Extensions to the Guaranteed Service Model for Industrial Applications of Multi-Echelon Inventory Optimization

Victoria G. Achkar[a,b], Braulio B. Brunaud[c], Héctor D. Pérez[d], Rami Musa[c], Carlos A. Méndez[a,b], Ignacio E. Grossmann[d]

[a] *Universidad Nacional del Litoral, Argentina*

[b] *INTEC (UNL – CONICET), Argentina*

[c] *Johnson & Johnson, USA*

[d] *Carnegie Mellon University, USA*

Keywords: inventory, optimization, guaranteed-service, multi-echelon

Abstract. Multi-echelon inventory optimization (MEIO) plays a key role in a supply chain seeking to achieve specified customer service levels with a minimum capital in inventory. This is especially true in real supply chains, which face supply and demand uncertainty. In this work, we propose a generalized MEIO model based on the Guaranteed Service approach to allocate safety stock levels across the network at the lowest holding cost. This model simultaneously accounts for several features that are usually present in real multi-echelon supply chains: review periods, manufacturing facilities, hybrid nodes (nodes with both internal and external demand), minimum order quantities (MOQ), and different service level performance indicators (fill rate and cycle service levels). We propose an integrated nonlinear programming model to support decision making, which can be reformulated as a Quadratically Constrained Program (QCP), which leads to order of magnitude reductions in computational time. The performance of the model is evaluated by solving illustrative and real-world cases, and is validated with simulation.

1. INTRODUCTION

On-time fulfilment of customer demand is critical in today's customer-centric supply chains. Achieving this goal depends in great part on the inventory levels and policies that are set along a supply chain. However, efficient inventory control is particularly challenging when customer demand is uncertain and retailers may not know the exact size of an order in advance. Other sources of uncertainty may increase the problem complexity, such as lead time variability. Moreover, the decision at one stage impacts inventory decisions at other stages. To overcome these challenges, having a safety stock serves to mitigate the risk of stock-outs in the system. The purpose of multi-echelon inventory optimization (MEIO) is to allocate safety stocks to meet customer service levels, while minimizing the total capital tied up in inventory throughout the supply chain, in contrast to single-echelon inventory optimization

(SEIO), which seeks to independently minimize cost at each echelon. MEIO has enabled companies to reduce their inventories up to 30% and improve item availability up to 5% by supporting supply chain segmentation, and providing a better balance between lead time, inventory and service under uncertainty (Payne, 2016). From an optimization perspective, making decisions about inventory in multi-echelon systems is a challenging task because the objective functions usually involve non-linearities, and decision variables affect more than one echelon.

MEIO approaches have been studied in the literature for allocating safety stock in supply chains. The intent of safety stock allocation is to determine an overall strategy for deploying inventory levels across the supply chain in order to buffer it against sources of uncertainty (Graves & Willems, 2003). De Kok et al. (2018) present a general typology and review stochastic MEIO models in which they classify the extensive research on multi-echelon inventory management under model assumptions, research goals, and different applied methodologies. They state that multi-echelon inventory systems are still a very active area of research because of their complexity and practical relevance. More recently, Gonçalves et al. (2020) present a systematic literature review describing the history and trends regarding the safety stock determination from an operations research perspective. They also highlight that the number of contributions to MEIO has seen a significant increase from the year 2005 onwards, and they list many potential directions and trends for future research.

There are two widely known approaches in MEIO to determine safety stock levels: the stochastic-service model (SSM) and the guaranteed-service model (GSM), which were introduced by Clark and Scarf (1960) and Simpson (1958), respectively. These two approaches differ in terms of general characteristics and provide an interesting contrast on how a supply chain is represented for the purpose of setting safety stocks. Detailed comparisons can be found in De Smet et al. (2019), Graves and Willems (2003), and Simchi-Levi and Zhao (2012).

The objective of this work is to develop a guaranteed-service optimization model to address the problem of optimizing multi-echelon supply chain network inventory management for complex systems. However, the GSM for multi-echelon supply chains developed in literature does not fully account for all the issues and characteristics arising in industrial practice. Many authors have developed some extensions to the GSM, but to the best of our knowledge, nobody has developed a model that can achieve optimum safety stocks on complex supply chains while integrating all the features typical of industrial environments presented in this work. We integrate them into a single model that enables an improved real-world supply chain representation in order to provide support to strategic decision-makers.

This paper presents a model for multi-echelon inventory systems of a multi-product supply chain with both demand and lead time uncertainty. The novelty in the proposed model is that it combines several features. First, demand can occur at any node in the network. This can result in hybrid nodes that have both dependent and independent demands. To the best of our knowledge, these characteristics, which represent the common operation mode of many real multi-echelon systems, has not been

addressed before, as most of the literature on supply chain inventory management considers only external demand at the final nodes of the network. Our proposed formulation also captures risk-pooling effects by consolidating the safety stock inventory of downstream nodes to the upstream nodes in the multi-echelon supply chain. Second, manufacturing plants can be placed at any location in the network, enabling the manufacture of any desired good at those locations. This feature allows generalizing and managing larger supply chains that have grown in their vertical integration. Capturing wider networks can significantly improve the inventory decision-making process across the process supply chains as is seen in those industries that produce both raw materials and finished goods. Third, fill rates can be used as an alternative customer service key performance indicator when setting safety stock levels. Fill rates are not considered in the standard GSM, which relies on cycle service levels instead. However, fill rate is the most widely applied service level measure in industry (Teunter et al., 2017).. We thus allow the modeler the flexibility of specifying the customer service metric to be used. In order to do this, we propose a quadratic regression to estimate the equivalent Cycle Service Level (CSL) when fill rates are used in the model as the desired customer service measure. In addition, minimum order quantities (MOQ) for replenishment orders are explicitly modelled. To the best of our knowledge, literature on the impact of MOQ on safety stock levels is scarce and mostly focuses on continuous-review inventory policies. Depending on the size of the order, using MOQ can cause overshoot in the inventory levels, impacting service levels and costs. This is frequently observed in almost all supply chains. Finally, the resulting nonconvex Nonlinear Programming (NLP) model is reformulated as a Quadratically Constrained Problem (QCP) by exploiting the structure of the constraints of the base model. Several computational examples for illustrative and real industrial systems are presented to illustrate the application of the proposed model and its resulting improved computational performance.

The outline of the paper is as follows. The literature review and background with the basic concepts of the GSM are presented in the following subsection. The problem statement is given in Section 2, followed by the model formulation in Section 3. Section 4 details the application of the model on illustrative and real-world case studies. We conclude this article in Section 5. A Nomenclature section is presented at the end to facilitate the model understanding. A Supporting Information Section is included to provide the data input used in the real case study and detail additional discoveries relating to the impact of MOQ on service metrics.

1.1 Literature and Background of the Guaranteed-service Model

The present paper relies on the GSM to optimize safety stocks. Although this approach was developed more than 50 years ago, 80% of the existing works on this topic have been published in the last 2 decades (Eruguz et al., 2016). The first multi-echelon serial system for the GSM model was proposed by Simpson (1958), and then it was extended to deal with different network topologies (Graves & Willems, 2000; Inderfurth, 1993; Inderfurth & Minner, 1998; Minner, 1998). Later, Magnanti et al. (2006) developed a guaranteed-service approach for general acyclic networks. The main

idea of the classic guaranteed service approach is that if the customer places an order of size $d_j(t)$ on node j at time t, this order will be fulfilled by time $t + S_j$ (Graves & Willems, 2000), with S_j being the guaranteed-service time of node j. Moreover, each node j receives a service commitment from its upstream node $i \in J$, called inbound service time SI_j ($SI_j = S_i$), and has an order processing time or lead time of LT_j. This lead time represents the time until the outputs are available to serve the demand, including material handling and transportation times (You & Grossmann, 2009). Both SI_j and LT_j are times that must be taken into account to define S_j. The Net Lead Time (NLT) is a concept that links them and represents the period of exposure that is not covered within the guaranteed service time and must be covered with safety stock. The *NLT* for node j is defined as $NLT_j = SI_j + LT_j - S_j$. Figure 1 displays examples for different values of S_j. The first example (1), also called "pull" or "full delay" scenario, is the case where node j promises to its customer a guaranteed service time equal to the worst-case replenishment time ($S_j = SI_j + LT_j$). This node places an order to its predecessor every time it receives an order from its customer, then it waits for the upstream node to process its order before processing the order without storing any inventory. In this case, $NLT_j = 0$. On the other hand, if the customer bounds the maximum possible service time ($S_j \leq maxS_j$) and this maximum is less than the worst-case replenishment time ($maxS_j \leq SI_j + LT_j$), node j should satisfy customer demand in less time that the required to place an order on the supplier and process it. Therefore, $NLT_j > 0$, meaning that there is a period of time that should be covered with safety stock, as shown in cases (2) and (3) in Figure 1.

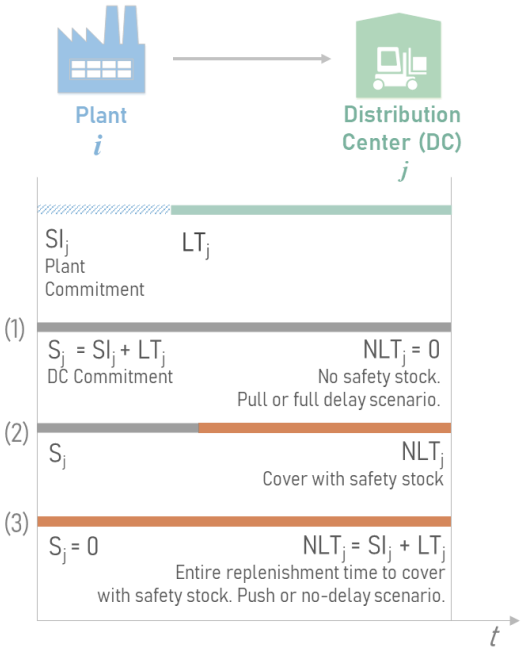

Figure 1: Examples of different values for *SIj*, *Sj*, and *NLTj*: 1) pull scenario, 2) intermediate scenario, and 3) push scenario.

The aim of the GSM is to define the values of SI_j and S_j in order to reduce the safety stock holding cost. This may seem a simple task for a single-echelon supply chain. However, as shown in Figure 2,

the guaranteed-service time defined for one node impacts the downstream stages in the network, because the guaranteed service time for the node becomes the inbound service time for its downstream successors ($SI_j = S_i$). In case (1), avoiding safety stocks in node j yields large inventory levels on the successor stage k (proportional to NLT_k), while in case (2) the inventory level in k is reduced by holding stock in j. Therefore, the GSM on multi-echelon systems is more complex and requires optimizing the network as a whole.

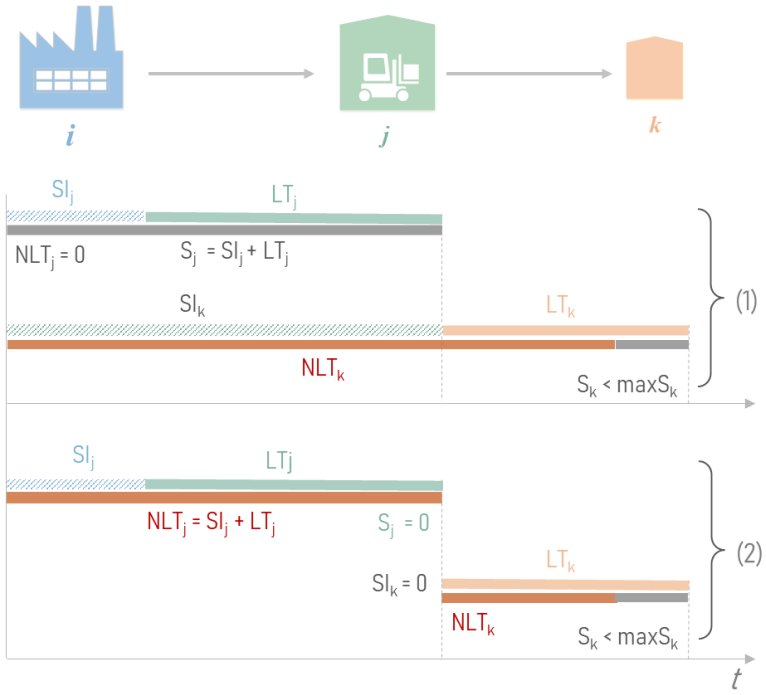

Figure 2: Guaranteed service approach in multi-echelon supply chains.

A key assumption of the basic GSM is that demand is bounded. If demand in a certain period exceeds the bound, it is assumed that other extraordinary measures such as overtime production are used to satisfy excess demand. Moreover, it is assumed that each stage of the supply chain operates under a periodic review inventory policy with a base-stock level, with the review period being common for all echelons. The demand is independent and identically distributed at each demand node, following a normal distribution. Lead times are constant, and independent demand only occurs at final nodes in the network. In addition, the service times at the initial and the final nodes are inputs. Finally, each plant has a coefficient that represents the bill of materials for product transformation and depends on location-location relationships (network arcs).

Over the years, many authors have worked on extending the original GSM assumptions to enable real-world supply chain characteristics to be captured, as presented in the survey by Eruguz et al. (2016). The authors in this survey summarize several extensions made to the basic model of Graves and Willems (2000). The main assumptions that were relaxed are related to the external demand, lead time variability, capacity constraints, service time customization, alternative replenishment policies, review periods, and extraordinary measures. Moreover, other authors have presented works about integrating

the classic GSM with other activities or approaches. You and Grossmann (2008) develop models and algorithms that simultaneously consider inventory optimization and supply chain network design under demand uncertainty. In a subsequent work, these authors present an integrated multi-echelon supply chain design and inventory management model under uncertainty using the GSM (You & Grossmann, 2009). Klosterhalfen et al. (2013) propose an integrated hybrid guaranteed-service and stochastic-service approach for inventory optimization, that allows selecting the approach that minimizes costs. Recently, the work by Ghadimi (2020) presents a model for joint optimization of production capacity and safety stocks under the GSM approach. Bendadou et al. (2021) analyze the impact of merging activities in a supply chain under the GSM.

In addition to the extensions mentioned in (Eruguz et al., 2016), the inclusion of MOQ and fill rate as a service level measure have significant importance for representing real-world supply chain dynamics and must be accounted for. Chopra and Meindl (2013) define the Cycle Service Level (CSL) as the fraction of replenishment cycles that end with all the customer demand being met, where the replenishment cycle is the interval between two successive replenishment deliveries. In other words, it is the probability that an order is fulfilled "on time in full" (OTIF). On the other hand, the product fill rate (*fr*) is the fraction of product demand that is fulfilled on time from the product in inventory. The latter one is known as the most widely applied service level measure in industry (Teunter et al., 2017). However, CSL is the measure required by the GSM. Chopra and Meindl (2013) describe how to introduce the fill rate given a continuous review inventory policy with a formula that links both indicators to obtain the equivalent CSL. They also describe how a large MOQ yields larger fill rates. Silver and Bischak (2011) present an exact fill rate in a periodic review base stock system under normally distributed demand, and they state that the fill rate depends on four parameters, safety factor, coefficient of variation, review period, and lead time, but not on the minimum order quantity. Other works (Park et al., 2018; Shen et al., 2019; Zhou et al., 2007; Zhu et al., 2015) focus on the impact of the MOQ on inventory control rather than on safety inventory allocation. De Smet et al. (2019) combine stochastic lead times with batching decisions for a distribution network based on the work of Humair et al. (2013) and calculate fill rates with an iterative procedure. More recently, Peeters (2020) accounts for MOQ to set safety stock levels and review periods integrated with stochastic lead times, based on the approach proposed by Humair et al. (2013), and using the Cycle Service Level (CSL) as a customer service measure. The present work is based on another approach proposed by Inderfurth (1993) to integrating stochastic lead times, and is flexible in the sense that it enables specifying either CSL or fill rates for each material at each location.

2. PROBLEM STATEMENT

Given a supply chain with a fixed design for a set of materials $p \in P$ that can be either raw materials or finished goods. The locations $j \in J$ belong to a set of plants, distribution centers, and retailers that can store different materials. Stock holding costs are incurred at all nodes; their unit costs

are given. We assume uncertain demand and lead times. The objective is to determine the guaranteed-service times for each material at each location, and consequently how much safety stock to maintain at each location to minimize the total holding costs and satisfy a specified customer service level. The guaranteed-service time is a variable that affects more than one location, as discussed in the background section (see Figure 1). Thus, the entire multi-echelon system must be considered for optimal decision-making.

Unlike most literature on the topic, this work does not assume there is a final customer demand zone. One the other hand, we assume that external or independent demand for any material p can be placed at any node j in the network. Each node j can have an uncorrelated independent normally distributed demand of material p with mean μ_{Ijp} and variance σ_{Ijp}. In addition, location j has an internal or dependent demand if it is required to satisfy replenishment orders from downstream nodes, which is normally distributed with mean μ_{Djp} and variance σ_{Djp}. Demand is propagated upstream considering the risk pooling assumptions described in You and Grossmann (2009). A node that satisfies both dependent and independent demands is called a hybrid node, and an example of it is presented in Figure 3. . If a node is a hybrid one, demand is propagated upstream by pooling independent and dependent demand separately.

Regarding the network topology, we assume divergent networks, as shown in Figure 3. In other words, a node that holds a material p can only receive this material from a single node and can distribute it to one or more locations, as is usual in finished goods supply chains. The same node can be supplied of another material $q \in P$ from another location, but this last one should be the only supplier of q for that node. The route that each material follows, as well as the lead time distributions between two connected nodes, are given. Lead times are assumed to follow an uncorrelated independent normal distribution with mean μ_{LTjp} and standard deviation σ_{LTjp}. They represent the delay that is under the responsibility of node j, including transportation, material handling, and other processing times until the material is ready to be shipped (i.e., is fulfilled).

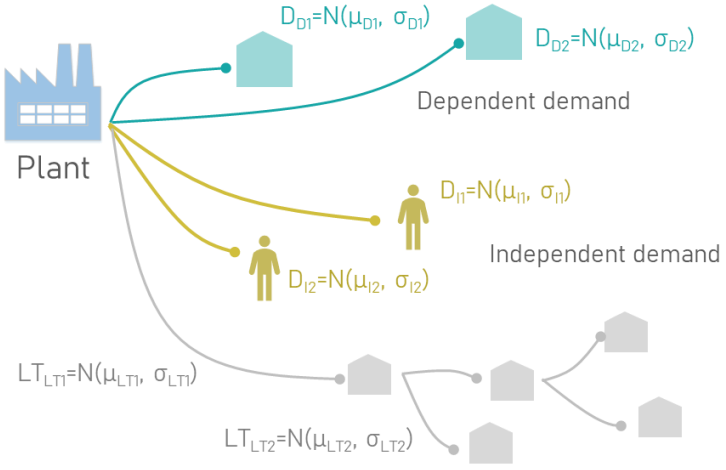

Figure 3: Example of a hybrid node and a divergent network.

Plants can be located at any node. Plant nodes can hold stock of both raw materials and finished goods. Nevertheless, the raw materials guaranteed-service time for raw materials is always required to be equal to zero to satisfy production demand immediately from stock. We introduce a general bill of materials based on a material-material relation, instead of a location-location relation as in Graves and Willems (2000). The value ϕ_{pq} determines the amount of material p required to produce a unit of material q, regardless of the plant location.

We assume a periodic review policy, with nested review periods (r_{jp}) as inputs to the model and common review days. Under a nested policy, every replenishment epoch of an upstream stage coincides with a shipment epoch towards its downstream stage (Eruguz et al., 2014). Furthermore, a minimum order quantity moq_{jp} may be enforced on replenishment orders. This means that if location j needs to place an order, it will need to order at least the moq_{jp}, which may force it to receive an amount larger than required.

The network topology is assumed to be fixed, so transportation costs are not evaluated in this study. The service time of the most upstream nodes in the network and the maximum service time of each final node are given. We assume decentralized information; therefore, each node makes its own replenishment decisions and has no delay in ordering. For each node, the safety stock factor k related to the CSL, which is represented by the standard normal distribution through the z safety factor is also given, reflecting the percentage of time that the safety stock covers the demand variation. Alternatively, the modeler can also ask for a fill rate to be considered as a target service measure.

3. MODEL FORMULATION

The multi-echelon safety inventory optimization problem can be formulated as a nonlinear program (NLP) that deals with the safety inventory planning in a given supply chain. The model proposed in Graves and Willems (2000) is used as a basis, and all sets, parameters, and variables of this model are presented in the Nomenclature section. First, we assume that external demand is propagated upstream to define internal demands through the complete network. If there are stages with more than one successor, we require a decision on how to combine the demand bounds for the downstream stages to obtain a relevant demand bound for the upstream stage to position the safety stock (Graves & Willems, 2000). There will be a relative reduction in variability as we combine demand streams due to risk pooling. Therefore, the dependent demand for material p at node j is obtained by converting the demand for all materials q that require p as an input at all successor nodes k. The conversion is done via the bill of materials ϕ_{pq} as a pre-processing step. The total demand is then the sum of the independent and dependent demands as shown in Equations (1)-(2). Note that the second term on both equations correspond to the dependent demand mean and deviation, that is μ_{Djp} and σ_{Djp}. For nodes where material p is distributed, rather than transformed into q, $p = q$ and $\phi_{pq} = 1$.

$$\mu_{jp} = \mu_{I\,jp} + \sum_{(p,q)\in\Phi} \sum_{(j,k)\in A} \phi_{pq}\mu_{kq} \qquad \forall\, j \in J, p \in P_j \tag{1}$$

$$\sigma_{jp} = \sqrt{\sigma_{I\,jp}^2 + \sum_{(p,q)\in\Phi} \sum_{(j,k,p)\in A} \phi_{pq}^2\sigma_{kq}^2} \qquad \forall\, j \in J, p \in P_j \tag{2}$$

3.1 Constraints

The first set of constraints is related to bounding the guaranteed-service time variables. Equation (3) defines the first inbound service time for the starting (source) nodes in the network J^0, where si^0 is a given input. Equation (4) links the inbound guaranteed-service time SI_{jp} and the guaranteed-service time of its upstream node S_{iq}. A is a set with elements (i,j,p) indicating that there is an enabled route for material p from i to j. Note that $q = p$ and $i \neq j$ if it is a distribution link (i to j) of the same product p, and $q \neq p$ and $i = j$ if node j is a plant location that produces p from q. If there is a maximum accepted delay for any material on a node, the inequality in (5) is active.

$$SI_{jp} = si_{jp}^0 \qquad \forall\, j \in J^0, p \in P_j \tag{3}$$

$$SI_{jp} \geq S_{iq} \qquad \forall\, (i,j,p) \in A, (q,p) \in \Phi, p \in P_j \tag{4}$$

$$S_{jp} \leq maxS_{jp} \qquad \forall\, j \in J, p \in P_j \tag{5}$$

3.1.1. Manufacturing locations

It is possible to understand the plant as two nodes connected by an arc that represents the manufacturing time. On one hand, it is required that the safety stock of raw materials is enough to satisfy production demand with no delay, requiring $maxS_{iq}$ (maximum possible service time) to be equal to zero for plant i and raw material q as shown in Figure 4. On the other hand, safety stocks of finished goods are optional. A represents an enabled production process to obtain product p at node i, and there is a production lead-time lt_{ip} to represent the manufacturing time.

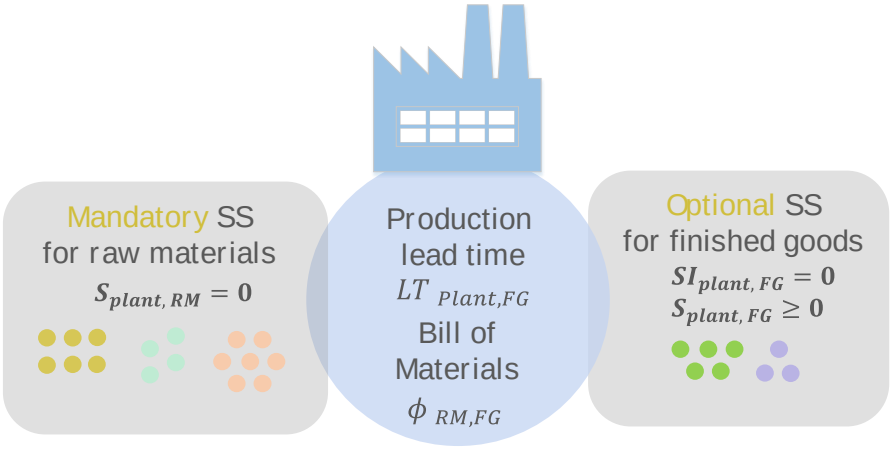

Figure 4: Service time requirements in the plant for raw materials (RM) and finished goods (FG).

3.1.2. Stochastic lead times

Concerning the incorporation of stochastic lead times to the GSM, our work is based on the approach by Inderfurth (1993). In that work, a serial network is proposed with external demand at the final nodes, called "demand nodes", and upstream nodes are called "non-demand nodes". Using the theory of single-echelon inventory optimization, Inderfurth proposes that the safety stock at a demand node be calculated as:

$$SS_{I_{jp}} = z\sqrt{NLT_{jp}\,{\sigma_{I_{jp}}}^2 + {\mu_{I_{jp}}}^2\,{\sigma_{LT_{jp}}}^2} \qquad \forall\, j \in J_p^I,\ p \in P_j \tag{6}$$

where $SS_{I_{jp}}$ represents the safety stock level to satisfy independent demand. The safety factor z multiplies the square root that involves the combination of the two random variables: independent demand ($\mu_{I_{jp}}, \sigma_{I_{jp}}$) and lead time ($lt_{jp}, \sigma_{LT_{jp}}$). At upstream stages, the stochastic lead time is converted into a deterministic lead time $\widehat{lt}_{jp} = lt_{jp} + z_{LT}\,\sigma_{LT_{jp}}$, where z_{LT} relates to the service level that denotes the probability that the lead time realization does not exceed the planned lead time $\widehat{lt}_{jp}$, and the safety stock for these non-demand nodes is calculated by:

$$SS_{D_{jp}} = z\,\sigma_{D_{jp}}\sqrt{SI_{jp} + \widehat{lt}_{jp} - S_{jp}} \qquad \forall\, j \in J_p^D,\ p \in P_j \tag{7}$$

In our work, we propose that the safety stock for each hybrid node is equal to the summation of safety stocks for independent and dependent demands, $SS_{jp} = SS_{I_{jp}} + SS_{D_{jp}}$. Hence, on each node, we define a safety stock to satisfy downstream orders and another safety stock for external orders. In general, these inventories are at the same location but have to be dedicated to each type of demand. Therefore, this limits pooling at this location. In practice, safety stock can be considered as a whole to satisfy demand if it does not mean a stockout on other customers at the same location.

3.1.3. Review periods

We assume nested review periods with common review days as in the work of Eruguz et al. (2014), and that a replenishment order is ready to satisfy the demand on its period of arrival. A node that faces external demand, needs to cover with safety stock the demand during net lead time $NLT_{jp} = SI_{jp} - S_{jp} + lt_{jp} + r_{jp}$, where r_{jp} represents the review period. The guaranteed-service approach assumes that there is no delay in placing an order and we assume that a replenishment is available to serve demand in its period of arrival. Hence, when an order is placed in a node, it is instantaneously propagated upstream. Therefore, in upstream nodes $NLT_{jp} = SI_{jp} - S_{jp} + lt_{jp} + r_{jp} - 1$, as stated in Eruguz et al. (2014). As mentioned above, we propose the alternative of hybrid nodes with both types of demand. However, we need to define unique guaranteed service times SI_{jp} and S_{jp}. Inequalities (8) and (9) account for the definition of the net lead times to be covered with safety stock to achieve the desired service level for independent and dependent demands, respectively. These equations combine review periods and the stochastic lead time approach developed above. Note that SI_{jp}, lt_{jp}, r_{jp}, and S_{jp} are assumed to be the same for both types of demands, and ARG_1 and ARG_2 are positive continuous

variables representing the terms in the square roots for the independent demand and dependent demand safety stocks, respectively. For the same values of SI_{jp} and S_{jp}, we can have different amounts of safety stocks for satisfying internal or external demands, because stochastic lead times are accounted for differently for both demand types.

$$ARG_{1\,jp} \geq SI_{jp} - S_{jp} + lt_{jp} + z\,\sigma_{LT_{jp}} + r_{jp} - 1 \qquad \forall\, j \in J^D, p \in P_j \tag{8}$$

$$ARG_{2\,jp} \geq (SI_{jp} - S_{jp} + lt_{jp} + r_{jp})\,\sigma_{I_{jp}}{}^2 + \mu_{I_{jp}}{}^2\,\sigma_{LT_{jp}}{}^2 \qquad \forall\, j \in J,\ p \in P_j \tag{9}$$

Similarly, we need to define the upper bound of S_{jp}, which occurs when the net lead time is zero. This upper bound is defined with inequalities (10) and (11). The former defines the upper bound as only for the case of pure-dependent demand nodes, while the latter one accounts for the upper bound in the case of potential hybrid nodes. The tightest of both will define S_{jp} the upper bound. Note that ub_{jp} is a parameter in the model derived from (6) and (7), which allows choosing the largest upper bound when there is a hybrid node.

$$S_{jp} \leq SI_{jp} + r_{jp} - 1 + lt_{jp} + z\,\sigma_{LT_{jp}} \qquad \forall\, j \in \left(J_p - J_p^{ID}\right),\ p \in P_j \tag{10}$$

$$S_{jp} \leq SI_{jp} + r_{jp} + lt_{jp} + ub_{jp} \qquad \forall\, j \in J_p^{ID},\ p \in P_j \tag{11}$$

$$ub_{jp} = \max\left\{z_{LT}\,\sigma_{LT_{jp}} - 1\,;\ \left(\frac{\mu_{I\,jp}}{\sigma_{I_{jp}}}\sigma_{LT_{jp}}\right)^2\right\} \tag{12}$$

3.1.4. Fill rate as a target service level

As described previously, the GSM uses the Cycle Service Level (CSL) as the customer service performance indicator when setting safety stocks. Since fill rate is more widely used in industry (Teunter et al., 2017), we extend the GSM to allow specifying fill rates if desired. Fill rates represent the fraction of demand that was met on-time from inventory. Chopra and Meindl (2013) propose a formula to obtain fill rates from a given safety stock value for continuous review policies. From this formula we can obtain the following constraint that links fill rate (*fr*) to the safety factor *z*, and consequently to the CSL:

$$fr \leq \frac{\sigma_{NLT}}{Q}\Big(z\,[1 - F_s(z)] - f_s(z)\Big) + 1 \tag{13}$$

Note that σ_{NLT} refers to the variability during net lead time, and in this work, we define $\sigma_{NLT} = \sigma_{D\,jp}\sqrt{ARG_{1\,jp}} + \sqrt{ARG_{2\,jp}}$. On the other hand, Q refers to the replenishment quantity. For periodic review inventory policies this amount is not fixed as in continuous review policies. We assume an average replenishment quantity, $Q_{jp} = \mu_{jp} r_{jp0}$. $F_s(z)$ and $f_s(z)$ correspond to the cumulative and density normal distributions functions, respectively. Therefore, the complex function $g(z) = z\,[1- F_s(z)] - f_s(z)$ needs to be included in the mathematical model. To overcome this difficulty, we propose a surrogate model through a second-order polynomial regression ($h(z) = az^2 + bz + c$) to generate an approximation

to *g(z)* in (13). The best-fit values obtained for the parameters in *h(z)* are a = -0.074700, b = 0.331986, c = -0.357195, with R^2 = 0.98. Figure 5 presents the original function $g(z)$ and the surrogate model function $h(z)$.

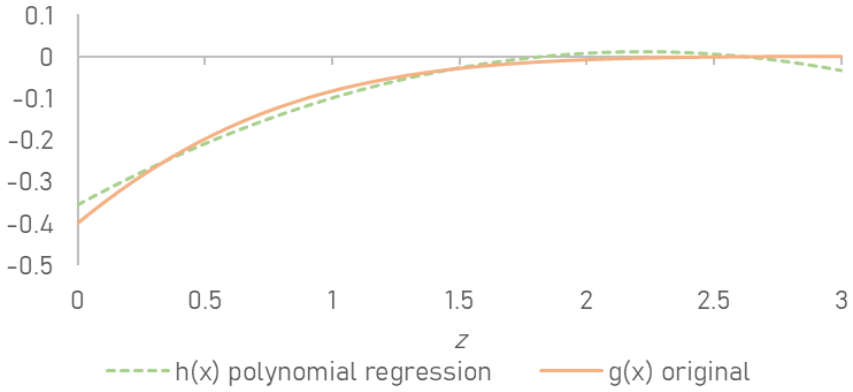

Figure 5: Surrogate model $h(z)$ and original function $g(z)$ curves.

Therefore, the constraint proposed in this model to find the minimum CSL needed to achieve the desired fill rate is given by Equation (14). Note that the safety factor z is now a continuous positive variable ZV_{jp} for those materials and locations that have fill rate levels active. The objective is to find the lowest CSL level that can meet a defined fill rate.

$$fr_{jp} \leq \frac{1}{Q_{jp}} \left(\sigma_{D\,jp} \sqrt{ARG_{1\,jp}} + \sqrt{ARG_{2\,jp}} \right) \left(-a\, ZV_{jp}^2 + b\, ZV_{jp} - c \right) + 1 \qquad (14)$$

$$\forall\, j \in J,\ p \in P_j,\ (j,p) \in F$$

3.1.5. Minimum Order Quantity (MOQ)

This requirement is frequently found in practice and studied in inventory control systems (Zhou et al., 2007; Zhu et al., 2015). Nevertheless, to the best of our knowledge, there is little literature that relates MOQ to safety inventories. When an MOQ is required, flexibility is reduced, because the customer needs to either order many units or none. However, this does not necessarily mean that the risk and safety stocks are increased. Figure 6 depicts the effect that MOQ has on inventories. Plot (a) presents inventory evolution through time for a periodic review policy with a review frequency of one week and no lead time (lt_k = 0). In grey color, we can see the safety stock level (SS_k) set to cover a proportion of the demand variability during the net lead time. The basestock level B denotes the order-up-to level that must be accounted for when a replenishment order is placed. The order quantity (Q) is equal to the demand during a review period ($\mu_{jp}\, r_{jp}$). Each replenishment cycle, that is, the time between two consecutive replenishments deliveries (Chopra & Meindl, 2013) has a probability of non-stocking out of 1 - α. The safety stock level is set to cover demand variability during the net lead time (1-α)100% of the times, this being the probability determined by the z factor.

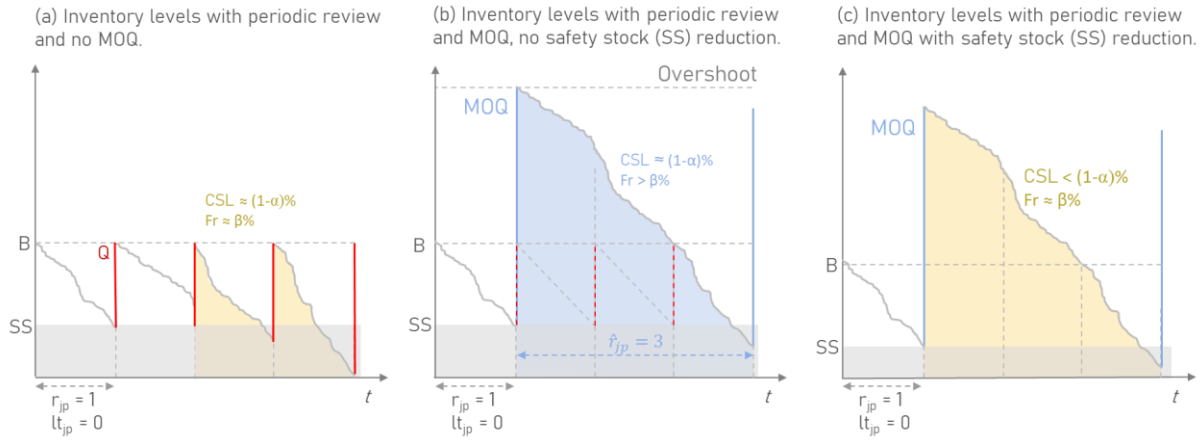

Figure 6: MOQ effect on inventory levels

If there is an MOQ required by the supplier to a node, and the MOQ is larger than the standard Q, the inventory level evolution will look like the one in Figure 6 (b). In this example, the MOQ size is three times the demand during the review period. Therefore, the first and the second periods have a low probability of stocking out, because there will be more inventory than is needed to satisfy the expected demand. However, the Cycle length is increased and a replenishment order is placed every three periods on average. The CSL measure will not be affected by the MOQ because there will be less replenishment cycles (Chopra & Meindl, 2013). On the other hand, fill rate levels will increase, which means that safety stocks can be reduced at the expense of increasing the cycle stock as a result of the MOQ requirement. The number of orders placed by the customer is not modified, and the overshoot in stock causes that many periods have more stock than necessary to fulfil the order. Safety stock levels decrease as shown in Figure 6 (c) if the MOQ is larger than the original *Q*. This reduction results in a decrease in the z factor, because now $Q_{jp} = MOQ_{jp}$ in (14).

Figure 7 depicts how fill rates are generally larger than CSL for a given value of *z*. In the green lines, it is possible to see different curves for fill rates for increasing MOQ sizes, being *MOQ1* the smaller one and *MOQ4* the larger one. The larger the MOQ is, the larger is the fill rate achieved for a specified safety factor *z*.

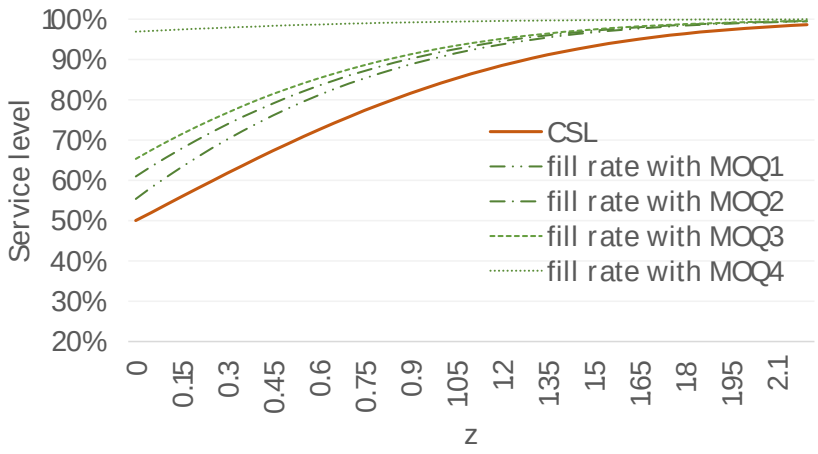

Figure 7: Fill rate sensitivity analysis with variations on replenishment quantities.

3.2 Objective Function

The objective function is to minimize safety stock holding cost as defined in Equation (15), where h_{jp} is the coefficient that represents holding cost for each material p at each location j.

$$min \sum_{j \in J} \sum_{p \in P_j} h_{jp}\, ZV_{jp} \left(\sigma_{D\,jp} \sqrt{ARG_{1\,jp}} + \sqrt{ARG_{2\,jp}} \right) \tag{15}$$

3.3 Solution Approach

The guaranteed service model, given by equations (3)-(5), (8)-(11), (14)-(15) is a nonconvex NLP with a concave objective function. Nonconvex NLP problems can in principle be solved with global optimization solvers like BARON. However, for medium or large-scale problem sizes, the computational time required to find a global solution may be very extensive. In order to improve the computational efficiency of the optimization, we propose a reformulation of the NLP model into a quadratically constrained problem (QCP), which solvers like CPLEX and Gurobi can solve quite effectively in reasonable computational times. To reformulate the problem we define a new variable Z that replaces all the square root terms in the problem, where $Z = \sqrt{\tau}$ for a general expression τ. Accordingly, the objective function of the NLP that is plotted in Figure 1 (a), will be reformulated as in Figure 1 (b), where Equation (16) is the reformulation of the objective function in (20).

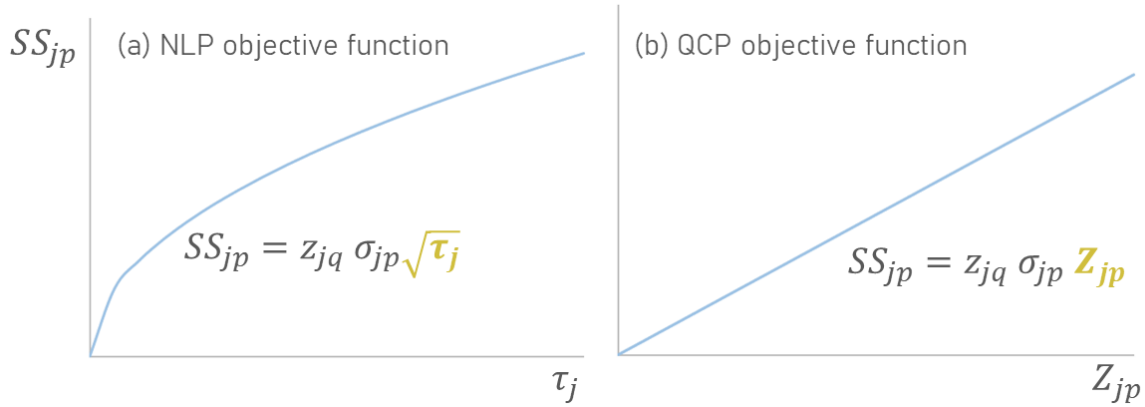

Figure 8: Objective function for NLP and QCP models.

$$min \sum_{j \in J} \sum_{p \in P_j} h_{jp}\, ZV_{jp} \left(\sigma_{D\,jp}\, Z1_{jp} + Z2_{jp} \right) \tag{16}$$

Inequalities (8) and (9) are reformulated by replacing the left-hand sides terms with variables $Z1_{jp}$ and $Z2_{jp}$, resulting in Equations (17) and (18).

$$Z1_{jp}^2 \geq SI_{jp} - S_{jp} + lt_{jp} + z\, \sigma_{LT_{jp}} + r_{jp} - 1 \qquad \forall\, j \in J^D, p \in P_j \tag{17}$$

$$Z2_{jp}^2 \geq (SI_{jp} - S_{jp} + lt_{jp} + r_{jp})\, \sigma_{I_{jp}}{}^2 + \mu_{I_{jp}}{}^2\, \sigma_{LT_{jp}}{}^2 \qquad \forall\, j \in J,\ p \in P_j \tag{18}$$

Equation (14) is replaced with Equations (19) and (20). Note that a new continuous positive variable U_{jp} replaces $ZV_{jp}{}^2$ to avoid a trilinear term as shown in Equation (19).

$$Q_{jp}(fr_{jp} - 1) \leq (-a\, U_{jp} + b\, ZV_{jp} - c)\left(\sigma_{D\,jp} Z1_{jp} + Z2_{jp}\right) \quad \forall\, j \in J,\ p \in P_j,\ (j,p) \in F \tag{19}$$

$$ZV_{jp}^2 - U_{jp} \leq 0 \quad \forall\, j \in J,\ p \in P_j,\ (j,p) \in F \tag{20}$$

The new mathematical reformulation (MQC) is a quadratically constrained program, composed of Equations (3)-(5), (10)-(11), (16)-(20).

4. APPLICATION AND RESULTS

4.1 Illustrative example

An illustrative example is presented in Figure 9 to understand the model results and how different considerations impact safety stock decisions. This case involves the production and distribution of a finished good (*SKU1*) obtained from two raw materials (*Raw1* and *Raw2*), and a Plant location that manufactures *SKU1* and delivers it to three retailers that satisfy external demand. The proportion of raw materials needed to obtain a unit of *SKU1* are $\phi_{Raw1,\,SKU1} = 1$ and $\phi_{Raw2,\,SKU1} = 0.014$. Raw materials in the plant need to have enough safety stock to satisfy production demand with no delay. Thus, $S_{jp} = 0$ for *Raw1* and *Raw2* at the *Plant*. *SKU1* storage in the *Plant* is optional, and $S_{jp} = 0$ for *SKU1* at the three retailers' locations. The production lead time is 2 weeks and it is represented by the loop above the plant. Table 1 displays the demand and lead time input data, maximum service time constraints, and unit holding costs.

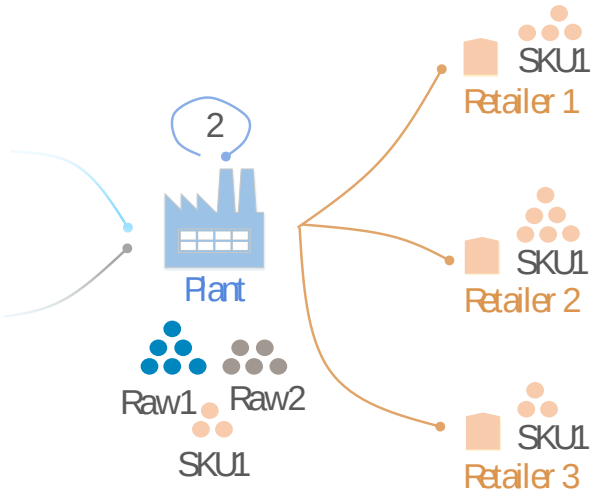

Figure 9: Illustrative example representation

Table 1: Illustrative example input data

Material		Raw_1	Raw_2	SKU_1	SKU_1	SKU_1	SKU_1
Location		*Plant*	*Plant*	*Plant*	$Retailer_1$	$Retailer_2$	$Retailer_3$
Demand (units)	μ_{jp}	425,717	5,913	425,717	162,379	67,284	196,054
	σ_{jp}	192,229	2,669	192,229	119,665	61,585	137,258
Coefficient of Variation (CV= σ_{jp} /μ_{jp})		0.45	0.45	0.45	0.73	0.91	0.70
Lead Time (weeks)	μ_{LTjp}	6	3	2	1	1	1
	σ_{LTjp}	1.9	0.7	0.0	0.3	0.6	0.4
Max Service Time S_{jp} (weeks)		0	0	-	0	0	0
h_{jp} ($/unit)		0.01171	0.00002	0.12	0.12	0.12	0.12

Results are detailed in Table 2. The computational tests are performed on an Intel® Core i7 CPU with 8 GB RAM and 4 parallel threads using Gurobi 9.1.2 as the QCP solver. The model (MQC) involves 34 continuous variables and 30 constraints. The CPU time required to obtain the optimal solution is less than 1 second and the total holding cost obtained is $13,181. It is possible to see that in the plant the decision is not to hold safety stock, and to select a guaranteed service of 2 weeks for supplying the retailers.

Table 2: Illustrative example results

Material	Raw_1	Raw_2	SKU_1	SKU_1	SKU_1	SKU_1
Location	*Plant*	*Plant*	*Plant*	$Retailer_1$	$Retailer_2$	$Retailer_3$
S_{jp} (weeks)	0	0	2	0	0	0
SS_{jp} (units)	1,125,289	10,600	0.0	459,359	243,782	536,961
Holding $cost_{jp}$ ($/unit)	13,181	0.2	0.0	55,123	29,253	64,435

It is worth mentioning that the guaranteed service time of SKU_1 in *Plant* affects the retailers' safety stock levels, which need to cover for 2 more weeks with stock, as this is the inbound service time ($SI_{Retauler,SKU1}$ = 2 weeks). If this inbound service time continues to increase by relaxing the service time constraint on the raw materials or increasing the production time, the safety stock at the retailers will also increase, and it is possible that the model decides to hold safety stock upstream in the plant so as to take advantage of system-wide risk-pooling, and have a lower total safety stock in the supply chain. Figure 10 (A) depicts the current case, with a production lead time of 2 weeks. The total safety stock in the supply chain is 2,092,387 units with a cost of $154,145. In case (B), the production lead time is increased to 10 weeks, and the optimal solution for this case is to hold stock of SKU_1 (1,143,300 units) in the plant. The total holding costs increase to $249,232 because there is more time to cover. However,

the decision of pooling in the plant yields a lower cost than if we maintained the decision of no safety stock in the plant for *SKU1*. In (C) we present the solution of case (B) if no safety stock for SKU_1 in the plant is allowed. In this last case, the total cost is increased to $259,393 in comparison to the solution in (B) because the opportunity of pooling and reducing the inbound service time for retailers is missed.

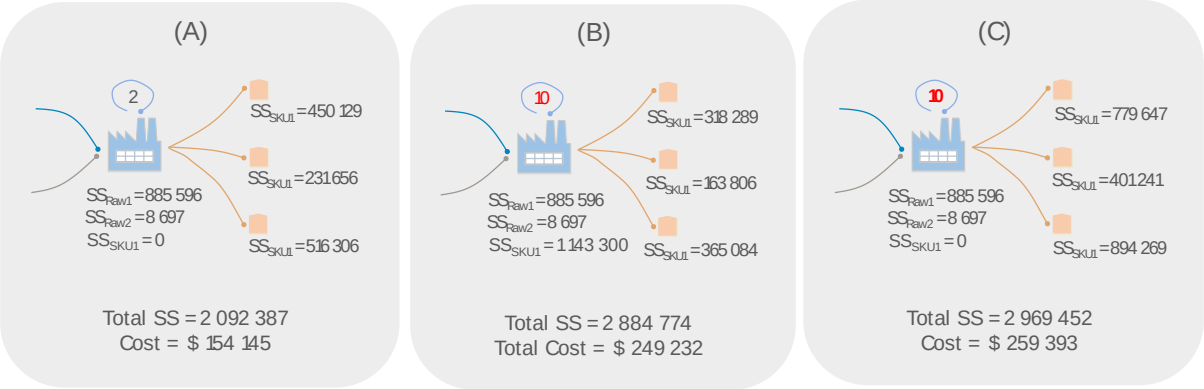

Figure 10: Safety stock levels for different lead time cases: A) 2-week production lead time, B) 10-week production lead time, C) 10-week production lead time with no safety stock for *SKU1* at the *Plant*

For this illustrative example, there is no MOQ requirement. Therefore, we assume that for a periodic review policy, the expected order size (Q_{jp}) is equal to the demand during the review period, that is $\mu_{jp}\, r_{jp}$. The last analysis of this illustrative example concerns the measurement of customer service. In the current case, there is a desired 97% CSL, which corresponds to a safety factor z of 1.88. If a location, for example, *Retailer 1*, has to change the customer service metric from CSL to fill rate for a given product, a different CSL can be required to achieve the expected fill rate, so z can change its value. In Figure 11, CSL and safety stocks are given for different cases varying target fill rates and MOQ constraints for *SKU1* on *Retailer 1*. The blue line indicates the expected fill rate, which is a given input in all cases except in the first one, in which the CSL is defined to set safety stocks as the original example, and the fill rate in this case is obtained through (19).

The following scenarios consist of target fill rates and the CSL is obtained by the model. The yellow dashed line represents the resulting CSL for each case, and the yellow bar is the correspondent safety stock (secondary vertical axis) for that z coverage. In addition, the brown dashed line and brown bars represent the resulting CSL and safety stock when an MOQ constraint is active. The first case (the left-most case) is the current illustrative example scenario. The desired CSL is 97%, and a near 100% fill rate is expected, with or without a required MOQ. However, safety stock levels decrease when a large MOQ (500,000 units) is active. The second case sets a 98% fill rate to define safety stock levels. The minimum required CSL to achieve this expected fill rate decreases together with its corresponding safety stock level, and a sharper decrease occurs when a large MOQ is required. In the subsequent scenarios, the desired fill rates decrease and consequently, the CSL is lower. This difference is even more remarkable in the presence of MOQs, having no safety stocks defined for fill rates less than or

equal to 80%, with a minimum required CSL of 50%. In general, CSLs are lower than fill rates for a given safety stock level.

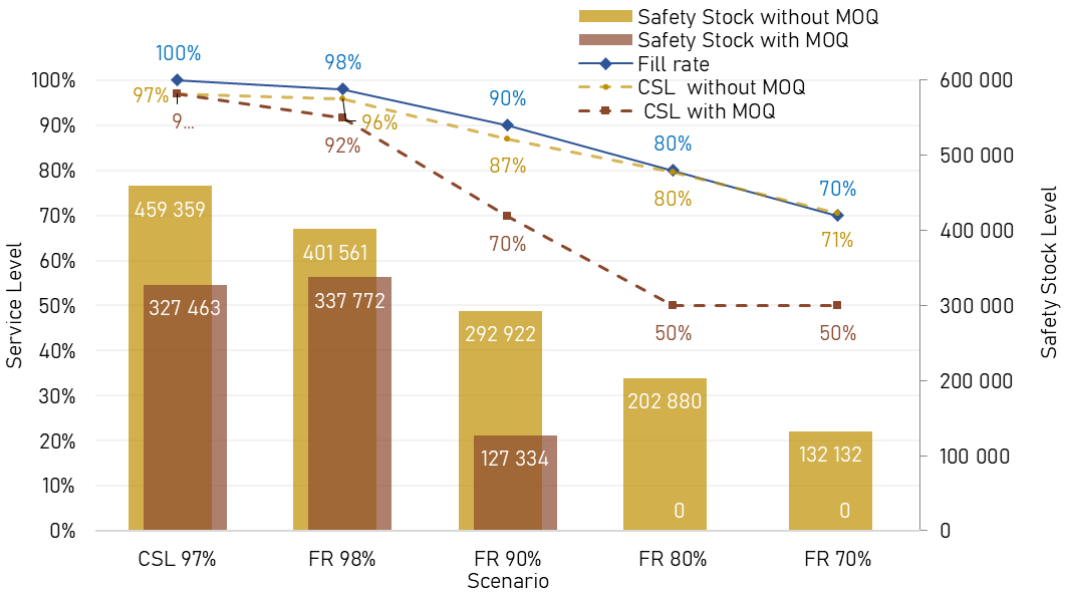

Figure 11: Effect of fill rate and MOQ on CSL and safety stock levels.

4.2 Small-size industrial case study

The MQC formulation is now applied to a small-size industrial case study (Figure 12), with two echelons, 4 SKUs, and 31 raw materials coming from different locations. The first echelon has one plant and the second echelon has three retailers, as in the illustrative example presented above. The input data is presented in Table S1 in the Supporting Information Section. Note that lead times have decimals because they are averages of historic data. For MEIO purposes, the ceiling of the lead time is $\lceil lt_{jp} \rceil$ used as input.

Figure 12: Small-size industrial case

The MQC model has 248 continuous variables and 280 constraints. While the NLP formulation using BARON is not able to find a feasible solution within 1000 seconds, the proposed QCP formulation finds the optimal solution using Gurobi in 0.03 seconds. The optimal solution is \$185,185 with S_{jp}=0 for raw materials in the plant and SKUs at the retailers as required, and S_{jp}=1 for SKUs in the plant. The results were compared with commercial software (not identified due to confidentiality reasons), and the current safety stock levels in the plant and costs at each location are summarized in Figure 13. Total holding costs and safety stocks levels for raw materials and finished goods are presented in Table 3. While the vendor software obtained a 10% reduction in holding costs regarding the current safety stock levels, the proposed model in this work yields a 17% reduction, clearly showing the advantage of this tool to achieve customer service levels with a minimum capital in inventory. Note that safety stocks at the retailers are slightly larger with the proposed model. A possible reason for this is that we use the ceiling value of the lead times to cover stock over discrete periods. On the other hand, the model seems to reduce the amount of inventory of raw materials, yielding the largest reduction in holding costs.

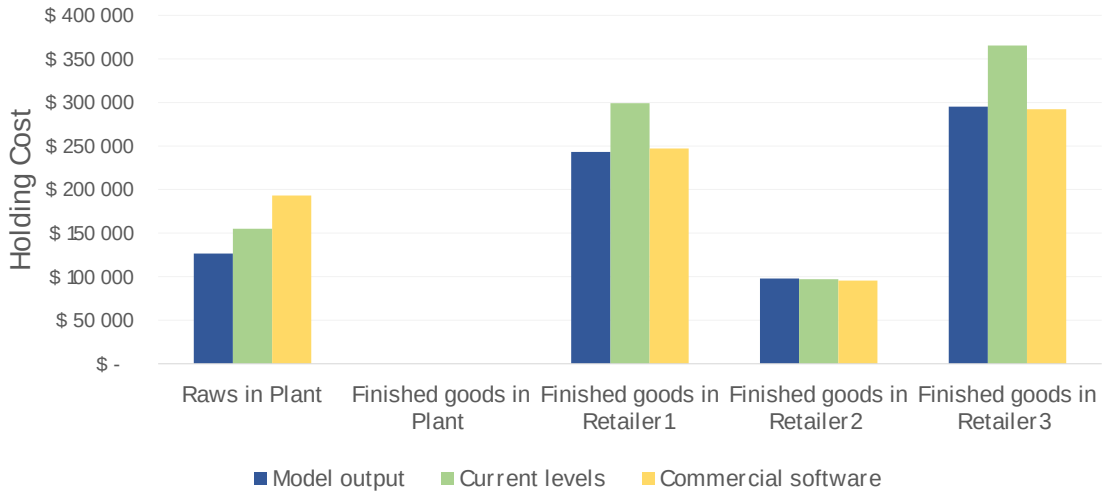

Figure 13: Small-size industrial safety stock results of the proposed tool, the commercial software, and current levels in the supply chain

Table 3: Small-size industrial safety stock levels and holding costs

		Model output	Baseline (Current level)	Commercial software
Holding cost	Raw materials	\$ 126,532	\$ 155,070	\$ 193,080
	Finished goods	\$ 635,971	\$ 761,525	\$ 634,950
Safety stock level	Raw materials	8,031,692	9,917,801	13,029,797
	Finished goods	2,740,254	3,042 031	2,685,615

4.3 Medium-size industrial case study

This case involves the same network as in Figure 9, but now it has 20 finished goods and 120 raw materials, requiring 196 safety stock decisions. The MOQs constraints (Equations (19) and (20)) are active for all nodes and materials, and the customer service measure of interest is the fill rate, which is different for each material. The size of the QCP model is 1,973 constraints and 1,427 continuous variables, and is solved to optimality within 3 seconds using Gurobi as the QCP solver. This further supports the usefulness of the proposed approach for solving real-world problems.

5. VALIDATION THROUGH SIMULATION

All results obtained from the developed model are validated using simulation. The aim is to evaluate if safety stocks can meet expected customer service levels (CSL and/or fill rate). This is done using an open-sourced discrete-time inventory simulation software package written in the Julia language: *InventoryManagement.jl* (Perez, 2021). This simulator allows modeling multiproduct supply networks of any topology (e.g., serial, divergent, convergent, tree, or general). Each of the features included in the extended GSM model can be simulated using this software: hybrid nodes, MOQ, bill of materials, stochastic demand, and stochastic lead times. The software allows evaluating any static (continuous or periodic review) and dynamic inventory control policy. For greater clarity, the validation of the illustrative example is presented with two extra scenarios to analyze how some features affect the system behavior in the simulation.

Demand and lead times are randomly generated using normal distributions for each period. Basestock levels are calculated following Equation (21). Note that the safety stock level SS_{jp} includes both safety stocks to satisfy dependent and independent demands at every location. The two following terms refer to the previously mentioned expected demands at each location during the Net Lead Time.

$$B_{jp} = SS_{jp} + \mu_{D_{jp}}\left(SI_{jp} - S_{jp} + \left\lceil lt_{jp} + z\,\sigma_{LT_{jp}}\right\rceil + r_{jp} - 1\right) + \mu_{I_{jp}}\left(SI_{jp} - S_{jp} + \left\lceil lt_{jp}\right\rceil + r_{jp}\right) \quad \forall\, j \in \left(J_p - J_p^{ID}\right),\ p \in P_j \tag{21}$$

In each period, demand is generated, orders are placed and delivered, and inventory levels are updated. During each review period, an order is placed if the inventory position of a material on a location is below the basestock level. The information is known at all levels but the management policy is decentralized: each location asks the amount they need to reach the basestock level, no matter how much the upstream node has on stock. If the available inventory is not able to meet demand, lost sales are considered (the extraordinary measures that are referred to in the GSM approach are ignored in the simulation). In addition, there is no delay in ordering, so all echelons place orders simultaneously. The period selected in this case is one day with 7,000 days (1,000 weeks) in each run, so as to simulate the stationary state at each location. Demand and review periods have a weekly basis. The sequence of steps in the simulation of each period is the following:

1. External demand is observed and discounted at each node with independent demand. Unfulfilled demands are marked as lost sales.
2. If the review period has been meet for a material, internal replenishment and lead times lt_{jp} are observed. Orders start being processed immediately with no delay (processing time includes material handling, order preparation, shipments and/or manufacture). Internal demand is discounted at each node. Unfulfilled replenishment orders are marked as lost sales.
3. Stocks are updated with the replenishment orders that arrive at each node (placed lt_{jp} periods before).

The illustrative case study was run and two additional scenarios were also tested, combining fill rates as a target measure and MOQs for finished goods at retailers. The results of each of the scenarios are shown in the following subsections.

5.1 Scenario 1: Illustrative example with 97% CSL as target

The first scenario simulates the results of the illustrative example presented in Section 4.1, with a 97% CSL target for every material/location combination and no MOQs required. Table 4 presents the safety stocks given by the model output and the basestock levels obtained.

Table 4: Input data for simulation of Scenario 1

Location	Material	Safety Stock level	Basestock level	*z* factor	Expected CSL
Plant	SKU_1	511,300	1,362,734	1.88	97%
Plant	raw_1	1,116,603	5,373,773	1.88	97%
Plant	raw_2	10,530	40,094	1.88	97%
Retailer1	SKU_1	331,214	655,972	1.88	97%
Retailer2	SKU_1	180,548	315,116	1.88	97%
Retailer3	SKU_1	393,753	785,861	1.88	97%

The simulation is run for 1,000 weeks. Figure 14 displays on-hand inventory and inventory position for each material at each location through the first 50 weeks (350 days) for greater clarity. While inventory levels are presented in dark colors, inventory positions are displayed in light colors. Calculated CSL and fill rates from the simulation output are shown in Figure 15.

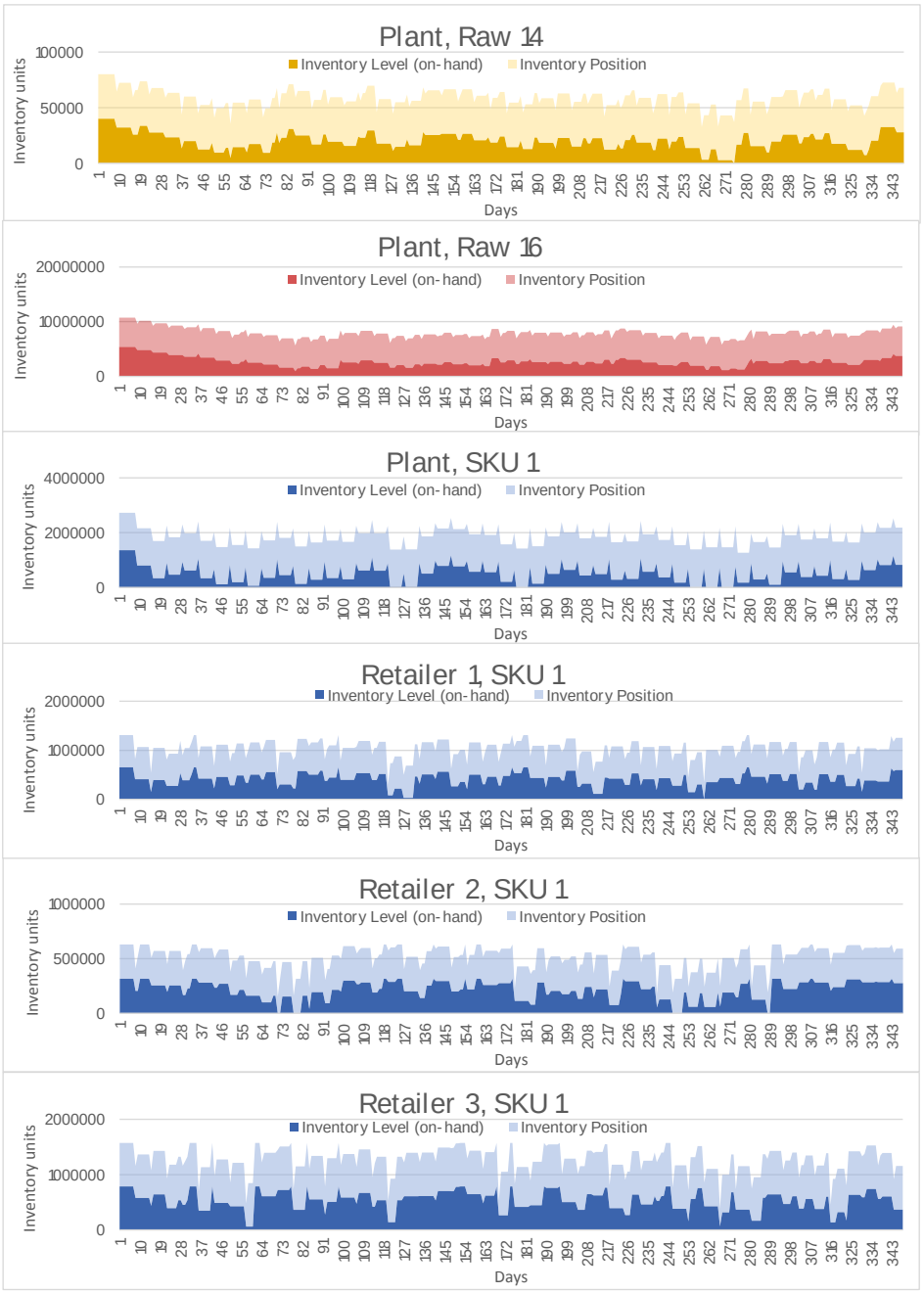

Figure 14: Inventory levels and inventory positions from the simulation of Scenario 1

It is possible to see that materials at all locations meet the expected 97% CSL approximately, with fill rates that in general surpass the CSLs. Service levels can be larger than expected because the approach has some conservative measures, like using the ceiling value of lead times for both basestock and safety stocks levels, and the stochastic approach proposed by Inderfurth also increases the mean lead time for the basestock level and pipeline inventory. Therefore, we consider that the safety stocks set for this case are able to meet the desired customer service levels.

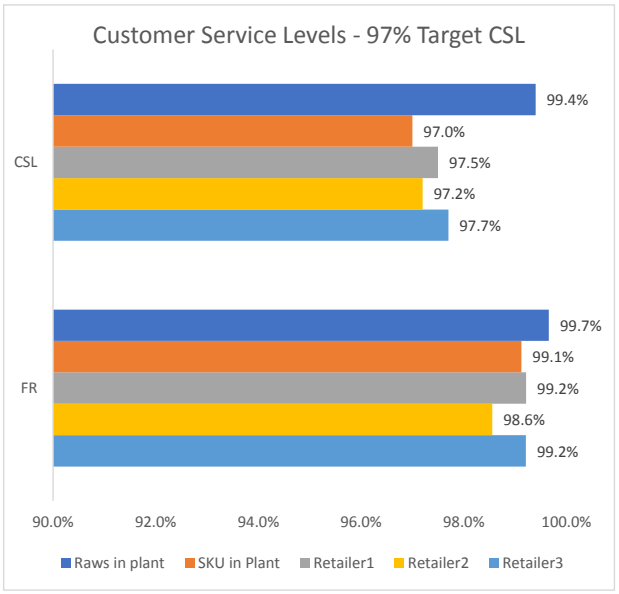

Figure 15: CSL and fill rates obtained from Scenario 1 simulation output.

5.1 Scenario 2: Illustrative example with 90% target fill rate

In Scenario 2, the service level target is changed to a 90% expected fill rate for all materials. Table 5 presents the basestock level and the minimum z factor required to meet the 90% fill rate, obtained from variable *ZV* in Equations (19) and (20) in the optimization model. Note that all safety factors, and thus expected CSLs, are strongly reduced, showing that a lower CSL is enough to meet the desired fill rate. In Figure 16 it is possible to verify that the resulting CSL is approximately the expected one, with some expected differences due to the quadratic approximation proposed to obtain z. For example, $z_{Retailer1,SKU1}$ = 1.04 is obtained using the surrogated model $h(x)$ displayed in Figure 5. It is possible to detect a small difference between the polynomial approximation $h(x)$ and the original formula $g(x)$, which indicates $z_{Retailer1,SKU1}$ = 0.95 to reach a 90% fill rate. A larger safety factor z yields a slightly larger CSL (+2%) and fill rates (+5%).

Table 5: Input data for simulation of Scenario 2

Location	Material	Safety Stock level	Basestock level	Expected fill rate	z factor	Expected CSL
Plant	SKU_1	196,088	1,047,522	90%	0.72	76%
Plant	raw_1	643,590	4,956,714	90%	1.17	83%
Plant	raw_2	5,012	35,048	90%	0.97	88%
Retailer1	SKU_1	183,666	508,424	90%	1.04	85%
Retailer2	SKU_1	112,890	247,458	90%	1.18	88%
Retailer3	SKU_1	216,640	608,748	90%	1.03	85%

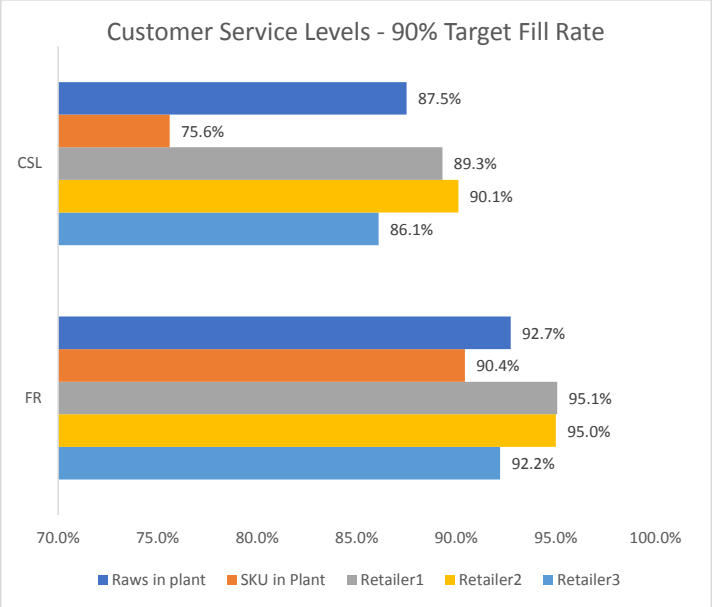

Figure 16: CSL and fill rates obtained from Scenario 2 simulation output.

5.2 Scenario 3: Illustrative example with 90% target fill rate and MOQ for retailers

This scenario adds a minimum order quantity of 500,000 units to Scenario 2. This minimum order quantity is required by the plant, assuming that this is the minimum necessary batch size to deliver an order from the plant to a retailer. In this case, we will only validate the results on the retailers. It is assumed that demand upstream will increase according to the MOQ size because the plant will need enough inventory to deliver at least three orders of 500,000 units simultaneously to the retailers. In this case, expected CSL (and *z* value) is reduced because the fill rate is chosen and also because of the MOQ as discussed in Section 3.1.5. Table 6 shows the values of the safety factors (*z*) obtained in the model output and the expected CSL to obtain the given target fill rate considering the MOQ constraint.

Table 6: Input data for simulation of Scenario 3

Location	Material	Safety Stock level	Basestock level	Expected fill rate	*z* factor	Expected CSL
Retailer1	SKU_1	41,016	365,774	90%	0.23	59%
Retailer2	SKU_1	0	134,568	90%	0.00	50%
Retailer3	SKU_1	81,842	473,950	90%	0.39	65%

Figure 17 presents the resulting performance indicator values. Note that fill rates are achieved with very low safety stock levels and no safety stock is necessary to account for the fill rate target at *Retailer2*. The MOQ effect can be detected on inventory levels through time in Figure 18, that presents inventory levels for Retailer 2. Retailer 2 has the lowest demand, and therefore the MOQ is able to cover more review periods, 7.4 weeks on average. In the zoomed-in rectangle, it is possible to detect that an MOQ can cover several weeks.

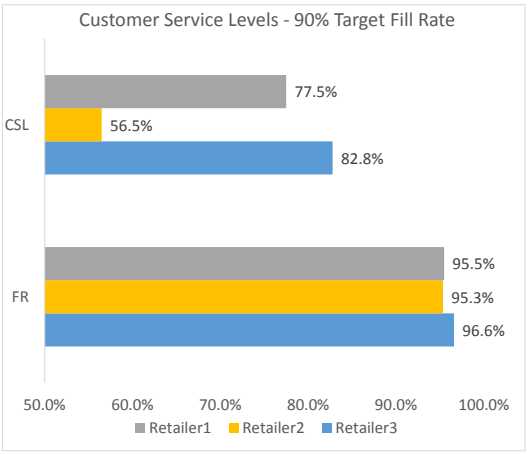

Figure 17: CSL and fill rates obtained from Scenario 3 simulation output.

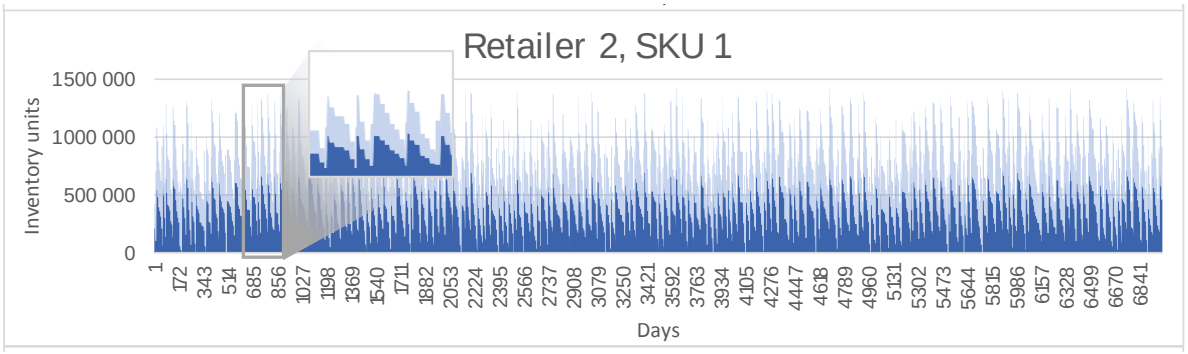

Figure 18: Inventory levels and inventory positions from the simulation of Scenario 3

It is also remarkable that the obtained CSL levels from simulation are significantly larger than those expected, and this also impacts the resulting fill rates. This agrees with the analysis by Peeters (2020), which states that the guaranteed service model becomes less accurate for lower service levels and also states that batching in a guaranteed service context leads to an excessive amount of inventory and significantly higher service levels. A simple simulation used to illustrate this is detailed in the Supporting Information section and the summarized results are presented in Figure 19. The simulations show that the safety stock settings tend to be more accurate for large expected CSLs in a system with lost sales. Lower values of expected CSL tend to be less accurate, yielding larger effective CSL's. When the coefficient of variation (CV) and the expected CSL are low, the CSL is generally underestimated and this effect is amplified when there are large MOQs active. In practice, target CSL's are generally larger than 90%, so approximations tend to be accurate in that range. Nevertheless, the equivalent CSL resulting from selecting a fill rate as target measure brings the opportunity to largely decrease *z* safety factors, and the gap in CSL estimations arises. Customer service is generally ensured while reductions in safety stocks can be performed. Future research may explore how safety stocks or basestock levels can be adjusted to deal with CV and MOQ to obtain more precise estimations. In conclusion, the validation of results through simulation allows demonstrating the accuracy of the model to obtain safety stocks that can meet specified customer service levels.

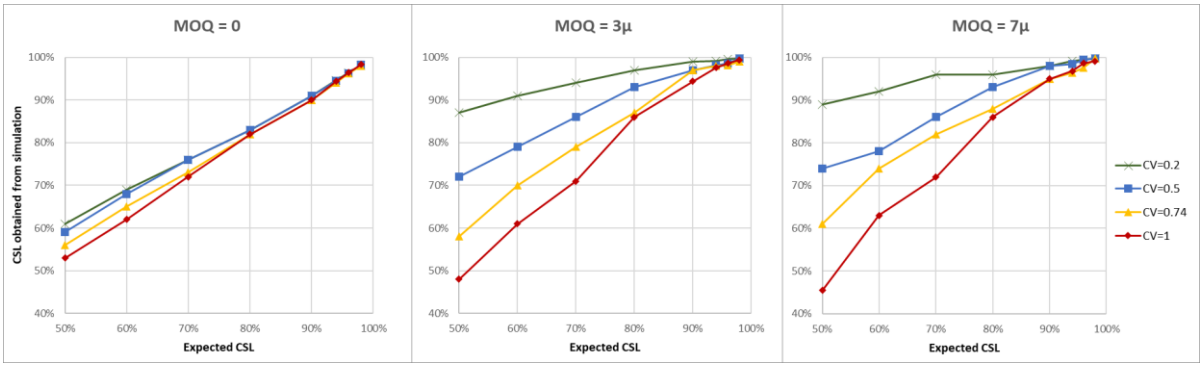

Figure 19: Effective vs. Expected CSL obtained for MOQ = 0 (left), MOQ = 3μ (middle), and MOQ = 7μ (right) and for different CVs.

6. CONCLUSIONS

In this paper, we have presented an NLP optimization model based on the guaranteed-service approach model that determines the optimal safety stock allocation in multi-echelon divergent networks. The proposed model integrates several features commonly found in industrial practice that have a strong impact on inventory levels. Therefore, the proposed tool is able to accurately represent real systems, and to set tight safety stock levels in order to achieve target customer service levels with a minimum capital in inventory. Real-world examples from the pharmaceutical industry are presented to illustrate the applicability of the proposed formulation. To solve the model efficiently, a QCP reformulation is proposed by exploiting the mathematical structure of the constraints. The QCP outperforms the NLP formulation by allowing the use of QCP solvers, which leads to order of magnitude reductions in computational time. Optimal solutions can be found with small computational expense for medium/large scale problems (less than 5 seconds). To the best of our knowledge, this is the first model that brings together multiple features typical of industrial practice, such as MOQs, hybrid nodes, and alternative service level measures to determine safety stock levels. It is also the first model to introduce the QCP reformulation to improve the computational efficiency of the optimization. The simulation of the results demonstrates that the model is valid for achieving target service levels.

Future work will address an extension of the present formulation for cases of non-normal demand, and a pre-processing procedure of input data in order to decide which mathematical formulation is appropriate to optimally determine safety stock levels. The effects of CV and MOQ on CSL estimation can also be analyzed to review other potential safety stocks reductions. This research can also be extended by including responsive characteristics to account for supply chain disruptions and by including storage capacity limitations. Another important extension is on constrained capacity on nodes.

7. NOMENCLATURE

7.1 Sets

J	Set of locations
P	Set of products
P_j	Subsets of products that can be stored at location *j*
J_p	Subsets of locations in the route of material *p*
J^0	Subset of starting locations in the network
J^I	Subset of locations that face external demand
J^D	Subset of locations that face internal demand
A	Subset of routes segments (from node *i* to node *j*) enabled for material *p*
F	Set of locations that have materials with an active fill rate as a target
Φ	Set of all valid material transformations (from material p to material q)

7.2 Parameters

μ_{jp}	Mean of the total demand of material *p* in location *j*
σ_{jp}	Standard deviation of the total demand of material *p* in location *j*
$\mu_{D\,jp}$	Mean of the dependent demand of material *p* in location *j*
$\sigma_{D\,jp}$	Standard deviation of the dependent demand of material *p* in location *j*
$\mu_{I\,jp}$	Mean of the independent demand of material *p* in location *j*
$\sigma_{I\,jp}$	Standard deviation of the independent demand of material *p* in location *j*
lt_{jp}	Lead time/order processing time of material *p* in location *j*
$\sigma_{LT\,jp}$	Standard deviation of the lead time/order processing time of material *p* in location *j*
h_{jp}	Holding cost of material *p* in location *j*
si_{jp}^0	Inbound service time for the source nodes in the network
ϕ_{pq}	Amount of material *p* required to produce material a unit of material *q*
ub_{jp}	Auxiliary parameter to calculate S_{jp} upper bound constraint according to comparisons in lead time variability
$maxS_{jp}$	Maximum service time accepted for material *p* in location *j*
r_{jp}	Stock review period for material *p* in location *j*
moq_{jp}	Minimum Order Quantity of material *p* that location *j* must place
Q_{jp}	Replenishment order size of material *p* at location *j*
fr_{jp}	Fill rate level of material *p* at location *j*
z_{jp}	Safety factor associated with CSL of material *p* at location *j*

7.3 Positive Variables

S_{jp}	Guaranteed service time within which location *j* will attend demand of material *p*
SI_{jp}	Inbound Guaranteed service time at location *j* of material *p*
$ARG_{1\,jp}$	Argument of square root for independent demand of material *p* at node *j*
$ARG_{2\,jp}$	Argument of square root for dependent demand of material *p* at node *j*
NLT_{jp}	Net Lead time of material *p* at node *j*
$Z_{1\,jp}$	Variable used for quadratic reformulation on dependent demand net lead time formula
$Z_{2\,jp}$	Variable used for quadratic reformulation on independent demand net lead time formula
ZV_{jp}	Variable used to replace *z* input factor when the fill rate is introduced to determine safety stocks
U_{jp}	Variable defined to replace ZV_{jp}^2 and avoid trilinear terms

8. ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support from Johnson and Johnson, the Fulbright Program and the Ministerio de Educación de Argentina, and the Center for Advanced Process Decision-making (CAPD) from Carnegie Mellon University. We would also like to thank Kyle Harshbarger for his useful comments during EWO meetings in CMU and Alev Kaya for her valuable insights on MEIO.

9. REFERENCES

Bendadou, A., Kalai, R., Jemai, Z., & Rekik, Y. (2021). Impact of merging activities in a supply chain under the Guaranteed Service Model: Centralized and decentralized cases. *Applied Mathematical Modelling*, *93*, 509–524. https://doi.org/10.1016/j.apm.2020.12.024

Chopra, S., & Meindl, P. (2013). Supply Chain Management. In *Pearson*. http://www.doiserbia.nb.rs/Article.aspx?ID=0013-32640670067A

de Kok, T., Grob, C., Laumanns, M., Minner, S., Rambau, J., & Schade, K. (2018). A typology and literature review on stochastic multi-echelon inventory models. *European Journal of Operational Research*, *269*(3), 955–983. https://doi.org/10.1016/j.ejor.2018.02.047

De Smet, N., Aghezzaf, E.-H., & Desmet, B. (2019). Optimising installation (R,Q) policies in distribution networks with stochastic lead times: a comparative analysis of guaranteed- and stochastic service models. *International Journal of Production Research*, *57*(13), 4148–4165. https://doi.org/10.1080/00207543.2018.1518606

Eruguz, A. S., Jemai, Z., Sahin, E., & Dallery, Y. (2014). Optimising reorder intervals and order-up- to levels in guaranteed service supply chains. *International Journal of Production Research*, *52*(1),

149–164. https://doi.org/10.1080/00207543.2013.831188

Eruguz, A. S., Sahin, E., Jemai, Z., & Dallery, Y. (2016). A comprehensive survey of guaranteed-service models for multi-echelon inventory optimization. *International Journal of Production Economics*, *172*, 110–125. https://doi.org/10.1016/j.ijpe.2015.11.017

Gonçalves, J. N. C., Sameiro Carvalho, M., & Cortez, P. (2020). Operations research models and methods for safety stock determination: A review. *Operations Research Perspectives*, *7*(April), 100164. https://doi.org/10.1016/j.orp.2020.100164

Graves, S. C., & Willems, S. P. (2000). Optimizing Strategic Safety Stock Placement in Supply Chains. *Manufacturing & Service Operations Management*, *2*(1), 68–83. https://doi.org/10.1287/msom.2.1.68.23267

Graves, S. C., & Willems, S. P. (2003). Supply Chain Design: Safety Stock Placement and Supply Chain Configuration. In *Handbooks in Operations Research and Management Science* (Vol. 11, Issue C, pp. 95–132). https://doi.org/10.1016/S0927-0507(03)11003-1

Humair, S., Ruark, J. D., Tomlin, B., & Willems, S. P. (2013). Incorporating Stochastic Lead Times Into the Guaranteed Service Model of Safety Stock Optimization. *Interfaces*, *43*(5), 421–434. https://doi.org/10.1287/inte.2013.0699

Inderfurth, K. (1993). Valuation of Leadtime Reduction in Multi-Stage Production Systems. In *Operations Research in Production Planning and Control*. https://doi.org/10.1007/978-3-642-78063-9

Inderfurth, K., & Minner, S. (1998). Safety stocks in multi-stage inventory systems under different service measures. *European Journal of Operational Research*, *106*(1), 57–73. https://doi.org/10.1016/S0377-2217(98)00210-0

Klosterhalfen, S. T., Dittmar, D., & Minner, S. (2013). An integrated guaranteed- and stochastic-service approach to inventory optimization in supply chains. *European Journal of Operational Research*, *231*(1), 109–119. https://doi.org/10.1016/j.ejor.2013.05.032

Magnanti, T. L., Max Shen, Z.-J., Shu, J., Simchi-Levi, D., & Teo, C.-P. (2006). Inventory placement in acyclic supply chain networks. *Operations Research Letters*, *34*(2), 228–238. https://doi.org/10.1016/j.orl.2005.04.004

Minner, S. (1998). *Strategic Safety Stocks in Supply Chains*.

Park, J. H., Kim, J. S., & Shin, K. Y. (2018). Inventory control model for a supply chain system with multiple types of items and minimum order size requirements. *International Transactions in Operational Research*, *25*(6), 1927–1946. https://doi.org/10.1111/itor.12262

Payne, T. (2016). Key Principles for Implementing MEIO to Cope With Supply Chain Variability. In *Gartner*.

Peeters, Y. A. M. (2020). *Safety stock setting in the global supply chain of a life science company the comparison of a guaranteed service model and stochastic service model* [Eindhoven University of Technology MASTER]. https://research.tue.nl/en/studentTheses/safety-stock-setting-in-the-

global-supply-chain-of-a-life-science

Perez, H. D. (2021). *hdavid16/InventoryManagement.jl: v0.3.2 (v0.3.2)*. Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.5725543

Shen, K., David, J., De Pessemier, T., Martens, L., & Joseph, W. (2019). An efficient genetic method for multi-objective continuous production scheduling in Industrial Internet of Things. *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA*, *2019-Septe*, 1119–1126. https://doi.org/10.1109/ETFA.2019.8869049

Silver, E. A., & Bischak, D. P. (2011). The exact fill rate in a periodic review base stock system under normally distributed demand. *Omega*, *39*(3), 346–349. https://doi.org/10.1016/j.omega.2010.08.003

Simchi-Levi, D., & Zhao, Y. (2012). Performance Evaluation of Stochastic Multi-Echelon Inventory Systems: A Survey. *Advances in Operations Research*, *2012*, 1–34. https://doi.org/10.1155/2012/126254

Simpson, K. F. . (1958). In-Process Inventories. *INFORMS*, *6*(6), 863–873.

Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2017). Stock keeping unit fill rate specification. *European Journal of Operational Research*, *259*(3), 917–925. https://doi.org/10.1016/j.ejor.2016.11.017

You, F., & Grossmann, I. E. (2008). Mixed-Integer Nonlinear Programming Models and Algorithms for Large-Scale Supply Chain Design with Stochastic Inventory Management. *Industrial & Engineering Chemistry Research*, *47*(20), 7802–7817. https://doi.org/10.1021/ie800257x

You, F., & Grossmann, I. E. (2009). Integrated multi-echelon supply chain design with inventories under uncertainty: MINLP models, computational strategies. *AIChE Journal*, *59*(4), NA-NA. https://doi.org/10.1002/aic.12010

Zhou, B., Zhao, Y., & Katehakis, M. N. (2007). Effective control policies for stochastic inventory systems with a minimum order quantity and linear costs. *International Journal of Production Economics*, *106*(2), 523–531. https://doi.org/10.1016/j.ijpe.2006.06.020

Zhu, H., Liu, X., & Chen, Y. (2015). Effective inventory control policies with a minimum order quantity and batch ordering. *International Journal of Production Economics*, *168*, 21–30. https://doi.org/10.1016/j.ijpe.2015.06.008