeks, sym	2 384	859	3 243
(9) **SGSM**(300 → 75), weeks, asym	2 230	608	2 838

tomers is lost. If the warehouses have not enough stock to deliver the ordered parts, there is the possibility to buy these parts from an external supplier. This recourse action causes costs depending on the distance between the warehouse and the external supplier. The supplier himself has limited stock so that the warehouses are not able to order any amount from them. If a demand at a warehouse can be neither delivered from stock nor ordered from an external supplier, the demand is lost.

Internal orders (from a warehouse to the master warehouse) are still backlogged, and the master warehouse delivers the demand as soon as possible to the ordering warehouse. The ordering costs and the capacities of the external suppliers are not included in the data of our partner, so we had to set them artificially.

As we can see in the results of Table 5, the decentralized model and the GSM perform much better in the case with only one kind of uncertainty (demand uncertainty) than in the case of both lead time and demand uncertainty. The SGSM still outperforms the deterministic models achieving 10–20% of cost savings. Table 6 show the resulting service levels of the different methods.

Table 6 Comparison of service levels (%)

Warehouse	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	84.9	87.2	85.0	87.6	88.2	88.2	88.0	88.7	88.8
1	94.2	94.3	94.2	94.3	94.5	94.5	94.6	94.6	94.6
2	94.1	94.1	94.0	94.1	94.3	94.3	94.3	94.3	94.4
3	94.4	94.5	94.4	94.5	94.7	94.7	94.7	94.7	94.7
4	93.9	94.0	93.9	94.0	94.1	94.1	94.1	94.2	94.2
5	94.3	94.5	94.3	94.5	94.7	94.8	94.8	94.8	94.8
6	93.6	94.0	93.6	93.9	94.0	94.0	94.0	94.0	94.0
7	94.1	94.2	94.1	94.3	94.4	94.5	94.4	94.5	94.5

Here the service levels of the SGSM approaches are very similar to those of the decentralized model and the GSM, both with a prescribed service level of 96%. The costs in table 5 tell us that the SGSM treats different parts differently, while the GSM and the decentralized model cover 96% of the demand for every part, no matter what the costs h_i and c_{ji} are. This is the reason why the order points calculated by the SGSM can lead to lower inventory costs and recourse costs at the same time.

Last we want to compare our model to a model that was introduced by Doğru, de Kok and van Houtum, see [4]. In the following we will refer to this model as DoKoHo. In the simulation we need to apply fix lead times as this is one assumption of the DoKoHo model. We simulate a situation that fits to the DoKoHo assumptions where demand is backlogged and there are penalty costs if a location is not able to deliver as demanded.

Table 7 shows the results of the simulation for the GSM, the SGSM, and DoKoHo. There are some different parameter settings for DoKoHo where the penalty costs used in the model are multiplied by a factor (γ). The DoKoHo model outperforms the GSM but causes higher costs than the SGSM. The simulation of the different models lead to the service levels that are shown in table 8.

Table 7 Results of simulation with Poisson distributed demand and fix lead time

Model	Inventory Cost	Recourse Cost	Total Cost
(1) **DoKoHo** ($\gamma = 1$)	1451	2511	3961
(2) **DoKoHo** ($\gamma = 5$)	1817	1535	3352
(3) **DoKoHo** ($\gamma = 10$)	1955	1387	3342
(4) **GSM 96%**	1835	1980	3815
(5) **SGSM** 300 → 75, weeks, asym	1058	1630	2688

Table 8 Comparison of service levels (%)

Warehouse	(1)	(2)	(3)	(4)	(5)
0	48.0	53.7	55.4	91.9	83.3
1	96.8	98.7	99.0	98.1	97.4
2	96.4	98.4	98.6	97.8	97.1
3	96.6	98.7	98.9	97.9	97.4
4	96.3	98.2	98.4	97.6	96.5
5	97.0	98.9	99.1	98.7	97.6
6	96.7	98.1	98.2	97.8	96.6
7	96.6	98.5	98.7	97.9	97.5

The reason for the very low service levels at the master warehouse using the DoKoHo model compared to the ones using GSM or SGSM can be explained easily. In the DoKoHo model there are no explicit service times guaranteed to the successors. But the lower performance of the master warehouse is considered when the successor's safety stock is calculated. In the simulation we use a service time of zero for this case thus the delivery of master warehouse is often late. As all penalty costs at the master warehouse are zero in the simulation, this does not affect the costs that we get applying the DoKoHo models.

6 Conclusion

We have provided additional evidence that the Stochastic Guaranteed Service Model (SGSM), stochastic programming version of the Guaranteed Service Model (GSM), first introduced in [14], induces policies that outperform other policies for the computation of safety stock levels in a multi-echelon spare part distribution system of a large German car manufacturer. Moreover, we enhanced the simple-recourse SGSM by a recourse model that covers outsourcing in a more appropriate fashion by a transportation problem.

In order to take advantage of the SGSM, we could show in our simulations that the stochasticity needs to be captured by sufficiently large sample sizes: in our example we generated 200 scenarios most of the time and reduced them to 50 applying modified scenario reduction techniques. The resulting MILP models could be solved straight-forwardly in our example.

The SGSM makes some assumptions that are only approximations of reality (complete recourse, piecewise linear demand, exogenous lead time and demand distributions in internal nodes). However, our simulation was not restricted by these assumptions; it only checked the resulting policies, no matter what they assumed, and accounted for all the occurring costs. And in this quite realistic simulation experiment, the policies calculated with the SGSM performed extremely well. One reason for this is that the SGSM can have *structurally* different optimal solutions than the GSM: not all optimal SGSM solutions are extreme in the space of variables of the GSM. Thus the SGSM sometimes finds solutions that the GSM can never provide, no matter which target service level. And such solutions dominated the GSM solutions in our simulations.

References

1. Birge, J.R., Louveaux, F.: Introduction to Stochastic Programming. Springer (1997)
2. Clark, A., Scarf, H.: Optimal policies for a multi-echelon inventory problem. Management Science **6**, 475–490 (1960)
3. Diaz, A., Fu, M.C.: Multi-echelon models for repairable items: A review. Document in Decision, Operations & Information Technologies Research Works http://hdl.handle.net/1903/2300, University of Maryland (2005). URL http://hdl.handle.net/1903/2300
4. Doğru, M., de Kok, A., van Houtum, G.: Optimal control of one-warehouse multi-retailer systems with discrete demand. Working paper (2005)
5. Graves, S., Willems, S.: Optimizing strategic safety stock placement in supply chains. Manufacturing & Service Operations Management **2**(1), 68–83 (2000)
6. Heitsch, H.: Stabilität und approximation stochastischer optimierungsprobleme. PhD dissertation, Humbold-Universität zu Berlin (2007)
7. Heitsch, H., Römisch, W.: Scenario reduction algorithms in stochastic programming. Computational Optimization and Applications **24**, 187–206 (2003)
8. Henrion, R., Küchler, C., Römisch, W.: Discrepancy distances and scenario reduction in two-stage stochastic mixed-integer programming. JOURNAL OF INDUSTRIAL AND MANAGEMENT OPTIMIZATION **4**(2), 363–384 (2008)
9. Inderfurth, K.: Safety stock optimization in multi-stage inventory systems. International Journal of Production Economics **24**(1-2), 103 – 113 (1991). DOI 10.1016/0925-5273(91)90157-O. URL http://www.sciencedirect.com/science/article/pii/092552739190157O
10. Inderfurth, K.: Safety stocks in multistage divergent inventory systems: A survey. international journal of production economics **35**, 321–329 (1994)

11. K. Schade: Lagerhaltungsstrategie für mehrstufige Lagerhaltung in der Automobilindustrie. Master's thesis, Universität Bayreuth (2008)
12. Magnanti, T., Shen, Z.J., Shu, J., Simchi-Levi, D., Teo, C.P.: Inventory placement in acyclic supply chain networks. Operations Research Letters **34**, 228–238 (2006)
13. Minner, S.: Dynamic programming algorithms for multi-stage safety stock optimization. OR Spectrum **19**, 261–271 (1997). URL http://dx.doi.org/10.1007/BF01539783. 10.1007/BF01539783
14. Rambau, J., Schade, K.: The stochastic guaranteed service model with recourse for multi-echelon warehouse management. In: Proceedings of the International Symposium on Combinatorial Optimization (ISCO 2010), *Electronic Notes in Discrete Mathematics*, vol. 36, pp. 783–790. Elsevier (2010). To appear
15. Shapiro, A.: Monte Carlo sampling methods. In: A. Ruszczynski, A. Shapiro (eds.) Stochastic Programming, *Handbooks in Operations Research and Management Science*, vol. 10, pp. 353 – 425. Elsevier (2003). DOI DOI: 10.1016/S0927-0507(03)10006-0
16. Sherbrooke, C.: Metric: A multi-echelon technique for recoverable item control. Operations Research **16**, 122–141 (1968)
17. Simpson: In-process inventory. Operations Research **6**, 863–873 (1958)

A Proof of Theorem 1

We repeat the theorem here for convenience.

Theorem 1 *Let (Ξ, p) be a finite lead time/demand distribution with the total-order property and positive demands. Let n^{target} be a target service level. Moreover, let $\big((s^{in})^{\text{GSM}}, (s^{in})^{\text{GSM}}, x^{\text{GSM}}, y^{\text{GSM}}\big)$ be optimal for the GSM with induced lead times $L_i^* \geq 0$ and demand rates $\alpha_i^* > 0$. Then there are marginal expediting costs t_i and marginal outsourcing costs c_i such that for the corresponding SGSM the following solution induced by the GSM is optimal:*

$$
\begin{aligned}
(s_i^{in})^{\text{SGSM}} &= (s_i^{in})^{\text{GSM}} && \forall i \in N(G), && (67)\\
(s_i^{out})^{\text{SGSM}} &= (s_i^{out})^{\text{GSM}} && \forall i \in N(G), && (68)\\
x_i^{\text{SGSM}} &= x_i^{\text{GSM}} && \forall i \in N(G), && (69)\\
y_i^{\text{SGSM}} &= y_i^{\text{GSM}} && \forall i \in N(G), && (70)\\
(r_i^{\omega})^{\text{SGSM}} &= \max\big(0, L_i^{\omega} - x_i^{\text{SGSM}} + (s_i^{in})^{\text{SGSM}} - (s_i^{out})^{\text{SGSM}}\big) && \forall i \in N(G), && (71)\\
(q_i^{\omega})^{\text{SGSM}} &= \max\big(0, \alpha_i^{\omega} x_i^{\text{SGSM}} - y_i^{\text{SGSM}}\big) && \forall i \in N(G). && (72)
\end{aligned}
$$

Proof We prove the assertion by constructing marginal expediting and outsourcing costs from the complementary slackness condition of a primal-dual optimal solution to the GSM. From this, we construct a primal-dual solution to the corresponding SGSM that satisfies complementary slackness condition of the SGSM.

We first list the GSM and its dual DGSM with the assumptions of this section (no integrality constraints, constant demand rates), where all variables appear on the left-hand side and all constants on the right-hand side. With this, the GSM reads as follows:

$$
\min \sum_{i \in N(G)} h_i y_i \tag{73}
$$

$$
\begin{aligned}
-s_i^{\text{out}} & & & & \geq -\bar{s}_i^{\text{out}} \quad & \forall i \in D(G), && (74)\\
s_j^{\text{in}} - s_i^{\text{out}} & & & & \geq 0 \quad & \forall ij \in A(G), && (75)\\
-s_i^{\text{in}} + s_i^{\text{out}} & & + x_i & & \geq L_i^* \quad & \forall i \in N(G), && (76)\\
& & -\alpha_i^* x_i & + y_i & \geq 0 \quad & \forall i \in N(G), && (77)\\
s_i^{\text{in}}, \quad s_i^{\text{out}}, & & x_i, & \quad y_i & \geq 0 \quad & \forall i \in N(G). && (78)
\end{aligned}
$$

With dual variables π_i, ρ_{ij}, σ_i, and τ_i corresponding in that order to the four sets of restrictions, the dual DGSM of the GSM, with restrictions ordered according to s_i^{in}, s_i^{out}, x_i, y_i reads as follows:

$$
\max \sum_{i \in D(G)} (-\bar{s}_i^{\text{out}}) \pi_i \qquad + \sum_{i \in N(G)} L_i^* \sigma_i \tag{79}
$$

$$\sum_{j:ji\in A(G)} \rho_{ji} \qquad - \sigma_i \qquad \leq 0 \;\; \forall i \in N(G), \tag{80}$$

$$-\pi_i - \sum_{j:ij\in A(G)} \rho_{ij} \qquad + \sigma_i \qquad \leq 0 \;\; \forall i \in N(G), \tag{81}$$

$$\sigma_i - \alpha_i^* \tau_i \leq 0 \;\; \forall i \in N(G), \tag{82}$$

$$\tau_i \leq h_i \;\; \forall i \in N(G), \tag{83}$$

$$\pi_i, \qquad \rho_{ij}, \qquad \sigma_i, \qquad \tau_i \geq 0 \;\; \forall i \in N(G), \;\; \forall ij \in A(G). \tag{84}$$

From this we derive the optimality conditions via complementary slackness in primal-dual pairs of feasible solutions:

$$\pi_i\big(-s_i^{\text{out}} + \bar{s}_i^{\text{out}}\big) = 0 \qquad \forall i \in D(G), \tag{85}$$

$$\rho_{ij}\big(s_j^{\text{in}} - s_i^{\text{out}}\big) = 0 \qquad \forall ij \in A(G), \tag{86}$$

$$\sigma_i\big(x_i - s_i^{\text{in}} + s_i^{\text{out}} - L_i^*\big) = 0 \qquad \forall i \in N(G), \tag{87}$$

$$\tau_i\big(-\alpha_i^* x_i + y_i\big) = 0 \qquad \forall i \in N(G), \tag{88}$$

$$s_i^{\text{in}}\big(\sum_{j:ji\in A(G)} \rho_{ji} - \sigma_i\big) = 0 \qquad \forall i \in N(G), \tag{89}$$

$$s_i^{\text{out}}\big(-\pi_i - \sum_{j:ij\in A(G)} \rho_{ij} + \sigma_i\big) = 0 \qquad \forall i \in N(G), \tag{90}$$

$$x_i\big(\sigma_i - \alpha_i^* \tau_i\big) = 0 \qquad \forall i \in N(G), \tag{91}$$

$$y_i\big(\tau_i - h_i\big) = 0 \qquad \forall i \in N(G). \tag{92}$$

Next we do the same for the SGSM. The primal reads as follows:

$$\min \quad \sum_{i\in N(G)} h_i y_i + \sum_{\substack{\omega\in\Omega \\ i\in N(G)}} \big(p^\omega t_i r_i^\omega + p^\omega c_i q_i^\omega\big) \tag{93}$$

$$-s_i^{\text{out}} \qquad \geq -\bar{s}_i^{\text{out}} \;\; \forall i \in D(G), \tag{94}$$

$$s_j^{\text{in}} - s_i^{\text{out}} \qquad \geq 0 \qquad \forall ij \in A(G), \tag{95}$$

$$-s_i^{\text{in}} + s_i^{\text{out}} \quad + x_i \qquad + r_i^\omega \qquad \geq L_i^* \quad \forall i \in N(G), \;\; \forall \omega \in \Omega, \tag{96}$$

$$-\alpha_i^\omega x_i + y_i \qquad + q_i^\omega \geq 0 \quad \forall i \in N(G), \;\; \forall \omega \in \Omega, \tag{97}$$

$$s_i^{\text{in}}, \quad s_i^{\text{out}}, \qquad x_i, \quad y_i, \qquad r_i^\omega, \qquad q_i^\omega \geq 0 \quad \forall i \in N(G), \tag{98}$$

$$\forall \omega \in \Omega. \tag{99}$$

With dual variables π_i, ρ_{ij}, σ_i^ω, and τ_i^ω corresponding in that order to the four sets of restrictions, the dual DSGSM of the SGSM, with restrictions ordered according to s_i^{in}, s_i^{out}, x_i, y_i, r_i^ω, q_i^ω, reads as follows:

$$\max \sum_{i\in D(G)} (-\bar{s}_i^{\text{out}})\pi_i \qquad + \sum_{i\in N(G)} L_i^* \sigma_i^\omega \tag{100}$$

$$\sum_{j:ji\in A(G)} \rho_{ji} \quad - \sum_{\omega\in\Omega} \sigma_i^\omega \qquad \leq 0 \qquad \forall i \in N(G), \tag{101}$$

$$-\pi_i - \sum_{j:ij\in A(G)} \rho_{ij} \quad + \sum_{\omega\in\Omega} \sigma_i^\omega \qquad \leq 0 \qquad \forall i \in N(G), \tag{102}$$

$$\sum_{\omega\in\Omega} \sigma_i^\omega - \sum_{\omega\in\Omega} \alpha_i^\omega \tau_i^\omega \leq 0 \qquad \forall i \in N(G), \tag{103}$$

$$\sum_{\omega\in\Omega} \tau_i^\omega \leq h_i \qquad \forall i \in N(G), \tag{104}$$

$$\sigma_i^\omega \qquad \leq p^\omega t_i \;\; \forall i \in N(G), \tag{105}$$

$$\forall \omega \in \Omega, \tag{106}$$

$$\tau_i^{\omega} \leq p^{\omega} c_i \; \forall i \in N(G), \tag{107}$$

$$\forall \omega \in \Omega, \tag{108}$$

$$\pi_i, \qquad \rho_{ij}, \qquad \sigma_i^{\omega}, \qquad \tau_i^{\omega} \geq 0 \quad \forall i \in N(G), \; \forall ij \in A(G), \; \forall \omega \in \Omega. \tag{109}$$

The resulting optimality conditions for the SGSM are as follows:

$$\pi_i\big(-s_i^{\text{out}} + \bar{s}_i^{\text{out}}\big) = 0 \qquad \forall i \in D(G), \tag{110}$$

$$\rho_{ij}\big(s_j^{\text{in}} - s_i^{\text{out}}\big) = 0 \qquad \forall ij \in A(G), \tag{111}$$

$$\sigma_i^{\omega}\big(x_i - s_i^{\text{in}} + s_i^{\text{out}} + r_i^{\omega} - L_i^{\omega}\big) = 0 \qquad \forall i \in N(G), \omega \in \Omega, \tag{112}$$

$$\tau_i^{\omega}\big(-\alpha_i^{\omega} x_i + y_i + q_i^{\omega}\big) = 0 \qquad \forall i \in N(G), \omega \in \Omega, \tag{113}$$

$$s_i^{\text{in}}\big(\sum_{j:ji \in A(G)} \rho_{ji} - \sum_{\omega \in \Omega} \sigma_i^{\omega}\big) = 0 \qquad \forall i \in N(G), \tag{114}$$

$$s_i^{\text{out}}\big(-\pi_i - \sum_{j:ij \in A(G)} \rho_{ij} + \sum_{\omega \in \Omega} \sigma_i^{\omega}\big) = 0 \qquad \forall i \in N(G), \tag{115}$$

$$x_i\big(\sum_{\omega \in \Omega} \sigma_i^{\omega} - \sum_{\omega \in \Omega} \alpha_i^{\omega} \tau_i^{\omega}\big) = 0 \qquad \forall i \in N(G), \tag{116}$$

$$y_i\big(\sum_{\omega \in \Omega} \tau_i^{\omega} - h_i\big) = 0 \qquad \forall i \in N(G), \tag{117}$$

$$r_i^{\omega}\big(\sigma_i^{\omega} - p^{\omega} t_i\big) = 0 \qquad \forall i \in N(G), \omega \in \Omega, \tag{118}$$

$$q_i^{\omega}\big(\tau_i^{\omega} - p^{\omega} c_i\big) = 0 \qquad \forall i \in N(G), \omega \in \Omega. \tag{119}$$

Consider now a primal-dual pair of optimal solutions to the GSM and DGSM, respectively, with target service level $n^{\text{target}} > 0$ and actual service level $n^* > 0$, denoted by

$$\big(x^{\text{GSM}}, y^{\text{GSM}}, (s^{in})^{\text{GSM}}, (s^{in})^{\text{GSM}}\big), \quad \big(\pi^{\text{GSM}}, \rho^{\text{GSM}}, \sigma^{\text{GSM}}, \tau^{\text{GSM}}\big). \tag{120}$$

Case 1: $n^* = 1$. In this case, the lead times and demand rates in all scenarios are bounded by L_i^* and α_i^*, respectively. We claim that for all $c_i > \frac{h_i}{p^{\omega^*}}$ and $t_i > \alpha_i^* c_i$ the given SGSM solution is optimal. We show first, that the SGSM optimality conditions imply that $r_i^{\omega} = 0$ and $q_i^{\omega} = 0$ for all $\omega \in \Omega$ and all $i \in N(G)$.

Indeed: Assume, for the sake of contradiction, that there is an optimal solution to the SGSM/DSGSM with $q_i^{\omega} > 0$. Then, by the total-order property, $q_i^{\omega^*} > 0$. Equation (119) implies $\tau_i^{\omega} = p^{\omega^*} c_i$. Therefore, we have

$$\sum_{\omega \in \Omega} \tau_i^{\omega} \geq \tau_i^{\omega^*} = p^{\omega^*} c_i > h_i, \tag{121}$$

which contradicts the feasibility of $\tau_i^{\omega^*}$ in DSGSM. Thus, $q_i^{\omega} = 0$ for all $\omega \in \Omega$.

Assume next, for the sake of contradiction, that there is an optimal solution to the SGSM/DSGSM with $r_i^{\omega} > 0$. Again, this implies that $r_i^{\omega^*} > 0$. Then equation (118), $\alpha_i^* \geq \alpha_i^{\omega}$, and the feasibility of τ_i^{ω} in DSGSM imply the following:

$$\sum_{\omega \in \Omega} \sigma_i^{\omega} - \sum_{\omega \in \Omega} \alpha_i^{\omega} \tau_i^{\omega} \geq \sigma_i^{\omega^*} - \sum_{\omega \in \Omega} \alpha_i^{\omega} \tau_i^{\omega} \tag{122}$$

$$\geq p^{\omega^*} t_i - \alpha_i^* \sum_{\omega \in \Omega} \tau_i^{\omega} \tag{123}$$

$$\geq p^{\omega^*} t_i - \alpha_i^* h_i \tag{124}$$

$$> p^{\omega^*} \alpha_i^* \frac{h_i}{p^{\omega^*}} - \alpha_i^* h_i \tag{125}$$

$$= 0, \tag{126}$$

which this time contradicts the feasibility of $\sigma_i^{\omega^*}$ in DSGSM.

Thus, in the SGSM with the given marginal costs for expediting and outsourcing, expediting and outsourcing quantities r_i^ω and q_i^ω, respectively, can be fixed to zero. The resulting SGSM is identical to the GSM. Hence, the SGSM solution from the assertion, whose first-stage part equals an optimal GSM solution, is optimal.
Case 2: $0 < n^* < 1$. As an abbreviation for the following, we define

$$\bar{n} := \sum_{\omega > \omega^*} p^\omega = 1 - n^* > 0 \quad \text{and} \quad \bar{\alpha}_i := \sum_{\omega > \omega^*} \alpha_i^\omega p^\omega > 0. \tag{127}$$

Define marginal costs for expediting and outsourcing as follows:

$$c_i := \frac{\tau^{\text{GSM}}}{\bar{n}} \quad \text{and} \quad t_i := \frac{\bar{\alpha}_i}{\bar{n}} c_i. \tag{128}$$

Note that the definition of c_i corresponds to the standard penalty to model chance constraints with stochastic right-hand side, whereas the definition of t_i is different because we have to keep under control the stochastic coefficient α_i^ω in front of x_i.

The SGSM solution from the assertion of the theorem is obviously feasible for the SGSM. We claim that the following is a solution to the DSGM, which together with the given SGSM solution satisfies the optimality conditions for the SGSM:

$$\pi_i^{\text{SGSM}} := \frac{\bar{\alpha}_i}{\alpha_i^* \bar{n}} \pi_i^{\text{GSM}} \qquad \forall i \in N(G), \tag{129}$$

$$\rho_{ij}^{\text{SGSM}} := \frac{\bar{\alpha}_i}{\alpha_i^* \bar{n}} \rho_{ij}^{\text{GSM}} \qquad \forall i \in N(G), \tag{130}$$

$$(\sigma_i^\omega)^{\text{SGSM}} := \begin{cases} 0 & \text{if } \omega \le \omega^*, \\ p^\omega t_i & \text{if } \omega > \omega^*; \end{cases} \qquad \forall i \in N(G), \tag{131}$$

$$(\tau_i^\omega)^{\text{SGSM}} := \begin{cases} 0 & \text{if } \omega \le \omega^*, \\ p^\omega c_i & \text{if } \omega > \omega^*; \end{cases} \qquad \forall i \in N(G). \tag{132}$$

Since $(\sigma_i^\omega)^{\text{SGSM}}$ and $(\tau_i^\omega)^{\text{SGSM}}$ are only positive for $\omega > \omega^*$, the validity of the SGSM optimality equations (112) and (113) follows from the definitions of $(r_i^\omega)^{\text{SGSM}}$ and $(q_i^\omega)^{\text{SGSM}}$ and the validity of the GSM optimality equations (87) and (88).

The validity of the SGSM optimality equations (118) and (119) follows directly from the definitions of $(\sigma_i^\omega)^{\text{SGSM}}$ and $(\tau_i^\omega)^{\text{SGSM}}$.

Furthermore, we have

$$\sum_{\omega \in \Omega} (\tau_i^\omega)^{\text{SGSM}} = \sum_{\omega > \omega^*} (\tau_i^\omega)^{\text{SGSM}} = \sum_{\omega > \omega^*} p^\omega c_i = \sum_{\omega > \omega^*} p^\omega \frac{\tau^{\text{GSM}}}{\bar{n}} = \tau^{\text{GSM}}. \tag{133}$$

Thus, since $y_i^{\text{SGSM}} = y_i^{\text{GSM}}$, the validity of all the remaining SGSM optimality equations (117) containing $\sum_{\omega \in \Omega} (\tau_i^\omega)^{\text{SGSM}}$ follows from the validity of the corresponding GSM optimality equations (92) with τ_i^{GSM}. Moreover:

$$\sum_{\omega \in \Omega} (\sigma_i^\omega)^{\text{SGSM}} = \sum_{\omega > \omega^*} (\sigma_i^\omega)^{\text{SGSM}} = \sum_{\omega > \omega^*} p^\omega t_i = \bar{n} t_i = \bar{\alpha}_i c_i = \sum_{\omega > \omega^*} \alpha_i^\omega (\tau_i^\omega)^{\text{SGSM}}. \tag{134}$$

This proves the validity of the SGSM optimality equations (116).

On the other hand, for all $i \in N(G)$:

$$\sum_{\omega \in \Omega} (\sigma_i^\omega)^{\text{SGSM}} = \bar{\alpha}_i c_i = \bar{\alpha}_i \frac{\bar{n}}{\bar{n}} c_i = \frac{\bar{\alpha}_i}{\bar{n}} (\bar{n} c_i) = \frac{\bar{\alpha}_i}{\bar{n}} \tau_i^{\text{GSM}} = \frac{\bar{\alpha}_i}{\alpha_i^* \bar{n}} \sigma_i^{\text{GSM}}. \tag{135}$$

Moreover, recall that for all $i \in N(G)$ we have defined:

$$\pi_i^{\text{SGSM}} = \frac{\bar{\alpha}_i}{\alpha_i^* \bar{n}} \pi_i^{\text{GSM}} \tag{136}$$

$$\rho_{ij}^{\text{SGSM}} = \frac{\bar{\alpha}_i}{\alpha_i^* \bar{n}} \rho_{ij}^{\text{GSM}}. \tag{137}$$

Now, scale by $\frac{\bar{\alpha}_i}{\alpha_i^* \bar{n}} > 0$ the optimal GSM solutions in the homogeneous GSM optimality equations (89) and (90) as well as (85), (86) and (87). In the resulting valid equations, using (135), (136), and (137), substitute $\frac{\alpha}{\alpha_i^* \bar{n}} \sigma_i^{\text{GSM}}$ by $\sum_{\omega \in \Omega} (\sigma_i^{\omega})^{\text{SGSM}}$, $\frac{\bar{\alpha}_i}{\alpha_i^* \bar{n}} \pi_i^{\text{GSM}}$ by π_i^{SGSM}, and $\frac{\bar{\alpha}_i}{\alpha_i^* \bar{n}} \rho_{ij}^{\text{GSM}}$ by ρ_{ij}^{SGSM}. The resulting valid equations are exactly the corresponding SGSM optimality equations with the SGSM/DSGSM values that we have defined. Therefore, the SGSM optimality equations (114) and (115) as well as (110), (111) and (112) are satisfied for the given DSGSM solution. Analogously, we can show that the given solution to the DSGSM is feasible for the DSGSM. Consequently, the asserted SGSM solution is optimal. □

B Results of the Individual Simulation Runs

In this section we show some of the numerical results in detail. The average costs of the ten simulation runs listed here are given in tables 1 and 3. The tables 9–12 include the results that lead to the average costs of table 2. Tables 13 and 14 include the results for the single runs of **GSM 96%** (4) and **SGSM** 200 → 50, weeks, asym (10) of table 3.

Table 9 No reduction (3 → 3)

Run	Inventory Costs	Recourse Costs	Total Costs
1	1 262	20 905	22 167
2	1 228	17 573	18 801
3	1 238	22 342	23 580
4	1 294	19 942	21 237
5	1 311	21 047	22 358
6	1 257	20 861	22 118
7	1 326	15 529	16 379
8	1 262	20 905	22 167
9	1 242	32 304	33 546
10	1 311	18 170	19 481
Average	**1 273**	**20 958**	**22 183**

Table 10 Symmetric reduction (50 → 3)

Run	Inventory Costs	Recourse Costs	Total Costs
1	1 344	5 886	7 230
2	1 375	6 232	7 606
3	1 357	5 264	6 622
4	1 358	5 693	7 052
5	1 378	6 038	7 415
6	1 357	5 469	6 826
7	1 376	5 947	7 323
8	1 401	5 516	6 917
9	1 358	5 828	7 187
10	1 379	6 118	749
Average	**1 368**	**5 799**	**7 168**

Table 11 Asymmetric reduction (50 → 3)

Run	Inventory Costs	Recourse Costs	Total Costs
1	1 486	1 981	3 467
2	1 478	1 768	3 246
3	1 477	1 833	3 310
4	1 509	1 744	3 252
5	1 510	1 861	3 371
6	1 475	1 866	3 341
7	1 502	1 889	3 391
8	1 496	1 910	3 406
9	1 481	1 827	3 307
10	1 458	2 089	3 546
Average	**1 487**	**1 877**	**3 364**

Table 12 No Reduction (50 → 50)

Run	Inventory Costs	Recourse Costs	Total Costs
1	1 651	1 532	3 182
2	1 657	1 392	3 049
3	1 655	1 520	3 175
4	1 652	1 512	3 164
5	1 660	1 511	3 171
6	1 658	1 577	3 235
7	1 664	1 561	3 226
8	1 673	1 509	3 182
9	1 670	1 525	3 195
10	1 655	1 642	3 297
Average	**1 660**	**1 528**	**3 188**

Table 13 GSM with a prescribed service level of 96%

Run	Inventory Costs	Recourse Costs	Total Costs
1	2 977	954	3 931
2	2 985	957	3 942
3	2 982	1 017	3 999
4	2 988	945	3 934
5	2 983	1 086	4 069
6	2 993	914	3 907
7	2 998	881	3 879
8	2 975	963	3 938
9	2 972	928	3 899
10	2 978	985	3 964
Average	**2 983**	**963**	**3 946**

Table 14 SGSM(200 → 3) asymmetric reduction with time discretization in weeks

Run	Inventory Costs	Recourse Costs	Total Costs
1	1 870	794	2 664
2	1 895	743	2 637
3	1 892	834	2 727
4	1 875	764	2 639
5	1 876	997	2 873
6	1 884	772	2 656
7	1 880	812	2 692
8	1 894	809	2 703
9	1 885	768	2 653
10	1 890	788	2 678
Average	**1 884**	**808**	**2 692**