sustainability

MDPI

Article

Research on Method for Multi-Category, Multi-Stage Delayed Shipment Consolidation Under a Differentiated Compensation Mechanism

Lei Zhou [1], Xiaonan Sun [2] and Tao Liu [2,*]

[1] School of Electrical Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China
[2] School of Economics and Management, Hebei University of Science and Technology, Shijiazhuang 050018, China
* Correspondence: hebustlt@163.com

Abstract

Improving transportation resource efficiency while meeting delivery service requirements is a key challenge in sustainable logistics. Shipment consolidation can increase capacity utilization; however, waiting for shipments to accumulate may lead to delivery delays and additional compensation costs. To address this trade-off, this paper develops a multi-category, multi-stage stochastic programming model for delayed shipment consolidation. The model captures the relationships among stochastic demand, dispatch decisions, and subsequent distribution states through the recursive evolution of interstage backlogs. Because the number of state–strategy combinations grows rapidly with the number of product categories and decision stages, a generalized expected value model and a corresponding recursive solution method are further developed. Numerical results show that higher transportation costs make delayed consolidation more economically attractive, whereas higher delay compensation rates favor earlier dispatch. Compared with dispatching at every stage, delaying consolidation across all stages reduces the expected number of containers by about 21–25%. The risk-preference parameter changes the optimal strategy mainly near the cost-indifference boundaries, and the direction of this change depends on how stage demand matches container capacity. This study supports decisions that balance transportation resource utilization, operating costs, and delivery delays under uncertain demand, thereby helping logistics firms improve transportation resource efficiency.

Keywords: stochastic programming; compensation mechanism; shipment consolidation; delivery delays; generalized expectation model

Check for updates

Academic Editor: Giada La Scalia

Received: 8 September 2026
Revised: 31 October 2026
Accepted: 3 November 2026
Published: 6 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Sustainable logistics requires efficient use of transportation resources while maintaining economically viable operations and meeting customer delivery needs. In China, total logistics costs totaled CNY 19.5 trillion in 2025, equivalent to 13.9% of GDP [1]. At this scale, even small gains in transport resource efficiency can yield substantial economic and resource savings, which is central to sustainable logistics. Underutilized transport capacity can increase both operating costs and resource consumption per unit of freight. Improving capacity utilization is therefore an important consideration in sustainable logistics management [2,3]. However, achieving higher utilization by consolidating shipments may prolong customer waiting times and create additional service costs. Logistics firms thus

need to balance resource efficiency with delivery requirements, particularly when different types of goods vary in their sensitivity to delays.

As logistics services move toward a wider range of products, smaller shipment sizes, and higher delivery frequencies, shipment consolidation has become an important way to reduce transportation costs and improve vehicle utilization. This problem is commonly referred to as the shipment consolidation and dispatching problem (SCDP) [4,5]. In practice, the arrival times and demand quantities for different types of goods are often uncertain [6]. When the goods waiting at a delivery point are insufficient to form a full load, logistics firms must choose between immediate dispatch and delayed dispatch to consolidate with later arrivals. Immediate dispatch can reduce customer waiting time, but it may increase the transportation cost per unit. Delayed dispatch can improve vehicle utilization, but it increases the waiting time for goods and may incur additional delay compensation costs [7–9]. Therefore, balancing transportation costs and delay compensation costs is an important issue when making shipment consolidation decisions under stochastic demand.

Compared with single-category delivery, delay decisions in multi-category shipment consolidation are more complex [10,11]. First, different types of goods may belong to different shippers. Their delivery-time requirements and tolerance for delays may therefore vary. As a result, the compensation costs caused by delays can differ greatly across product types. Second, the dispatch decision at the current stage affects not only the transportation cost of that stage but also the available loading capacity and transportation cost in later stages. This creates a continuous link between decisions across stages. When the delivery horizon is divided into multiple decision stages, stochastic demand, backlog status, and dispatch decisions interact. The problem therefore exhibits clear characteristics of a multi-stage stochastic decision problem [12].

Existing studies have examined logistics delivery under uncertainty from the perspectives of vehicle routing, distribution network design, and stochastic programming. They have also developed solution frameworks based on expected value models, risk measures, and multi-stage stochastic programming [13–17]. Existing multi-class consolidation studies [10,11] mainly focus on dispatch capacity rationing or delivery time windows, whereas category-specific delay compensation in multi-stage consolidation has received less attention. These differences affect the service costs incurred when shipments are delayed to improve capacity utilization. Neglecting them therefore limits the assessment of the trade-off between transport resource efficiency and delivery service requirements in sustainable logistics operations. In addition, as the number of product types and decision stages increases, the combinations of backlog states and dispatch decisions grow rapidly, making direct stochastic optimization computationally demanding. Moreover, decisions based solely on the expected total cost do not reflect cost variability or differences in decision-makers' risk preferences. This limits the joint evaluation of cost efficiency and cost stability, both of which are relevant to the economic sustainability of logistics operations. It is therefore necessary to develop a delivery model that retains both demand uncertainty and differences across product types and accounts for cost variability. The model should also capture multi-stage delay decisions while remaining computationally tractable.

Motivated by the compensation-based delayed delivery model [9], this study investigates a multi-category, multi-stage delayed shipment consolidation problem with differentiated compensation. We consider stochastic arrivals of different product types and varying delay compensation rates. When the available goods are insufficient to form a full load, the logistics firm may delay dispatch. The loading priority of each product type is set according to its compensation rate. Based on this setting, we first develop a stochastic decision model for multi-category, multi-stage delayed delivery. The model captures the recursive relationships among backlog states, stochastic demand, and dispatch

decisions across stages. Second, to address the rapid growth of state combinations in direct stochastic optimization, we introduce a generalized expected value model that accounts for both expected cost and cost variability. This approach transforms the stochastic delivery problem into a deterministic decision problem, allowing for direct comparison of alternatives. We then develop a recursive solution procedure and analyze the relationships among transportation costs, differentiated delay compensation parameters, and optimal delivery strategies. Finally, numerical experiments are conducted for two-stage, three-stage, and multi-category multi-stage cases. The delivery strategies and their responses to different parameter settings are examined to demonstrate the applicability and interpretability of the proposed model. The main contributions of this study are summarized as follows:

- We extend compensation-based shipment consolidation modeling to a multi-category, multi-stage setting with shared transportation capacity. The recursive stochastic formulation incorporates category-specific delay costs and interstage decision dependence, providing a framework for examining how product heterogeneity affects the trade-off between transportation resource utilization and delivery service costs.
- We extend the GEM formulation to the multi-category setting, accounting for both expected cost and cost variability. By combining this deterministic evaluation criterion with recursion, the proposed method reduces the computational complexity of multi-category, multi-stage planning compared with direct enumeration of stochastic states.
- We characterize the conditions under which immediate dispatch or delayed consolidation is preferred through analytical and numerical comparisons. By identifying how transportation costs, differentiated compensation rates, and risk preferences influence these conditions, the study provides a basis for assessing its suitability for sustainable logistics operations.

The remainder of this paper is organized as follows: Section 2 reviews the related studies. Section 3 describes the problem, formulates the multi-category, multi-stage delayed-shipment consolidation model, and presents the solution method. Section 4 describes the numerical experiments. Section 5 presents the conclusions.

2. Literature Review

2.1. Shipment Consolidation and Delivery Delays

The shipment consolidation and dispatching problem (SCDP) aims to combine less-than-truckload shipments into full-truckload shipments. This improves vehicle or container utilization and achieves economies of scale in transportation. Its main objective is to minimize total logistics costs by balancing transportation capacity and delivery efficiency. The problem has attracted significant attention in freight forwarding, air cargo management, and supply chain coordination [3–7,18]. Existing studies have identified three main shipment consolidation policies. Time-based policies consolidate and dispatch shipments at predetermined time intervals. They maintain a stable dispatch schedule and facilitate transportation planning [5,6] and are widely used in less-than-truckload networks where linehauls between depots are scheduled in advance [19]. However, dispatch decisions under these policies are independent of the actual shipment volume. As a result, vehicle utilization may be low when demand is insufficient to form a full load. In contrast, quantity-based policies trigger dispatch only when the accumulated shipment volume reaches a preset threshold [5,20,21]. This allows transportation capacity to be used more efficiently. Although such policies can better exploit economies of scale, their dispatch timing is highly sensitive to demand arrivals. When demand is sparse, delivery may be excessively delayed. To overcome the limitations of these two policies, hybrid policies have been proposed. These policies consider both time and quantity conditions and seek to balance capacity utilization and delivery timeliness [3,5,22]. The choice among these

policies reflects a fundamental trade-off between transportation cost efficiency and service reliability. This trade-off has been widely studied under different demand patterns and cost structures [23]. In these studies, consolidation policies are typically compared by their expected or long-run average cost [5,20], which does not reflect the variability of costs under uncertain demand.

Although the three consolidation policies use different dispatch triggers, they all involve a trade-off between transportation and waiting costs. When consolidation is delayed to accumulate more shipments, customers incur waiting costs. These costs may take the form of contractual penalties, loss of goodwill, or deterioration of time-sensitive goods [24,25]. This trade-off gives rise to the delayed delivery problem. The decision maker must decide whether to dispatch the available goods immediately or delay the shipment in anticipation of future demand. Delayed delivery is not limited to shipment consolidation. It also appears in production scheduling, inventory replenishment, and service operations. This broad application underscores its general relevance as a decision-making problem. A key structural feature of delayed delivery is the trade-off between the fixed dispatch cost incurred for each shipment and the penalty cost accumulated over each period of delay. Such problems also naturally involve decision dependence across multiple stages. In most of these models, the waiting cost is charged at a uniform rate for all shipments, so the dispatch decision depends only on the total accumulated volume rather than on the composition of the waiting goods.

In the literature on delayed delivery, early studies developed basic cost structures under deterministic demand. They identified the conditions under which immediate dispatch is preferred to delayed dispatch and vice versa. Later studies introduced demand uncertainty, recognizing that in many practical settings, the number of future orders is unknown when a dispatch decision is made. Under stochastic demand, the optimal dispatch policy depends not only on the current inventory level but also on the probability distribution of future arrivals. Stochastic programming and dynamic decision-making methods are therefore needed [26,27]. However, when shipments of multiple types must be tracked separately, the state space of such dynamic models grows rapidly, limiting their application to multi-category consolidation. Another stream of research has examined compensation mechanisms in delayed delivery. Under such mechanisms, carriers or logistics service providers incur penalties or compensation costs when the agreed delivery time is not met [9,28]. These costs can significantly affect shipment consolidation and dispatch policies. Compensation-based modeling is particularly relevant in third-party logistics, where service-level agreements clearly specify the economic cost of delivery delays. However, existing compensation-based models mainly involve a single shipper or a uniform compensation rate [9,29]. Multi-class consolidation models distinguish priorities among demand classes [10], but they do not relate these priorities to delay compensation in multi-stage dispatch decisions.

The review reveals three research gaps. First, waiting costs are typically charged at a uniform rate, and compensation-based studies mainly involve a single shipper. How shared container capacity should be allocated among goods with different compensation rates across multiple dispatch stages therefore remains unclear. This study formulates a multi-category, multi-stage delayed shipment consolidation model in which loading priority is determined by category-specific compensation rates. Second, stochastic dispatch models become difficult to solve when the backlog of each product type is tracked separately. This study evaluates dispatch plans using the expectation and variance of the delivery cost, reducing the problem to a comparison among a limited number of plans. Third, consolidation policies are mostly compared by expected cost, which does not distinguish plans with similar average costs but different cost variability. This study incorporates cost

variability and the decision-maker's risk preference through the generalized expected value model introduced in Section 2.2.

Table 1 compares the main research streams relevant to this study. Studies on consolidation policies and time-constrained consolidation rarely consider category-specific delay costs, while compensation-based studies are limited to a single compensation rate. This study combines category-specific compensation with multi-stage stochastic dispatch decisions and incorporates cost variability into the evaluation of dispatch plans.

Table 1. Comparison of related research streams.

Research	Multi-Category	Multi-Stage	Uncertainty	Delay-Related Cost	Evaluation Criterion
Stochastic consolidation policies [5,6,10]	Partly	√	Stochastic demand or arrivals	Uniform waiting or holding cost	Expected or long-run average cost
Time-constrained consolidation [11,19,22,23]	—	√	Mostly deterministic	Time windows, lead times or committed delivery dates	Total cost
Compensation-based delayed delivery [9,28]	—	√	Stochastic demand	Uniform compensation rate	Expected value and variance
Sustainability-oriented consolidation [21,29]	Partly	√	Stochastic demand or disruption scenarios	—	Expected cost
This study	√	√	Stochastic demand	Category-specific compensation	Expected value and variability

2.2. Stochastic Programming and the Generalized Expected Value Model

Stochastic programming provides a general framework for sequential decision-making under uncertainty, in which uncertain parameters are represented by probability distributions and decisions are adapted to the information available at each stage [30]. A simple way to obtain a deterministic decision criterion is to evaluate random costs using their expectations. Although computationally convenient, expectation-based evaluation captures only the average level of uncertain outcomes and does not explicitly distinguish alternatives with different degrees of variability. This limitation has motivated risk-sensitive formulations that supplement expected performance with additional measures of uncertainty.

Chance-constrained programming manages uncertainty by requiring selected constraints to hold with a prescribed probability [31], whereas robust optimization protects decisions against realizations within a specified uncertainty set [32]. Mean-risk formulations offer another alternative by combining expected performance with measures of dispersion or downside risk [33]. These approaches offer different interpretations of decision risk, but incorporating them into multistage problems can substantially increase computational requirements because decisions and uncertainty evolve sequentially and are commonly represented by expanding scenario structures [34,35].

This issue is particularly relevant to multistage shipment decisions. At each dispatch epoch, the decision depends on accumulated cargo and newly realized demand, while the current dispatch action determines the backlog entering subsequent stages. Hence, the evaluation criterion should preserve information about demand uncertainty while remaining sufficiently tractable for recursive computation. The Generalized Expected Value (GEM) model offers such a formulation by replacing the pure expectation objective with a composite criterion of the form $E(\xi) + C\sqrt{D(\xi)}$, where $E(\xi)$ denotes the expected cost, $D(\xi)$ denotes its variance, and $C \geq 0$ is a non-negative parameter reflecting the decision-maker's degree of risk aversion. When $C = 0$, GEM reduces to the standard expected value model; as C increases, greater weight is placed on the variability of outcomes, penalizing solutions that are optimal on average but highly volatile in practice. This formulation has a natural statistical interpretation: under normality, $E(\xi) + C\sqrt{D(\xi)}$ corresponds to a quantile of the cost distribution, so that C can be calibrated to a target service level.

Compared with chance-constrained and robust approaches, GEM retains the analytical tractability of the expected value model while incorporating a meaningful measure of uncertainty, making it well suited to multi-stage shipment consolidation problems where closed-form or recursive solution structures are available. In this paper, GEM is applied to the multi-category delayed shipment consolidation problem, in which the dispatch decision at each stage depends on the accumulated cargo of multiple freight categories under stochastic demand, and the optimal policy is characterized through analytical boundary conditions in the parameter space of transportation cost and delay penalties.

3. Problem Description and Model

3.1. Problem Description and Assumptions

Given the characteristics of shipment consolidation, demand for each product type arrives randomly. Therefore, the delivery problem is formulated as a stochastic programming problem. During the delivery process, the freight forwarder needs to load different types of goods into the same container. When the total volume of all product types meets the full-container requirement, the shipment must be dispatched. When the total volume is insufficient to fill a container, the freight forwarder may delay the shipment and wait for future arrivals. This can help form a full container and reduce the transportation cost per unit.

Delayed delivery introduces an additional constraint to the delivery process. The decision becomes more complex when multiple product types share the same container. Unlike single-category delivery, different product types may belong to different shippers and have varying tolerance for delay. This increases the complexity of delay decisions and has received considerable attention in related studies. In this section, we consider a multi-category full-container delivery problem under stochastic demand and discuss how the delivery mechanism is determined. The relevant agreements and requirements are as follows:

1. Each shipper pays the freight forwarder according to the agreed terms and the actual quantity shipped. If the freight forwarder fails to deliver the goods on time, it must pay delay compensation to the corresponding shipper.

2. All orders of all product types must be delivered within the same delivery cycle.

3. Each delivery cycle is divided into n time periods. In each period, if the total shipment volume meets the full-load requirement, the goods must be dispatched. If the total volume is insufficient for a full load, the shipment may be delayed. However, each delay incurs a penalty.

4. When several product types need to be delayed in the same period, the freight forwarder determines their loading priority based on their delay compensation rates. Product types with higher compensation rates receive higher loading priority.

The main symbols used in this paper are summarized in Table 2.

For simplicity, we make the following assumptions:

(1) The container has a fixed capacity Q. If the shipment volume reaches the full-load requirement, the goods must be dispatched.

(2) There are m product types. At the end of the first stage, the backlog of product type i that cannot form a full container is denoted by k^l, satisfying $\sum_{l=1}^{m} k^l < Q$.

(3) If the total volume of all product types is insufficient for a full-container shipment, delivery may be delayed. Otherwise, delivery cannot be delayed.

(4) The transportation cost for each dispatched container is c.

(5) In each stage, the delay compensation cost per unit of product type l is p_l. The compensation rates differ across product types. Without loss of generality, assume that $p_1 > p_2 > \ldots > p_m$.

(6) The loading priority is determined by the delay compensation rates. Product types with higher compensation rates are loaded first. Specifically, the product types are

loaded sequentially in the order $p_1, p_2, \ldots, p_m$, and the remaining capacity is allocated to the subsequent product types in turn.

(7) The demands for different product types are mutually independent, and the demands across stages are also mutually independent. The demand distribution for each product type at each stage is known, for example, from historical order records, and may differ across stages.

(8) Only the transportation cost for each dispatched container and the delay compensation cost are considered.

Table 2. Notation and Description.

Notation	Description
m	Number of product types; l denotes a product type, $l = 1, 2, \ldots, m$
n	Number of stages in a delivery cycle; i denotes a stage, $i = 1, 2, \ldots, n$
Q	Container capacity
c	Transportation cost per dispatched container
p_l	Delay compensation cost per unit of product type l per stage, $p_1 > p_2 > \ldots > p_m$
k^l	Backlog of product type l at the end of the first stage that cannot form a full container
k	Total first-stage backlog, $\sum_{l=1}^{m} k^l < Q$
C	Non-negative risk preference parameter in the generalized expected value model
ξ_i^l	Demand of product type l arriving in stage i
ξ_i	Total demand of all product types in stage i, $\xi_i = \sum_l \xi_i^l$
F_A, F_B	Demand distributions of product types A and B in the three-stage model
x_i	Dispatch decision in stage i; $x_i = 1$ if the waiting goods are dispatched, and $x_i = 0$ if dispatch is delayed
η_i^l	Quantity of product type l waiting for shipment in stage i; $\eta_i = \sum_l \eta_i^l$
d_i^l	Quantity of product type l dispatched in stage i under the loading priority rule
r_1^l, r_2^l	Remaining backlog of product type l after full containers are dispatched in stage 2, when the first-stage backlog is delayed or dispatched, respectively
z_i	Total delivery cost accumulated up to stage i
Z_{ij}	Delivery cost in the three-stage model, where $i = 1$ (2) indicates that dispatch is delayed (not delayed) in the first stage, and j is defined likewise for the second stage

3.2. Three-Stage Decision Mechanism for Multi-Category Delayed Delivery

This section takes the case of $m = 2$ and $n = 3$, that is, two product types A and B and three stages, as an example to analyze the multi-category delayed delivery decision mechanism. Let k^A, k^B denote the backlog of each product type that is insufficient to form a full container at each stage. The total backlog is $k = k^A + k^B$, satisfying $k < Q$. Suppose that the demands arriving in the second and third stages are (ξ_2^A, ξ_2^B) and (ξ_3^A, ξ_3^B), respectively. The distribution of (ξ_2^A, ξ_3^A) is $\mathcal{F}_A$, and that of (ξ_2^B, ξ_3^B) is $\mathcal{F}_B$. The demands satisfy Assumption 7. Let $\xi_2 = \xi_2^A + \xi_2^B$ and $\xi_3 = \xi_3^A + \xi_3^B$ denote the total demand of all product types in the second and third stages, respectively. Let $Z_{11}(k^A,\ k^B, \xi_2, \xi_3)$ denote the delivery cost when delayed delivery is adopted in both the first and second stages. Let $Z_{12}(k^A,\ k^B, \xi_2, \xi_3)$ denote the delivery cost when delivery is delayed in the first stage but not in the second stage. Let $Z_{21}(k^A,\ k^B, \xi_2, \xi_3)$ denote the delivery cost when delivery is delayed in the second stage but not in the first stage. Let $Z_{22}(k^A,\ k^B, \xi_2, \xi_3)$ denote the delivery cost when no delayed delivery occurs in either the first or the second stage. Under

these four cases, $Z_{ij}(k^A,\ k^B, \xi_2, \xi_3),\ i = 1,\ 2,\ j = 1,\ 2$ is a random variable, abbreviated as Z_{ij}. Its specific form is given as follows:

$$Z_{11}(k^A,\ k^B, \xi_2, \xi_3) = \begin{cases} k^A p_A + k^B p_B + (k + \xi_2^A) p_A + (k + \xi_2^B) p_B \\ +\text{int}\left(\frac{k+\xi_2+\xi_3}{Q}\right) c + \delta\left(\frac{k+\xi_2+\xi_3}{Q}\right) c, & k + \xi_2 < Q, \\ k^A p_A + k^B p_B + \text{int}\left(\frac{k+\xi_2}{Q}\right) c + r_2^A p_A + r_2^B p_B \\ +\text{int}\left(\frac{r_2+\xi_3}{Q}\right) c + \delta\left(\frac{r_2+\xi_3}{Q}\right) c, & k + \xi_2 \geq Q \end{cases},$$

$$Z_{12}(k^A,\ k^B, \xi_2, \xi_3) = k^A p_A + k^B p_B + \text{int}\left(\frac{k+\xi_2}{Q}\right) c + \delta\left(\frac{k+\xi_2}{Q}\right) c + \text{int}\left(\frac{\xi_3}{Q}\right) c + \delta\left(\frac{\xi_3}{Q}\right) c,$$

$$Z_{21}(k^A,\ k^B, \xi_2, \xi_3) = \begin{cases} c + \xi_2^A p_A + \xi_2^B p_B + \text{int}\left(\frac{\xi_2+\xi_3}{Q}\right) c + \delta\left(\frac{\xi_2+\xi_3}{Q}\right) c, & \xi_2 < Q, \\ c + \text{int}\left(\frac{\xi_2}{Q}\right) c + s_2^A p_A + s_2^B p_B + \text{int}\left(\frac{\text{mod}(\xi_2, Q)+\xi_3}{Q}\right) c \\ +\delta\left(\frac{\text{mod}(\xi_2, Q)+\xi_3}{Q}\right) c, & \xi_2 \geq Q \end{cases},$$

$$Z_{22}(k^A,\ k^B, \xi_2, \xi_3) = c + \text{int}\left(\frac{\xi_2}{Q}\right) c + \delta\left(\frac{\xi_2}{Q}\right) c + \text{int}\left(\frac{\xi_3}{Q}\right) c + \delta\left(\frac{\xi_3}{Q}\right) c.$$

where $\delta(t) = \begin{cases} 0,\ t\ is\ an\ integer \\ 1,\ otherwise \end{cases}$, $r_2^A = \max\left(k^A + \xi_2^A - \text{int}\left(\frac{k+\xi_2}{Q}\right) \cdot Q,\ 0\right)$, $r_2^B = \text{mod}(\xi_2, Q) - r_2^A$, $s_2^A = \max\left(\xi_2^A - \text{int}\left(\frac{\xi_2}{Q}\right) \cdot Q,\ 0\right)$ and $s_2^B = \text{mod}(\xi_2 - Q) - s_2^A$. Here, $\text{int}(x)$ denotes the integer part of x, and $\text{mod}(a, b)$ denotes the remainder of a divided by b.

Based on the above analysis, when only the expected cost $E(\cdot)$ is considered, we obtain the following results, where $a_{ij}^A,\ a_{ij}^B, b_{ij}, i = 1,\ 2,\ j = 1,\ 2$ are constants:

$$E\left(Z_{11}(k^A,\ k^B, \xi_2, \xi_3)\right) = a_{11}^A p_A + a_{11}^B p_B + b_{11} c,$$

$$E\left(Z_{12}(k^A,\ k^B, \xi_2, \xi_3)\right) = a_{12}^A p_A + a_{12}^B p_B + b_{12} c,$$

$$E\left(Z_{21}(k^A,\ k^B, \xi_2, \xi_3)\right) = a_{21}^A p_A + a_{21}^B p_B + b_{21} c,$$

$$E\left(Z_{22}(k^A,\ k^B, \xi_2, \xi_3)\right) = a_{22}^A p_A + a_{22}^B p_B + b_{22} c.$$

The expressions of the corresponding coefficients are given in Table 3.

Table 3. Coefficient Expressions for the Three-Stage Delayed Delivery Model.

	a_{ij}^A	a_{ij}^B	b_{ij}
Z_{11}	$k^A + E[r_2^A]$	$k^B + E[r_2^B]$	$E\left[\text{int}\left(\frac{k+\xi_2}{Q}\right)\right] + E\left[\text{int}\left(\frac{r_2+\xi_3}{Q}\right)\right] + E\left[\delta\left(\frac{r_2+\xi_3}{Q}\right)\right]$
Z_{12}	k^A	k^B	$E\left[\text{int}\left(\frac{k+\xi_2}{Q}\right)\right] + E\left[\delta\left(\frac{k+\xi_2}{Q}\right)\right] + E\left[\text{int}\left(\frac{\xi_3}{Q}\right)\right] + E\left[\delta\left(\frac{\xi_3}{Q}\right)\right]$
Z_{21}	$E[s_2^A]$	$E[s_2^B]$	$1 + E\left[\text{int}\left(\frac{\xi_2}{Q}\right)\right] + E\left[\text{int}\left(\frac{s_2+\xi_3}{Q}\right)\right] + E\left[\delta\left(\frac{s_2+\xi_3}{Q}\right)\right]$
Z_{22}	0	0	$1 + E\left[\text{int}\left(\frac{\xi_2}{Q}\right)\right] + E\left[\delta\left(\frac{\xi_2}{Q}\right)\right] + E\left[\text{int}\left(\frac{\xi_3}{Q}\right)\right] + E\left[\delta\left(\frac{\xi_3}{Q}\right)\right]$

When $E(Z_{11}) = \min\{E(Z_{12}), E(Z_{21}), E(Z_{22})\}$ holds, delaying delivery in both the first and second stages is the optimal delivery strategy. In this case, the following linear inequalities must hold:

$$\begin{cases} E\big(Z_{11}(k^A,\ k^B,\xi_2,\xi_3)\big) \le E\big(Z_{12}(k^A,\ k^B,\xi_2,\xi_3)\big), \\ E\big(Z_{11}(k^A,\ k^B,\xi_2,\xi_3)\big) \le E\big(Z_{21}(k^A,\ k^B,\xi_2,\xi_3)\big), \\ E\big(Z_{11}(k^A,\ k^B,\xi_2,\xi_3)\big) \le E\big(Z_{22}(k^A,\ k^B,\xi_2,\xi_3)\big). \end{cases}$$

Thus:

$$\begin{cases} (a_{11}^A - a_{12}^A)p_A + (a_{11}^B - a_{12}^B)p_B \le (b_{12} - b_{11})c, \\ (a_{11}^A - a_{21}^A)p_A + (a_{11}^B - a_{21}^B)p_B \le (b_{21} - b_{11})c, \\ (a_{11}^A - a_{22}^A)p_A + (a_{11}^B - a_{22}^B)p_B \le (b_{22} - b_{11})c. \end{cases}$$

In practice, $p_A \ge 0,\ \ p_B \ge 0,\ c \ge 0$. The three linear inequalities above define three half-spaces in the three-dimensional parameter space $(p_A,\ \ p_B,\ c)$. Their intersection forms a convex cone. Therefore, we obtain the following theorem.

Theorem 1. *In the multi-category three-stage delayed shipment consolidation model, when the expected value is used as the evaluation criterion, the transportation costc per dispatched container, the delay compensation rate p_A for product type A, and the delay compensation rate p_B for product type B jointly define a three-dimensional convex parameter region.*

Remark 1. In the two-stage delayed shipment consolidation model, there is a linear relationship among $p_A,\ \ p_B,\ c$, as discussed in detail in Section 4.1. When the model is extended to n stages ($n \ge 3$), the relationship among $p_A,\ \ p_B,\ c$ remains linear. Therefore, Theorem 1 also holds for the n-stage delayed shipment consolidation model ($n \ge 1$).

The above analysis shows that the same conclusion holds regardless of whether delayed delivery occurs at each stage. Based on this result, the three-stage delayed delivery model, denoted Model (1), can be formulated as follows:

$$\begin{cases} \min\ z_3 = \left(\begin{array}{l} x_2\Big(x_1c + (1-x_1)(k^Ap_A + k^Bp_B) + \mathrm{int}\Big(\frac{\xi_3+(1-x_1)k}{Q}\Big)c + \delta\Big(\frac{\xi_3+(1-x_1)k}{Q}\Big)c\Big) + (1-x_2)\Big[\mathrm{int}\Big(\frac{\xi_2+(1-x_1)k}{Q}\Big)c \\ +\delta\Big(\frac{\xi_2+(1-x_1)k}{Q}\Big)(r_2^Ap_A + r_2^Bp_B)\Big] + \Big[\mathrm{int}\Big(\frac{\xi_3+(1-x_2)(\xi_2+(1-x_1)k)}{Q}\Big) + \delta\Big(\frac{\xi_3+(1-x_2)(\xi_2+(1-x_1)k)}{Q}\Big)\Big]c \end{array} \right) \\ \textbf{\textit{s.t.}} \qquad x_i = 0,\ 1,\ i = 1,\ 2. \end{cases} \tag{1}$$

where Q denotes the container capacity, and c denotes the transportation cost for each dispatched container. Within each delivery cycle, if no delayed delivery occurs in the first stage, then $x_1 = 1$; otherwise, $x_1 = 0$. Similarly, if no delayed delivery occurs in the second stage, then $x_2 = 1$; otherwise, $x_2 = 0$. k^A, k^B denotes the amount of demand remaining after the first-stage demand fails to form a full load, and the total backlog is $k = k^A + k^B$. ξ_2 and ξ_3 denote the demand in the second and third stages, respectively. p_A and p_B denote the delay compensation rates for product types A and B, respectively. r_2^A and r_2^B denote the remaining backlog of each product type after a full-container shipment is dispatched in the second stage and the available capacity is allocated according to the loading priority rule. All orders of all product types must be fulfilled within each delivery cycle, which consists of three stages.

Example 1. When $Q = 12, k^A = 2, k^B = 1$ holds, suppose that the remaining quantity that cannot form a full load at the end of the first stage is 3. Under the three-stage delayed delivery mechanism, there are 13 possible cases in which the demand in the second stage is insufficient to form a full load, namely, $\xi_2 \in \{0,\ 1,\ \ldots,\ 12\}$. For each case, the distribution center must decide whether to carry the backlog from the first stage, k^A, k^B, over to the second stage. For instance, suppose that the total quantity remaining in the second stage is 5. If the first-stage backlog has not been dispatched ($x_1 = 1$), the total quantity waiting for shipment in the second stage becomes $3 + 5 = 8$. Since this quantity is still insufficient to form a full load, the distribution center must further decide whether to dispatch it in

the second stage. It must also determine how the remaining container capacity is allocated between product types A and B according to the loading priority rule. In contrast, if the first-stage backlog has already been dispatched ($x_1 = 0$), the distribution center only needs to decide whether to dispatch the five newly arrived units in the second stage. In summary, for the two-category three-stage delayed shipment consolidation model, there are 156 possible combinations of states and decisions in the first and second stages. For each combination, there are another 78 possible states in the third stage. Thus, the total number of possible cases is 156^2. By extension, for a two-category n-stage model, the number of possible cases is 156^n. When the model is further extended to m product types, the number of states in each stage is $\binom{Q+m-1}{m}$, and the total number of possible cases becomes $\left(2\binom{Q+m-1}{m}\right)^n$. Therefore, the computational complexity grows exponentially with both the number of product types m and the number of stages n. A direct analytical solution is no longer practical. To address this problem, we treat the demand for each product type as a random variable characterized by its mean and variance. We then use the generalized expected value model (GEM) to transform the stochastic programming problem into a deterministic programming problem for solution.

3.3. Generalized Expected Value Model for Delayed Delivery Under Stochastic Demand

We develop a model that considers the costs of m product types at each stage over n stages, with the objective of minimizing the total cost. In this study, we consider only two major cost components: the transportation cost for each dispatched container and the delay compensation cost. Based on the above analysis, the multi-category-n-stage delayed delivery model, denoted Model (2), is formulated as follows:

$$\begin{cases} \min & z_n = z_{n-1}x_{n-1} + (1-x_{n-1})\left[\mathrm{int}\left(\frac{\eta_{n-1}}{Q}\right)c + \delta\left(\frac{\eta_{n-1}}{Q}\right)\sum_{l=1}^{m}\eta_{n-1}^{l}p_l\right] + \left[\mathrm{int}\left(\frac{\eta_n}{Q}\right) + \delta\left(\frac{\eta_n}{Q}\right)\right]c \\ \textbf{\textit{s.t.}} & \eta_1 = k,\ \eta_{i+1} = \xi_{i+1} + (1-x_i)\eta_i,\ i = 1,\ 2,\ 3,\ \ldots,\ n-1, \\ & \eta_1^l = k_i,\ \eta_{i+1}^l = \xi_{i+1}^l + (1-x_i)\eta_i^l - d_i^l,\ l = 1,\ 2,\ \ldots, m,\ i = 1,\ 2,\ 3,\ \ldots,\ n-1 \\ & \sum_{l=1}^{m} d_i^l = \mathrm{int}\left(\frac{\eta_i}{Q}\right)\cdot Q,\ d_i^l \geq 0, l = 1,\ 2,\ \ldots, m \\ & d_i^l = \min\left(\eta_i^l + \xi_{i+1}^l,\ \max\left(\mathrm{int}\left(\frac{\eta_i}{Q}\right)\cdot Q - \sum_{j<l}\eta_i^j,\ 0\right)\right) \\ & z_2 = x_1 c + (1-x_1)\sum_{l=1}^{m} k^l p_l + \mathrm{int}\left(\frac{\xi_2 + (1-x_1)k}{Q}\right)c + \delta\left(\frac{\xi_2 + (1-x_1)k}{Q}\right)c, \\ & x_i = 0,\ 1,\ i = 1,\ 2,\ 3,\ \ldots,\ n-1. \end{cases} \tag{2}$$

where Q denotes the container capacity, and c denotes the transportation cost for each dispatched container. p_l denotes the delay compensation cost per unit of product type l, satisfying $p_1 > p_2 > \ldots > p_m > 0$. k^l denotes the backlog of product type l in the first stage, and $k = \sum_{l=1}^{m} k^l < Q$ denotes the total backlog. z_i denotes the cost of delayed delivery in stage i. $\xi_i = \sum_{l=1}^{m}\xi_i^l$ denotes the demand for product type l in stage i, and ξ_i denotes the total demand in that stage. η_i^l denotes the quantity of product type l waiting for shipment in stage i, while d_i^l denotes the actual quantity of product type l dispatched according to the loading priority rule. If dispatch occurs in stage i, then $x_i = 1$; otherwise, $x_i = 0$. All orders must be fulfilled by the end of each delivery cycle.

Model (2) provides a formal description of the m-category, n-stage delayed delivery problem. It can be solved using the recursive method introduced in Section 3.4. This recursive method is also used for following models.

Since random variables cannot be directly compared in the same way as real numbers, a key issue in solving above model is how to compare different random outcomes. A common approach is to transform the stochastic programming problem into a deterministic programming problem. If the expected value is used to characterize the random variables, the corresponding expected value model, denoted Model (3), can be formulated as:

$$\begin{cases} \min & E(z_n) = E(z_{n-1})x_{n-1} + (1-x_{n-1})E\left(\left[\text{int}\left(\frac{\eta_{n-1}}{Q}\right)c + \delta\left(\frac{\eta_{n-1}}{Q}\right)\sum_{l=1}^{m}\eta_{n-1}^{l}p_l\right] + \left[\text{int}\left(\frac{\eta_n}{Q}\right) + \delta\left(\frac{\eta_n}{Q}\right)\right]c\right) \\ \textbf{s.t.} & \eta_1 = k,\ \ \eta_{i+1} = \xi_{i+1} + (1-x_i)\eta_i,\ i = 1,\ 2,\ 3,\ \ldots,\ n-1, \\ & \eta_1^l = k_i,\ \eta_{i+1}^l = \xi_{i+1}^l + (1-x_i)\eta_i^l - d_i^l,\ l = 1,\ 2,\ \ldots, m,\ i = 1,\ 2,\ 3,\ \ldots,\ n-1 \\ & \sum_{l=1}^{m} d_i^l = \text{int}\left(\frac{\eta_i}{Q}\right)\cdot Q,\ d_i^l \geq 0, l = 1,\ 2,\ \ldots, m \\ & E(z_2) = x_1 c + (1-x_1)\sum_{l=1}^{m} k^l p_l + E\left(\text{int}\left(\frac{\xi_2 + (1-x_1)k}{Q}\right)c + \delta\left(\frac{\xi_2 + (1-x_1)k}{Q}\right)c\right), \\ & x_i = 0,\ 1,\ i = 1,\ 2,\ 3, \ldots,\ n-1. \end{cases} \tag{3}$$

Model (3) is solved recursively. To obtain the solution to a three-stage delayed delivery model, we first need to solve the corresponding two-stage model. Similarly, solving a four-stage model requires the solution to the three-stage model. This procedure can be continued for models with more stages. Model (3) has 2^{n-1} possible solutions. As shown in Example 1, the number of state and decision combinations in Model (2) grows exponentially with both the number of product types and the number of stages, whereas the number of possible solutions in Model (3) depends only on the number of stages. Therefore, the computational complexity of Model (3) is much lower than that of Model (2).

However, the expected value captures only one characteristic of a random variable. A random variable is fundamentally different from its expected value. Therefore, the expected value model may not fully represent stochastic decisions in the original problem. The main limitation is that the reliability of the resulting decision cannot be guaranteed. If the minimum reliability of ξ is β, then $\Pr(\xi \leq a) \geq \beta$ and $\Pr(\xi \leq a) \geq \beta$ must be satisfied. Following Zhou (2022) [9], in the multi-category shipment consolidation problem, ξ corresponds to the total demand across all product types, denoted by $\xi_t = \sum_{l=1}^{m} \xi_t^l$. Since the demands of different product types are mutually independent, the variance $D(\cdot)$ satisfies

$$D\left(\sum_{l=1}^{m} \xi_i^l\right) = \sum_{l=1}^{m} D\left(\xi_i^l\right)$$

If the demands of different product types are correlated, for example, through shared customers or common market conditions, the variance becomes $D\left(\sum_{l}^{m} \xi_i^l\right) = \sum_{l=1}^{m} D\left(\xi_i^l\right) + 2\sum_{l<l'} Cov(\xi_i^l, \xi_i^{l'})$. In this case, the joint distribution of $(l_i^1, l_i^2, \ldots, l_i^m)$ replaces the product of the marginal distributions in the recursive procedure, and the remaining steps are unchanged. Positive correlation increases the variance of total demand and may change the relative variability of alternative dispatch plans, thereby affecting the dispatch decision under a given risk preference parameter C.

This property makes the generalized expected value model applicable to the multi-category setting. The value of C can be determined by jointly considering the distributional characteristics of demand for different product types. Based on the above analysis, the generalized expected value provides a composite measure that captures both the magnitude and uncertainty of a random variable. Therefore, Model (2) can be transformed into Model (4).

$$\begin{cases} \min & z = E(z_n) + C\sqrt{D(z_n)}, \\ \textbf{\textit{s.t.}} & z_n = z_{n-1}x_{n-1} + (1 - x_{n-1})\left[\text{int}\left(\frac{\eta_{n-1}}{Q}\right)c + \delta\left(\frac{\eta_{n-1}}{Q}\right)\sum_{l=1}^{m}\eta_{n-1}^{l}p_l\right] + \left[\text{int}\left(\frac{\eta_n}{Q}\right) + \delta\left(\frac{\eta_n}{Q}\right)\right]c, \\ & \eta_1 = k,\ \eta_{i+1} = \xi_{i+1} + (1 - x_i)\eta_i,\ i = 1,\ 2,\ 3,\ \ldots,\ n-1, \\ & \eta_1^l = k_i,\ \eta_{i+1}^l = \xi_{i+1}^l + (1 - x_i)\eta_i^l - d_i^l,\ l = 1,\ 2,\ \ldots, m,\ i = 1,\ 2,\ 3,\ \ldots,\ n-1 \\ & \sum_{l=1}^{m} d_i^l = \text{int}\left(\frac{\eta_i}{Q}\right)\cdot Q,\ d_i^l \geq 0, l = 1,\ 2,\ \ldots, m \\ & z_2 = x_1 c + (1 - x_1)\sum_{l=1}^{m} k^l p_l + \text{int}\left(\frac{\xi_2 + (1 - x_1)k}{Q}\right)c + \delta\left(\frac{\xi_2 + (1 - x_1)k}{Q}\right)c, \\ & x_i = 0,\ 1,\ i = 1,\ 2,\ 3, \ldots,\ n-1. \end{cases} \tag{4}$$

The above discussion leads to the following observations: (1) When $C = 0$, Model (4) is equivalent to Model (3). Therefore, the two models have the same computational complexity. (2) Model (4) has a clear structural form. Different values of C can be used to represent different levels of uncertainty in the decision-making process. (3) Model (4) has a statistical interpretation. For example, $(z_n - E(z_n))/\sqrt{D(z_n)}$ represents a reliability quantile. In practical applications, the value of C can be selected according to the distributional characteristics of $(z_n - E(z_n))/\sqrt{D(z_n)}$. (4) When $m = 1$ holds, $\sum_{l=1}^{m} k^l p_l = kp$, and Model (4) reduces to the GEM for a single product type.

Based on the above analysis, Model (4) incorporates the traditional stochastic treatment as a special case. Different forms of the combined effect function may lead to different decision rules. This also increases the computational complexity of the model.

3.4. Model Solution

According to Model (2), the solution of z_3 can be obtained from the solution of z_2, and the solution of z_4 can then be obtained from z_3. This procedure continues recursively for later stages. Since the solution at each stage depends only on the solution from the previous stage, the sequence of delivery costs has the Markov property. Therefore, the following algorithm can be used to solve the m-category, n-stage delayed delivery problem, as illustrated in Figure 1.

The detailed steps are as follows:

Step 1. Input the initial delayed delivery model (2).

1. Given the parameters c, $p_1, p_2, \ldots, p_m$, Q, and $k^1, k^2, \ldots, k^m$, obtain $k = \sum_{l=1}^{m} k^l$, where $\eta_1 = k$ and $\eta_1^l = k^l$.

2. Compute the probability distribution of the total demand $\xi_t = \sum_{l=1}^{m} \xi_t^l$ from the marginal demand distributions $\mathcal{F}^1, \mathcal{F}^2, \ldots, \mathcal{F}^m$ of all product types.

3. During the recursive process, the distribution function of η_{i+1} is jointly determined by η_i, ξ_{i+1} and the dispatch decision x_i. The backlog η_{i+1}^l of each product type is determined according to the loading priority rule.

4. Calculate the expected cost $E(z_i)$ and variance $D(z_i)$ at each stage using $E(\eta_i^l)$ and $D(\eta_i^l)$.

Step 2. Select an appropriate generalized expected value function and determine the value of parameter C.

Step 3. Transform the stochastic programming model into the generalized expected value model.

Step 4. Evaluate all 2^{n-1} possible combinations of $x_i (i = 1, 2, \ldots, n-1)$ and select the optimal value of $E(z_n) + C\sqrt{D(z_n)}$ to obtain the solution of the generalized expected value model.

Step 5. Obtain the optimal delayed delivery strategy. This includes the dispatch decision $(x_1^*, x_2^*, \ldots, x_{n-1}^*)$ at each stage and the corresponding loading plan for each product type.

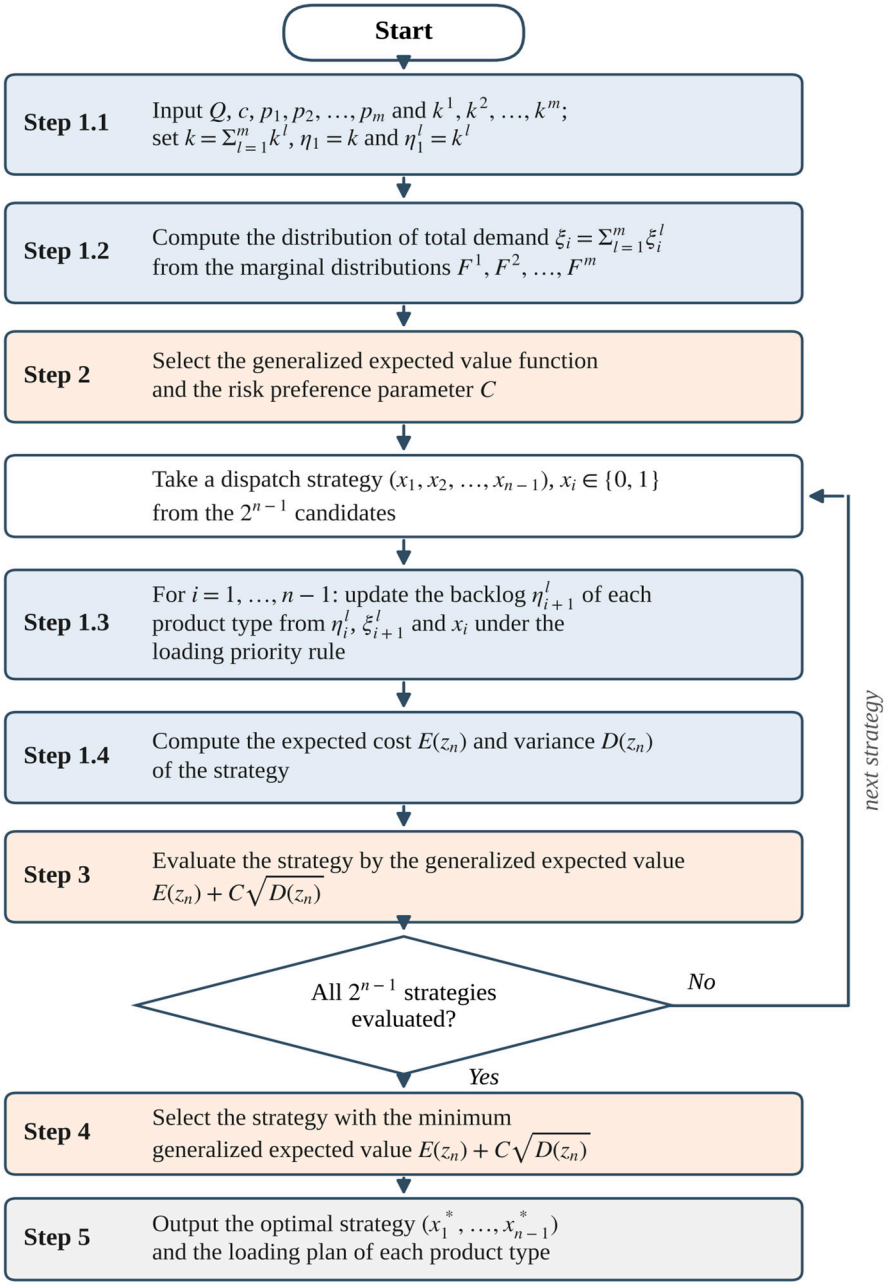


Figure 1. Process of the recursive solution procedure.

Remark 2. If the GEM is convex, it can be solved using convex optimization methods. When the objective function is linear, and the selected combined effect function defines a convex constraint, the resulting GEM is convex. Under Assumption 7, the demands of different product types are mutually independent. Therefore, $D\left(\sum_{l=1}^{m} z_n^l\right) = \sum_{l=1}^{m} D\left(z_n^l\right)$ is additive, which simplifies the calculation of the variance. In general, however, the GEM is not necessarily convex. In such cases, conventional convex optimization methods cannot be directly applied.

4. Case Studies

This section presents three numerical examples of increasing complexity: two product types over two and three stages, and three product types over four stages. The first two cases allow the strategy boundaries to be derived and visualized, whereas the third examines the

method when the parameter space can no longer be visualized. Section 4.4 evaluates the computational performance of the proposed method, and Section 4.5 discusses the results in terms of capacity utilization, risk preference, and previous studies.

4.1. Case 1: Two-Category, Two-Stage Delayed Delivery Problem

A freight forwarder consolidates shipments of two product types, A and B, into a single container with a capacity of $Q = 12$. There are two fixed dispatch dates each month: the 15th and the 30th. For orders received before the 15th of the current month, the backlogs for product types A and B are $k^A = 4$ and $k^B = 2$, respectively. Their delay compensation rates are $p_A > p_B > 0$, and the transportation cost per dispatched container is c. All orders received in the current month must be delivered by the 30th. The objective is to determine a reasonable delivery plan that minimizes the total cost.

There are two possible delivery decisions: Z_1 (delay) and Z_2 (dispatch). If delivery is postponed to the next dispatch date, one of the following two situations may occur: (1) The freight forwarder pays compensation for the delay, but total demand remains insufficient to form a full load. In this case, no transportation cost is saved, and the total cost increases. (2) The freight forwarder pays compensation for the delay, and total demand reaches the full-load requirement. In this case, transportation costs can be reduced through shipment consolidation.

Suppose that the freight forwarder has estimated the demand distributions for the two product types in the second stage using historical order records, as shown in Table 4.

Table 4. Predicted Demand Probabilities for Each Product Type in Stage 2.

ξ_2^A	3	6	7	8	9	11	12	13	14	16
$\Pr(\xi_2^A = i)$	0.03	0.04	0.12	0.18	0.26	0.12	0.06	0.1	0.05	0.04
ξ_2^B	2	4	5	6	7	8	9	10	11	12
$\Pr(\xi_2^B = i)$	0.04	0.08	0.13	0.20	0.22	0.15	0.10	0.05	0.02	0.01

By Assumption 7, the demands for the two product types are mutually independent, and the total demand is $\xi_2 = \xi_2^A + \xi_2^B$. From Table 4, we obtain $E(\xi_2) = 16.36$. Based on the demand distribution, the costs associated with a backlog of 6 units at this stage can be calculated, as shown in Table 5.

Table 5. Total Cost over Two Stages ($k = 6$).

ξ_2	5	7	8	9	10	11
$\Pr(\xi_2 = i)$	0.0012	0.0024	0.0055	0.0108	0.0170	0.0297
$Z_1(k, \xi_2)$	$c + (4p_A + 2p_B)$	$2c + (4p_A + 2p_B)$	$2c + (4p_A + 2p_B)$	$2c + (4p_A + 2p_B)$	$2c + (4p_A + 2p_B)$	$2c + (4p_A + 2p_B)$
$Z_2(k, \xi_2)$	$2c$	$2c$	$2c$	$2c$	$2c$	$2c$
ξ_2	12	13	14	15	16	17
$\Pr(\xi_2 = i)$	0.0410	0.0833	0.1052	0.1275	0.1206	0.1036
$Z_1(k, \xi_2)$	$2c + (4p_A + 2p_B)$	$2c + (4p_A + 2p_B)$	$2c + (4p_A + 2p_B)$	$2c + (4p_A + 2p_B)$	$2c + (4p_A + 2p_B)$	$2c + (4p_A + 2p_B)$
$Z_2(k, \xi_2)$	$2c$	$3c$	$3c$	$3c$	$3c$	$3c$
ξ_2	18	19	20	21	22	23
$\Pr(\xi_2 = i)$	0.0948	0.0755	0.0632	0.0458	0.0309	0.0212
$Z_1(k, \xi_2)$	$2c + (4p_A + 2p_B)$	$3c + (4p_A + 2p_B)$	$3c + (4p_A + 2p_B)$	$3c + (4p_A + 2p_B)$	$3c + (4p_A + 2p_B)$	$3c + (4p_A + 2p_B)$
$Z_2(k, \xi_2)$	$3c$	$3c$	$3c$	$3c$	$3c$	$3c$
ξ_2	24	25	26	27	28	
$\Pr(\xi_2 = i)$	0.0111	0.0060	0.0025	0.0008	0.0004	
$Z_1(k, \xi_2)$	$3c + (4p_A + 2p_B)$	$3c + (4p_A + 2p_B)$	$3c + (4p_A + 2p_B)$	$3c + (4p_A + 2p_B)$	$3c + (4p_A + 2p_B)$	
$Z_2(k, \xi_2)$	$3c$	$4c$	$4c$	$4c$	$4c$	

From Table 5, we obtain $E(Z_1) = 4p_A + 2p_B + 2.2562c$ and $E(Z_2) = 2.9021c$. For given p_A, p_B, c, the boundary condition is

$$4p_A + 2p_B = 0.6549c$$

If $4p_A + 2p_B < 0.6549c$ holds, delayed delivery is preferred; if $4p_A + 2p_B > 0.6549c$ holds, immediate dispatch is preferred.

Using the same analysis, we further examine the effects of different values of c and θ, as shown in Table 6.

Table 6. Expected Costs under Different Values of c, p_A and p_B.

c	p_A	p_B	$E(Z_1)$	$E(Z_2)$	Optimal Strategy
500	100	50	1628.10	1451.05	Z_2
500	150	100	1928.10	1451.05	Z_2
750	100	50	2192.15	2176.58	Z_2
750	150	100	2492.15	2176.58	Z_2
1000	100	50	2756.20	2902.10	Z_1
1000	150	100	3056.20	2902.10	Z_2
1250	100	50	3320.25	3627.63	Z_1
1250	150	100	3620.25	3627.63	Z_1
1500	100	50	3884.30	4353.15	Z_1
1500	150	100	4184.30	4353.15	Z_1
1750	100	50	4448.35	5078.68	Z_1
1750	150	100	4748.35	5078.68	Z_1
2000	100	50	5012.40	5804.20	Z_1
2000	150	100	5312.40	5804.20	Z_1

Using the above algorithm, we obtain the solutions of the GEM, as shown in Table 7. The expected cost $E(Z_2(k, \xi_2))$ and the generalized expected cost can also be calculated.

Table 7. Solutions of the Generalized Expected Value Model for $m = 2$ and $n = 2$.

c	p_A,p_B	$C = 0$	$C = 0.5$	$C = 1$	$C = 2$
500	100, 50	Z_2	Z_2	Z_2	Z_2
	150, 100	Z_2	Z_2	Z_2	Z_2
750	100, 50	Z_2	Z_2	Z_2	Z_2
	150, 100	Z_2	Z_2	Z_2	Z_2
1000	100, 50	Z_1	Z_1	Z_1	Z_2
	150, 100	Z_2	Z_2	Z_2	Z_2
1250	100, 50	Z_1	Z_1	Z_1	Z_1
	150, 100	Z_1	Z_2	Z_2	Z_2
1500	100, 50	Z_1	Z_1	Z_1	Z_1
	150, 100	Z_1	Z_1	Z_1	Z_2
1750	100, 50	Z_1	Z_1	Z_1	Z_1
	150, 100	Z_1	Z_1	Z_1	Z_2
2000	100, 50	Z_1	Z_1	Z_1	Z_1
	150, 100	Z_1	Z_1	Z_1	Z_1

Using the proposed algorithm, the GEM solutions under different parameter settings are obtained, as shown in Table 7. When $C = 0$, the GEM reduces to the expected value model, and the results are consistent with the boundary condition above. In the two-stage case, the compensation cost of delayed delivery is deterministic, and the variability of the delivery cost arises only from the number of dispatched containers. The standard deviation of the delivery cost is therefore proportional to c, and the boundary condition of the GEM becomes $k^A p_A + k^B p_B = [(b_2 - b_1) - C(\sigma_1 - \sigma_2)]c$, where σ_1 and σ_2 denote the standard deviations of the number of containers under delayed and immediate dispatch, respectively. In this case, delayed dispatch leads to greater variability in the number of containers. As C increases, the slope of the boundary decreases, and a higher transportation cost is required

for delayed delivery to be preferred. Table 8 reflects this effect. For parameter settings close to the expected value boundary, the optimal strategy changes from delayed delivery to immediate dispatch as C increases, whereas for settings far from the boundary, the optimal strategy remains unchanged for all values of C considered. Risk aversion therefore mainly affects dispatch decisions near the boundary between delayed and immediate dispatch, and a risk-averse freight forwarder requires a higher transportation cost before choosing to wait for consolidation.

Table 8. Predicted Demand Distributions in Stages 2 and 3.

ξ_2^A	2	4	5	6	7	8	9	10	11	12
$\Pr(\xi_2^A = i)$	0.04	0.08	0.1	0.2	0.21	0.15	0.09	0.05	0.06	0.02
ξ_3^A	3	5	6	7	8	9	10	11	12	14
$\Pr(\xi_3^A = i)$	0.03	0.07	0.12	0.18	0.24	0.16	0.10	0.05	0.03	0.02
ξ_2^B	1	3	4	5	6	7	8	9	10	11
$\Pr(\xi_2^B = i)$	0.05	0.08	0.12	0.15	0.15	0.17	0.11	0.06	0.04	0.07
ξ_3^B	2	3	4	5	6	7	8	9	10	12
$\Pr(\xi_3^B = i)$	0.05	0.09	0.13	0.18	0.20	0.15	0.10	0.05	0.03	0.02

Delayed delivery also reduces the expected number of dispatched containers and thus improves container utilization. When the transportation cost is high enough to justify waiting, consolidation lowers both the operating cost and the transport capacity required to deliver the same volume of goods.

4.2. Case 2: Two-Category, Three-Stage Delayed Delivery Problem

This case is similar to Case 1, except that each month has three fixed dispatch dates: the 10th, 20th, and 30th. Suppose that the demand distributions for each product type in the second and third stages have been estimated from historical order records, as shown in Table 8.

From Table 8, the expected total demands $\xi_t = \xi_t^A + \xi_t^B$ in the three stages are $E(\xi_2) = 13.18$ and $E(\xi_3) = 13.80$, respectively. In the three-stage delivery model, the distribution center has two decision variables, denoted by $x_1, x_2 \in \{0, 1\}$. These variables correspond to four delivery strategies: Z_{11} (delay in both stages), Z_{12} (delay in the first stage and dispatch in the second stage), Z_{21} (dispatch in the first stage and delay in the second stage), and Z_{22} (dispatch in both stages). When only the expected cost is considered, the expected cost of each strategy can be expressed as a linear function of p_A, p_B, c:

$$E(Z_{ij}) = a_{ij}^A p_A + a_{ij}^B p_B + b_{ij} c,\ i, j \in \{1, 2\}$$

The expected cost coefficients of the four strategies obtained from the numerical calculations are shown in Table 9.

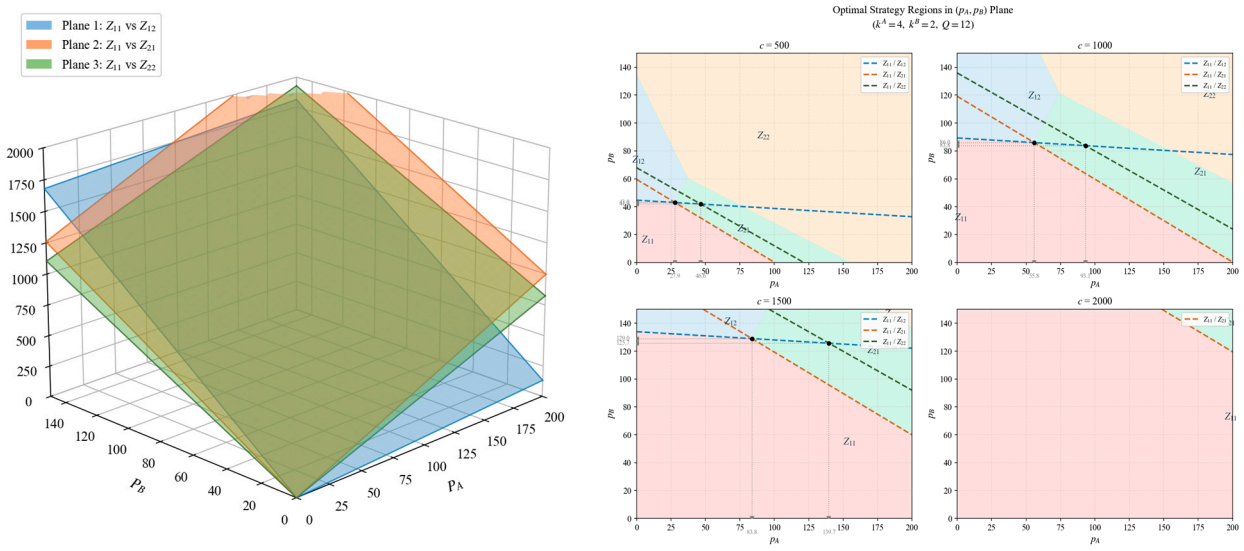
Table 9. Expected Cost Coefficients of the Four Strategies.

Strategy	a_{ij}^A	a_{ij}^B	b_{ij}
Z_{11}	4.3409	7.7555	3.1865
Z_{12}	4.0000	2.0000	3.7004
Z_{21}	1.6456	3.2208	3.7269
Z_{22}	0	0	4.2404

When $E(Z_{11}) = \min\{E(Z_{11}), E(Z_{12}), E(Z_{21}), E(Z_{22})\}$ holds, the strategy of delaying delivery in both the first and second stages is optimal. The corresponding boundary conditions are given:

$$\begin{cases} 0.3409p_A + 5.7555p_B \leq 0.5139c \\ 2.6953p_A + 4.5347p_B \leq 0.5404c \\ 4.3409p_A + 7.7555p_B \leq 1.0539c \end{cases}$$

Based on the analysis in Section 3.2 and Theorem 1, the functional relationships between c and (p_A, p_B) are illustrated in Figure 2.



(**a**) (**b**)

Figure 2. The functional relationships of case 2: (**a**) Strategy boundary surface under transportation cost and compensation rates; (**b**) optimal strategy regions under different transportation costs and compensation rates. Different background colors are used to distinguish the optimal strategy regions in each subfigure, with each region labeled by its corresponding strategy.

When only the expected value is considered, a small transportation cost c (e.g., $c = 500$) makes immediate dispatch relatively inexpensive. In this case, Z_{22} (dispatch in both stages) occupies most of the parameter region, while the region for delayed delivery is very limited. As c increases, the feasible region for delayed delivery gradually expands, and the region corresponding to Z_{11} (delay in both stages) becomes larger. The effects of p_A and p_B on the decision boundary are asymmetric. For Boundary 1 (Z_{11} vs. Z_{12}), the coefficient of p_B ($a_{ij}^B = 5.7555$) is much larger than that of p_A ($a_{ij}^A = 0.3409$). This result directly reflects the loading priority rule. Since product type A is loaded first, a larger share of the remaining backlog consists of product type B. As a result, the accumulated penalty for product B is more sensitive to changes in its compensation rate. Using the proposed algorithm, the GEM solutions are obtained, as shown in Table 10.

As shown in Table 10, when C = 0, the GEM reduces to the expected value model, and the results are consistent with the boundary conditions. When the transportation cost is small, dispatching in both stages is preferred. As the transportation cost increases, the optimal strategy shifts to dispatching the initial backlog and delaying in the second stage, and then to delaying in both stages. As C increases, the freight forwarder places greater emphasis on cost variability, and the optimal strategy changes for parameter settings near the boundary between strategies. For example, when c = 1000, $p_A = 150$ and $p_B = 100$, the optimal strategy changes from Z_{22} to Z_{21} as C increases. In this setting, Z_{22} has the lowest expected cost but the largest cost variability because the expected demand in each

stage only slightly exceeds one container load, so that the number of containers required in each stage frequently changes when goods are dispatched separately. Z_{21} dispatches the initial backlog and consolidates the second-stage goods with later arrivals, which yields a slightly higher expected cost but markedly lower variability. Figure 3 further shows how the optimal strategy regions in the parameter space of the compensation rates change with C when the transportation cost is fixed. Compared with the expected value model, an increase in C enlarges the region of Z_{21} in both directions. Its upper boundary moves upward into the region of Z_{22}, and its lower boundary moves downward into the region of Z_{11}. The parameter setting with higher compensation rates in Table 11 lies in the region of Z_{22} when $C = 0$, lies on the boundary between Z_{22} and Z_{21} when $C = 0.5$, and falls inside the region of Z_{21} as C increases further. The region of Z_{12} changes only slightly. These results indicate that the effect of risk preference is not limited to individual parameter settings but shifts the strategy boundaries over a wide range of compensation rates. In Case 2, risk aversion therefore favors partial consolidation over dispatching in every stage. The relative values of p_A and p_B also affect the optimal strategy. For the same value of c, the higher compensation rates are more likely to result in a dispatch strategy than the lower ones, indicating that a higher delay compensation rate encourages the freight forwarder to reduce delivery delays.

Table 10. Solutions of the GEM for $m = 2$ and $n = 3$.

c	p_A,p_B	$C = 0$	$C = 0.5$	$C = 1$	$C = 2$
500	100, 50	Z_{22}	Z_{22}	Z_{22}	Z_{22}
	150, 100	Z_{22}	Z_{22}	Z_{22}	Z_{22}
750	100, 50	Z_{21}	Z_{21}	Z_{21}	Z_{21}
	150, 100	Z_{22}	Z_{22}	Z_{22}	Z_{22}
1000	100, 50	Z_{11}	Z_{11}	Z_{11}	Z_{11}
	150, 100	Z_{22}	Z_{21}	Z_{21}	Z_{21}
1250	100, 50	Z_{11}	Z_{11}	Z_{11}	Z_{11}
	150, 100	Z_{21}	Z_{21}	Z_{21}	Z_{21}
1500	100, 50	Z_{11}	Z_{11}	Z_{11}	Z_{11}
	150, 100	Z_{21}	Z_{21}	Z_{21}	Z_{21}
1750	100, 50	Z_{11}	Z_{11}	Z_{11}	Z_{11}
	150, 100	Z_{11}	Z_{11}	Z_{11}	Z_{11}
2000	100, 50	Z_{11}	Z_{11}	Z_{11}	Z_{11}
	150, 100	Z_{11}	Z_{11}	Z_{11}	Z_{11}

Table 11. Effect of demand correlation between the two product types. (standard deviations for c = 1000, $p_A = 150$ and $p_B = 100$; changed solutions counted over the 56 settings in Table 10).

Correlation Coefficient	Variance of Total Demand in Stage 2	Standard Deviation of Z_{21}	Standard Deviation of Z_{22}	Number of Changed Solutions
0	10.84	564.66	684.06	—
0.3	14.00	612.65	695.52	2
0.6	17.16	654.97	706.53	8

To assess the robustness of the results to the independence assumption, positively correlated demands for the two product types are constructed by mixing the independent and comonotonic joint distributions, which leaves the marginal distributions in Table 9 unchanged. As shown in Table 11, positive correlation increases the variance of total demand and the cost variability of all strategies, but the main findings of Case 2 are preserved. Under a correlation of 0.3, the optimal strategy changes in only two of the fifty-six parameter settings, both on the boundary between Z_{22} and Z_{21}. Under a correlation of 0.6, the region of Z_{21} shrinks because its cost variability increases the most, whereas the solutions for the lower compensation rates remain unchanged. The independence assumption therefore has little effect on decisions away from the strategy boundaries, and

strongly correlated demand can be handled by using the joint distribution directly, as discussed in Section 3.3.

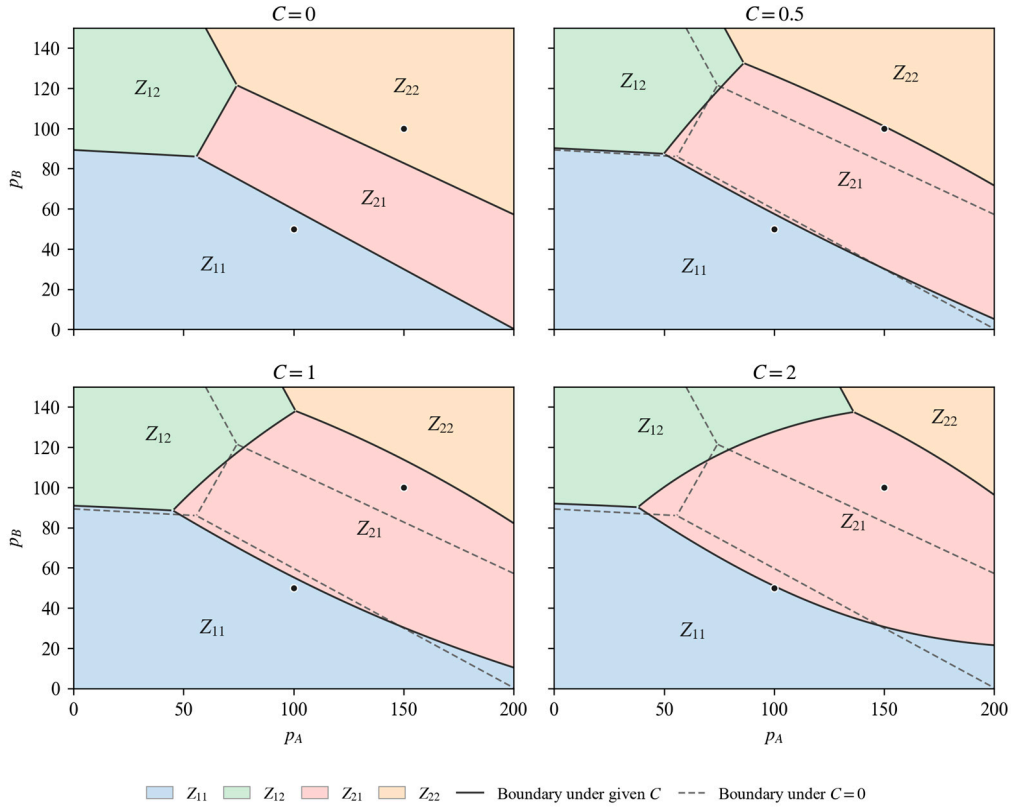


Figure 3. Optimal strategy regions of the GEM in the (p_A, p_B) plane under different risk preferences in Case 2 (c = 1000). Solid lines are the boundaries under the given C; dashed lines are the boundaries under C = 0. Dots mark the parameter settings in Table 10.

4.3. Case 3: Three-Category, Four-Stage Delayed Delivery Problem

This case is similar to Case 2, but the number of product types is increased from two to three and the number of delivery stages is increased from three to four. The purpose is to examine the applicability of the proposed model in a more complex setting. The freight forwarder handles consolidated shipments for three product types: A, B, and C. The three product types share the same container. Their delay compensation rates satisfy $p_A > p_B > p_C > 0$. The loading priority is determined by the compensation rates in descending order, namely A, B, and C. There are four fixed dispatch dates each month: the 8th, 16th, 24th, and 30th. All orders received in the current month must be delivered by the 30th. For orders received before the 8th, the backlogs of the three product types are $k^A = 3, k^B = 2, k^C = 1$, respectively. Suppose that the demand distributions of the three product types in the second, third and fourth stages have been estimated from historical order records, as shown in Table 12.

In the four-stage delivery model, the distribution center has three decision variables, denoted by $x_1, x_2, x_3 \in \{0, 1\}$. These variables correspond to $2^3 = 8$ possible delivery strategies, denoted by Z_{ijk}. Similar to Case 2, the expected cost of each strategy can be expressed as a linear function of p_A, p_B, p_C and c:

$$E\left(Z_{ijk}\right) = a^A_{ijk} p_A + a^B_{ijk} p_B + a^C_{ijk} p_B + b_{ij} c,\ i, j, k \in \{1, 2\}$$

As the number of strategies increases from 4 to 8, the number of boundary conditions increases from 3 to at most 28. The parameter space also expands from three to four

dimensions, making direct visualization infeasible. Therefore, in this case, we use the GEM to numerically determine the optimal strategy for different parameter combinations. The results are shown in Table 13.

Table 12. Predicted Demand Distributions in Stages 2, 3, and 4.

ξ_2^A	2	4	5	6	7	8	9	10	11	12
$\Pr(\xi_2^A = i)$	0.04	0.08	0.1	0.2	0.21	0.15	0.09	0.05	0.06	0.02
ξ_3^A	3	5	6	7	8	9	10	11	12	14
$\Pr(\xi_3^A = i)$	0.03	0.07	0.12	0.18	0.24	0.16	0.10	0.05	0.03	0.02
ξ_4^A	2	4	6	7	8	9	10	11	12	14
$\Pr(\xi_4^A = i)$	0.03	0.06	0.12	0.19	0.23	0.16	0.10	0.06	0.03	0.02
ξ_2^B	3	5	6	7	8	9	10	11	12	14
$\Pr(\xi_2^B = i)$	0.03	0.07	0.12	0.18	0.24	0.16	0.10	0.05	0.03	0.02
ξ_3^B	2	3	4	5	6	7	8	9	10	12
$\Pr(\xi_3^B = i)$	0.05	0.09	0.13	0.18	0.20	0.15	0.10	0.05	0.03	0.02
ξ_4^B	1	2	3	4	5	6	7	8	9	10
$\Pr(\xi_4^B = i)$	0.04	0.08	0.13	0.18	0.21	0.16	0.10	0.05	0.03	0.02
ξ_2^C	1	2	3	4	5	6	7	8	9	10
$\Pr(\xi_2^C = i)$	0.06	0.10	0.15	0.18	0.20	0.14	0.09	0.05	0.02	0.01
ξ_3^C	1	2	3	4	5	6	7	8	9	10
$\Pr(\xi_3^C = i)$	0.05	0.10	0.16	0.20	0.18	0.13	0.09	0.05	0.03	0.01
ξ_4^C	1	2	3	4	5	6	7	8	9	10
$\Pr(\xi_4^C = i)$	0.06	0.11	0.15	0.19	0.17	0.13	0.09	0.05	0.03	0.02

Table 13. Solutions of the GEM for $m = 3$ and $n = 4$.

c	p_A,p_B	$C = 0$	$C = 0.5$	$C = 1$	$C = 2$
500	100, 50	Z_{222}	Z_{222}	Z_{222}	Z_{222}
	150, 100	Z_{222}	Z_{222}	Z_{222}	Z_{222}
750	100, 50	Z_{211}	Z_{222}	Z_{222}	Z_{222}
	150, 100	Z_{222}	Z_{222}	Z_{222}	Z_{222}
1000	100, 50	Z_{211}	Z_{211}	Z_{211}	Z_{211}
	150, 100	Z_{222}	Z_{222}	Z_{222}	Z_{222}
1250	100, 50	Z_{211}	Z_{211}	Z_{211}	Z_{111}
	150, 100	Z_{222}	Z_{222}	Z_{222}	Z_{222}
1500	100, 50	Z_{111}	Z_{111}	Z_{111}	Z_{111}
	150, 100	Z_{211}	Z_{211}	Z_{211}	Z_{222}
1750	100, 50	Z_{111}	Z_{111}	Z_{111}	Z_{111}
	150, 100	Z_{211}	Z_{211}	Z_{211}	Z_{111}
2000	100, 50	Z_{111}	Z_{111}	Z_{111}	Z_{111}
	150, 100	Z_{111}	Z_{111}	Z_{111}	Z_{111}

Table 13 presents the GEM solutions for Case 3. When C = 0, a small transportation cost leads to dispatching in every stage; an intermediate cost leads to dispatching the initial backlog and delaying in subsequent stages; and a high cost leads to delaying in all stages. Higher compensation rates shift these transitions toward higher transportation costs. The effect of risk preference again appears near the boundaries between strategies, but its direction differs from that in Case 2. In Case 3, the expected total demand in each stage lies between one and two container loads, so the number of containers required in each stage is relatively stable, and dispatching in every stage yields the lowest cost variability. Near the boundary with dispatching in every stage, risk aversion shifts the optimal strategy from Z_{211} to Z_{222}. Near the boundary with delaying in all stages, Z_{111} has lower cost variability

than Z_{211}, and risk aversion shifts the optimal strategy to Z_{111}. These results indicate that risk aversion favors the strategy with lower cost variability, and which strategy this is depends on how the demand in each stage matches the container capacity.

4.4. Computational Complexity Analysis

To evaluate the computational performance of the proposed method, problems with two to four product types and two to six stages are solved, and the results are reported in Table 14. The number of state and decision combinations in direct stochastic optimization grows exponentially with both the number of product types and the number of stages, quickly exceeding any size that can be enumerated. Scenario enumeration, which evaluates each dispatch plan along every demand path, can be completed only for the smallest problems, and even in these cases it is about two orders of magnitude slower than the proposed method. In contrast, the GEM recursion merges paths with identical backlog states at each stage, so its computational effort depends on the number of backlog states rather than on the number of demand paths. All problems are solved by the GEM recursion, and the largest problem with four product types and six stages is solved in about two minutes. Computation time increases more rapidly with the number of product types than with the number of stages because the number of joint demand realizations in each stage grows with the number of product types. For problems with many product types, efficient approximation of the joint demand distribution is therefore a direction for future research.

Table 14. Computation complexity performance.

m,*n*	States per Stage	Combinations in Direct Optimization	Plans	Scenario Enumeration (s)	GEM Recursion (s)
2, 2	78	2.43×10^4	2	<0.001	<0.001
2, 3	78	3.80×10^6	4	0.014	0.004
2, 4	78	5.92×10^8	8	2.436	0.017
2, 5	78	9.24×10^{10}	16	—	0.050
2, 6	78	1.44×10^{13}	32	—	0.132
3, 2	364	5.30×10^5	2	0.001	0.001
3, 3	364	3.86×10^8	4	1.664	0.131
3, 4	364	2.81×10^{11}	8	—	0.625
3, 5	364	2.05×10^{14}	16	—	2.041
3, 6	364	1.49×10^{17}	32	—	5.858
4, 2	1365	7.45×10^6	2	0.014	0.016
4, 3	1365	2.04×10^{10}	4	—	2.356
4, 4	1365	5.56×10^{13}	8	—	9.990
4, 5	1365	1.52×10^{17}	16	—	37.999
4, 6	1365	4.14×10^{20}	32	—	115.869

"—" indicates that scenario enumeration is not performed because the number of scenario paths exceeds 2×10^6. Computations were performed using Python 3.10 on a computer equipped with a 12th Gen Intel(R) Core(TM) i5-12450H CPU (2.00 GHz; Intel Corporation, Santa Clara, CA, USA).

4.5. Discussion

Figure 4 compares the average load factor and the expected number of containers across all dispatch strategies in the three cases. These two indicators measure how efficiently transport capacity is used. The average load factor is the share of container capacity actually occupied by goods, and the expected number of containers is the number of container trips required to deliver the same volume of goods. A lower load factor means that more container trips are needed for the same freight volume, which increases resource consumption per unit of freight, as discussed in Section 1. When goods are dispatched at every stage, the average load factor is 64.2%, 64.8%, and 71.9% in Cases 1, 2, and 3, respectively, leaving more than one-third of container capacity unused in Cases 1 and 2 and more than one-quarter in Case 3. Delaying dispatch at all stages raises the load factor to 82.6%, 86.2%, and 91.6% and reduces the expected number of containers by 22.3%, 24.9%, and 21.5%, respectively. Consolidation therefore delivers the same volume of goods with about one-fifth fewer container trips.

The load factor is also higher in Case 3 than in Cases 1 and 2 under comparable strategies because the larger volume of goods in each stage reduces the share of capacity lost in the last, partially filled container of each dispatch. This indicates that consolidating more product types in shared containers improves capacity utilization. Combined with the results in Table 8, Table 12, and Table 14, consolidation is selected when transportation costs are high relative to the compensation rates, and in this range the cost-minimizing strategy coincides with the resource-efficient strategy. Differentiated compensation, in turn, limits waiting for goods with high delay costs, so that capacity utilization is improved without sacrificing the delivery requirements of time-sensitive goods. Since transportation emissions and storage-related resource consumption are not modeled explicitly, the environmental benefits of fewer container trips require further quantification.

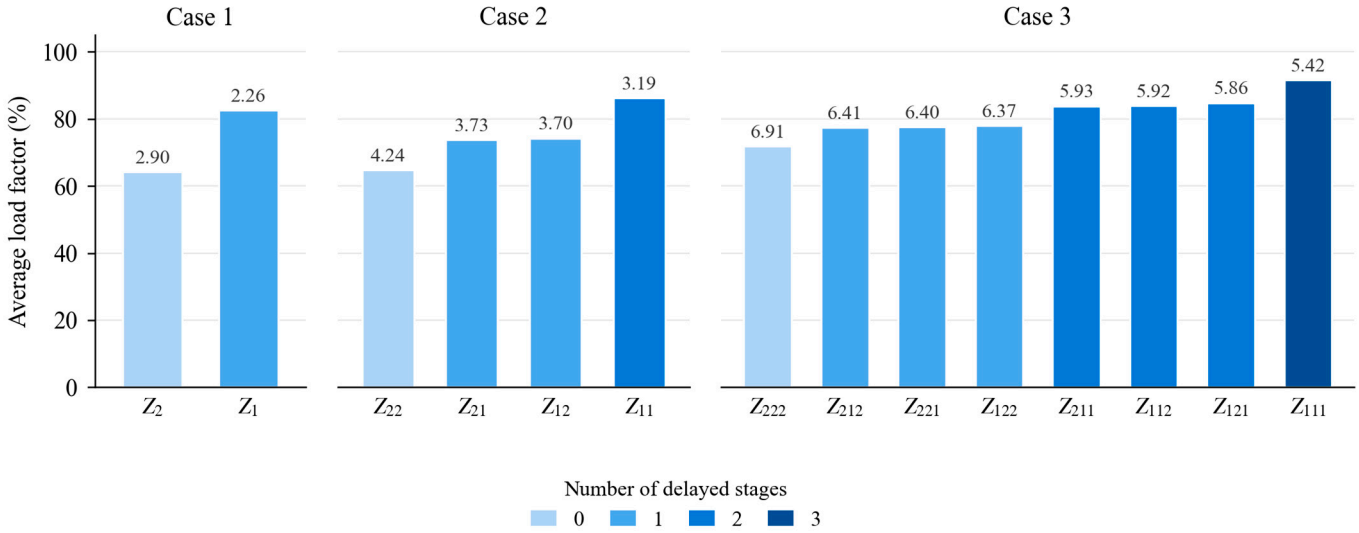


Figure 4. Average load factor of different dispatch strategies in the three cases. The numbers above the bars indicate the expected number of containers.

The three cases also clarify the role of risk preference. For parameter settings far from the strategy boundaries, the GEM yields the same decisions as the expected value model, and risk preference changes the optimal strategy only near these boundaries. The direction of the change is determined by the cost variability of the competing strategies rather than by the extent of delay. In Case 1, delayed dispatch has higher variability, and risk aversion raises the transportation cost required for consolidation. In Case 2, the expected demand in each stage only slightly exceeds one container load, so dispatching in every stage has the highest variability, and risk aversion favors partial consolidation, as illustrated in Figure 3. In Case 3, the expected demand in each stage lies between one and two container loads, so dispatching in every stage has the lowest variability, and risk aversion shifts decisions toward either dispatching in every stage or delaying in all stages, depending on which boundary the parameter setting is close to.

Compared with previous studies, the results extend our understanding of delayed delivery in three respects. First, when there is only one product type, the proposed model reduces to the compensation-based n-stage model of Zhou et al. [9], in which the strategy boundary depends on the transportation cost and a single compensation rate. With multiple product types, the loading priority rule makes the effects of the compensation rates asymmetric, as reflected in the different coefficients of p_A and p_B on the strategy boundaries in Case 2. Second, Satır et al. [10] distinguished two demand classes and showed that threshold policies can be optimal for rationing dispatch capacity. In this study, the difference between product types is expressed through their compensation rates, and

the loading priority rule links this difference directly to the allocation of shared container capacity across stages. Third, threshold-based consolidation policies [5] dispatch shipments when the accumulated quantity or waiting time reaches a preset level. In contrast, the dispatch boundaries obtained in this study shift with the compensation rates and the risk preference, so the same accumulated quantity may lead to different dispatch decisions for different shippers and decision-makers.

5. Conclusions

This study investigates the multi-category, multi-stage delayed shipment consolidation problem under stochastic demand and differentiated compensation. We develop a multi-stage delayed delivery model that captures interactions among backlog states, stochastic demand, and stage-wise dispatch decisions, and transform it into a generalized expected value model solved by a recursive procedure. The case studies show that transportation cost, delay compensation rates, and risk preference jointly determine the optimal dispatch strategy. As transportation cost increases, delaying dispatch becomes more attractive because it improves container utilization, whereas a higher compensation rate reduces the benefit of waiting. Under the loading priority rule, compensation rates for different product types affect the strategy boundaries asymmetrically. Risk preference changes the optimal strategy mainly near these boundaries, and its direction depends on how demand in each stage matches container capacity. Because the backlog carried between stages is always less than one container load, the number of states remains bounded, and problems with four product types and six stages are solved within a few minutes.

Delaying dispatch delivers the same volume of goods with about one-fifth to one-quarter fewer container trips. When transportation costs are high relative to compensation rates, the cost-minimizing strategy aligns with this resource-efficient approach. For practitioners, the strategy boundaries can guide freight forwarders in adjusting the consolidation window as transportation costs change, and compensation terms in service agreements should be set alongside loading priority. Risk-averse forwarders should compare the cost variability of alternative plans rather than follow a fixed rule of always waiting or always dispatching.

Future research may consider temporally dependent demand, demand distributions estimated from real-world order data, dynamic compensation mechanisms, and adaptive policies that revise dispatch decisions based on realized demand. This study evaluates consolidation decisions mainly from an economic perspective, based on transportation costs, compensation costs, and capacity utilization. Environmental aspects, such as transportation emissions and storage energy use, and social aspects, such as the effects of delivery delays on customers, are not explicitly modeled. Incorporating them would help provide a more comprehensive evaluation of the sustainability of delayed consolidation.

Author Contributions: Conceptualization, L.Z. and T.L.; methodology, X.S.; software, X.S.; validation, L.Z., X.S., and T.L.; formal analysis, L.Z. and X.S.; writing—original draft preparation, X.S. and T.L.; writing—review and editing, L.Z. and T.L.; visualization, T.L.; supervision, L.Z.; project administration, L.Z.; funding acquisition, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (U24A20271). The APC was funded by the same grant.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be shared on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. National Development and Reform Commission; China Federation of Logistics and Purchasing. *Bulletin on National Logistics Operations in 2025*; National Development and Reform Commission: Beijing, China, 2026.
2. Kasmi, M.; Rhalimi, A.; El-khanchoufi, M.; Ammi, A. Sustainability Synergies in Logistics: The Role of Green Practices and Social Performance in Driving Competitiveness in Morocco. *Sustainability* **2026**, *18*, 9207. [CrossRef]
3. Gao, Y.; Huang, S.; Zheng, Y.; Dai, X.; Lin, Q. Sustainability-Oriented Policy–Terrain-Coupled Mixed-Fleet Routing for Scenario-Based Green Urban Freight Logistics. *Sustainability* **2026**, *18*, 9178. [CrossRef]
4. Büyükdeveci, Ö.; Özpeynirci, S.; Özpeynirci, Ö. Multi-objective shipment consolidation and dispatching problem. *Comput. Oper. Res.* **2024**, *169*, 106728. [CrossRef]
5. Çetinkaya, S.; Bookbinder, J.H. Stochastic models for the dispatch of consolidated shipments. *Transp. Res. Part B Methodol.* **2003**, *37*, 747–768. [CrossRef]
6. Sonntag, D.R.; Schrotenboer, A.H.; Kiesmüller, G.P. Stochastic inventory routing with time-based shipment consolidation. *Eur. J. Oper. Res.* **2023**, *306*, 1186–1201. [CrossRef]
7. Niu, B.; Lai, C.; Wang, H. Shipment consolidation in assembly supply chains: Delivery delay, competition dynamics and resilience analysis. *Ocean Coast. Manag.* **2026**, *271*, 107928. [CrossRef]
8. Shu, S.; Xu, Z.; Baldacci, R. Incorporating holding costs in continuous-time service network design: New model, relaxation, and exact algorithm. *Transp. Sci.* **2024**, *58*, 412–433. [CrossRef]
9. Zhou, L.; Qi, Y.; Li, F. Research on the n-Stage Delay Distribution Method Based on a Compensation Mechanism in a Random Environment. *Axioms* **2022**, *11*, 67. [CrossRef]
10. Satır, B.; Erenay, F.S.; Bookbinder, J.H. Shipment consolidation with two demand classes: Rationing the dispatch capacity. *Eur. J. Oper. Res.* **2018**, *270*, 171–184. [CrossRef]
11. Hanbazazah, A.S.; Abril, L.; Erkoc, M.; Shaikh, N. Freight consolidation with divisible shipments, delivery time windows, and piecewise transportation costs. *Eur. J. Oper. Res.* **2019**, *276*, 187–201. [CrossRef]
12. Huang, K.; Ahmed, S. The value of multistage stochastic programming in capacity planning under uncertainty. *Oper. Res.* **2009**, *57*, 893–904. [CrossRef]
13. Schneider, M.; Timo, G.; Vigo, D. Advances in Vehicle Routing and Logistics Optimization: Exact Methods. *EURO J. Transp. Logist.* **2019**, *8*, 117–118. [CrossRef]
14. Ozbaygin, G.; Savelsbergh, M. An iterative re-optimization framework for the dynamic vehicle routing problem with roaming delivery locations. *Transp. Res. Part B Methodol.* **2019**, *128*, 207–235. [CrossRef]
15. Lin, B.; Zhao, Y.; Lin, R. Optimization for courier delivery service network design based on frequency delay. *Comput. Ind. Eng.* **2020**, *139*, 106144. [CrossRef]
16. Noyan, N. Risk-averse two-stage stochastic programming with an application to disaster management. *Comput. Oper. Res.* **2012**, *39*, 541–559. [CrossRef]
17. Abdelaziz, F.B. Solution approaches for the multi-objective stochastic programming. *Eur. J. Oper. Res.* **2012**, *216*, 1–16. [CrossRef]
18. Zhou, G.; Li, D.; Bian, J.; Zhang, Y. Airfreight forwarder's shipment planning: Shipment consolidation and containerization. *Comput. Oper. Res.* **2024**, *161*, 106443. [CrossRef]
19. Bazelmans, C.H.; ter Braake, S.L.; Schrotenboer, A.H.; van Lieshout, R.N.; Van Woensel, T. The vehicle routing problem with pickups, deliveries, and scheduled linehauls. *Eur. J. Oper. Res.* **2026**, *330*, 995–1008. [CrossRef]
20. Yücel, E.; Salman, F.S.; Erdoğan, G. Optimizing two-dimensional vehicle loading and dispatching decisions in freight logistics. *Eur. J. Oper. Res.* **2022**, *302*, 954–969. [CrossRef]
21. De Moor, B.J.; Creemers, S.; Boute, R.N. Breaking truck dominance in supply chains: Proactive freight consolidation and modal split transport. *Int. J. Prod. Econ.* **2023**, *257*, 108764. [CrossRef]
22. Greening, L.M.; Dahan, M.; Erera, A.L. Lead-time-constrained middle-mile consolidation network design with fixed origins and destinations. *Transp. Res. Part B Methodol.* **2023**, *174*, 102782. [CrossRef]
23. Xu, Z.; Li, F.; Chen, Z.L. Shipment consolidation with multiple shipping methods under nonlinear cost structures. *Manag. Sci.* **2024**, *70*, 2823–2841. [CrossRef]
24. Awaysheh, A.; Frohlich, M.T.; Flynn, B.B.; Flynn, P.J. To err is human: Exploratory multilevel analysis of supply chain delivery delays. *J. Oper. Manag.* **2021**, *67*, 882–916. [CrossRef]
25. Chen, J.; Dong, M.; Xu, L. A perishable product shipment consolidation model considering freshness-keeping effort. *Transp. Res. Part E Logist. Transp. Rev.* **2018**, *115*, 56–86. [CrossRef]
26. Minkoff, A.S. A Markov decision model and decomposition heuristic for dynamic vehicle dispatching. *Oper. Res.* **1993**, *41*, 77–90. [CrossRef]
27. Klapp, M.A.; Erera, A.L.; Toriello, A. The one-dimensional dynamic dispatch waves problem. *Transp. Sci.* **2018**, *52*, 402–415. [CrossRef]

28. Çömez-Dolgan, N.; Moussawi-Haidar, L.; Jaber, M.Y. A buyer-vendor system with untimely delivery costs: Traditional coordination vs. VMI with consignment stock. *Comput. Ind. Eng.* **2021**, *154*, 107009. [CrossRef]
29. Goodarzi, A.H.; Jabbarzadeh, A.; Fahimnia, B.; Paquet, M. Evaluating the sustainability and resilience of an intermodal transport network leveraging consolidation strategies. *Transp. Res. Part E Logist. Transp. Rev.* **2024**, *188*, 103616. [CrossRef]
30. Shapiro, A.; Dentcheva, D.; Ruszczynski, A. *Lectures on Stochastic Programming: Modeling and Theory*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2021.
31. Charnes, A.; Cooper, W.W. Deterministic equivalents for optimizing and satisficing under chance constraints. *Oper. Res.* **1963**, *11*, 18–39. [CrossRef]
32. Ben-Tal, A.; Nemirovski, A. Robust convex optimization. *Math. Oper. Res.* **1998**, *23*, 769–805. [CrossRef]
33. Ogryczak, W.; Ruszczyński, A. On consistency of stochastic dominance and mean–semideviation models. *Math. Program.* **2001**, *89*, 217–232. [CrossRef]
34. Heitsch, H.; Römisch, W.; Strugarek, C. Stability of multistage stochastic programs. *SIAM J. Optim.* **2006**, *17*, 511–525. [CrossRef]
35. Füllner, C.; Rebennack, S. Stochastic dual dynamic programming and its variants: A review. *Siam. Rev.* **2025**, *67*, 415–539. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.