Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrating Dynamic Programming and Q-Learning for Multi-Echelon Ordering Under Limited Information - A Hybrid Adaptive Framework

Leila Hosseini
Yazd University
Mohammad Saber Fallahnezhad
fallahnezhad@yazd.ac.ir
Yazd University
Vali Derhami
Yazd University
Mohammad Saleh Owlia
Yazd University

Research Article

Keywords: Supply chain management, multi-echelon ordering, Dynamic programming, Reinforcement learning, Q-learning, Markov decision process, Hybrid decision system

Posted Date: June 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9599083/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Integrating Dynamic Programming and Q-Learning for Multi-Echelon Ordering Under Limited Information -A Hybrid Adaptive Framework

Leila Hosseini
PhD Student, Department of Industrial Engineering, Yazd University, Yazd, Iran
lysahoseini@stud.yazd.ac.ir
Mohammad Saber Fallahnezhad * (Corresponding Author)
Professor, Department of Industrial Engineering, Yazd University, Yazd, Iran
fallahnezhad@yazd.ac.ir
Building of Faculty of Engineering 1, Room 317 | Tel: +98 31 31232548
Vali Derhami
Professor, Department of Computer Engineering, Yazd University, Yazd, Iran
vderhami@yazd.ac.ir
Building of Faculty of Engineering 1, Room 223 | Tel: +98 31 31232365
Mohammad Saleh Owlia
Professor, Department of Industrial Engineering, Yazd University, Yazd, Iran
owliams@yazd.ac.ir
Building of Faculty of Engineering 1 | Tel: +98 31 31232416

Integrating Dynamic Programming and Q-Learning for Multi-Echelon Ordering Under Limited nformation: A Hybrid Adaptive Framework

Leila Hosseini
PhD Student, Department of Industrial Engineering, Yazd University, Yazd, Iran
lysahoseini@stud.yazd.ac.ir

Mohammad Saber Fallahnezhad * (Corresponding Author)
Professor, Department of Industrial Engineering, Yazd University, Yazd, Iran
fallahnezhad@yazd.ac.ir

Vali Derhami
Professor, Department of Computer Engineering, Yazd University, Yazd, Iran
vderhami@yazd.ac.ir

Mohammad Saleh Owlia
Professor, Department of Industrial Engineering, Yazd University, Yazd, Iran
owliams@yazd.ac.ir

Abstract

Purpose
This study aims to propose a hybrid decision-making framework for multi-echelon ordering from multiple suppliers under conditions of limited information, where only current-period data are available and future values exhibit stochastic behavior.

Design/methodology/approach
The problem is first formulated as a Markov decision process (MDP), and the optimal policy under full information is derived using dynamic programming. Subsequently, a Q-learning-based reinforcement learning algorithm is employed as a model-free approach to learn policies under partial environmental knowledge. The framework adaptively selects between the two policies in each period using available information and probabilistic forecasts. Real-world industrial data are used for computational experiments.

Findings

The proposed framework effectively reduces total system costs and delivers robust performance under limited information conditions. The intelligent integration of model-based (dynamic programming) and data-driven (reinforcement learning) methods provides an efficient and adaptive decision mechanism for complex and uncertain supply chain environments.

Originality/value
This research introduces a novel adaptive decision mechanism that dynamically switches between dynamic programming and Q-learning policies based on real-time information and probabilistic forecasts. Unlike conventional approaches that rely solely on either model-based or model-free methods, the proposed hybrid framework leverages the strengths of both to handle multi-echelon ordering problems under information constraints.

Keywords: Supply chain management, multi-echelon ordering, Dynamic programming, Reinforcement learning, Q-learning, Markov decision process, Hybrid decision system

1. Introduction

In recent decades, the increasing complexity of global supply chains and the pervasive volatility of business environments have elevated ordering planning and optimal resource allocation to among the most challenging problems in industrial engineering and operations management (Hosseini et al., 2022). This complexity is further amplified when exchange rates, trade policies, and governmental regulations undergo rapid shifts (Kawakami et al., 2021).In such environments, fluctuations in ordering costs, raw material prices, and supply conditions create significant decision-making challenges, underscoring the need for dynamic and adaptive approaches (Chen & Rossi, 2021).

The multi-echelon ordering problem from multiple suppliers has attracted considerable attention in the research literature as a key supply chain management issue (Maulidi et al., 2022). In this problem, the decision-maker must determine the total order quantity and its allocation across multiple suppliers in each period, such that total costs—including ordering and inventory holding costs—are minimized (Zhang et al., 2024). Despite notable advances in supplier selection and order allocation, the integration of machine learning with optimization models, the replacement of binary delay classification with probabilistic approaches, and the assessment of Industry 4.0 readiness remain understudied; accordingly, this research proposes a dual-objective framework based on machine learning, TOPSIS, SWARA, and multi-objective linear programming to address these gaps (Shidpour et al., 2025).This problem is particularly significant in material-dependent industries such as steel, petrochemicals, and polymers (Li et al., 2009). In polymer industries, raw material price volatility and demand fluctuations render order quantity and supplier allocation decisions a multi-period, dynamic, and stochastic problem (Cheraghalipour & Farsad, 2018).

The literature indicates that sustainable supplier selection methods primarily fall into three categories: multi-criteria decision-making (MCDM), artificial intelligence, and hybrid MCDM-AI (Gholamazad & Eydi, 2026). However, existing studies suffer from limitations including small datasets, a limited number of criteria, static evaluation, and a lack of scalability; consequently, the integration of data envelopment analysis (DEA) with machine learning has been proposed as a more comprehensive approach (Gholamazad & Eydi, 2026).

Classical inventory management models, such as the Economic Order Quantity (EOQ) model, rely on the assumption of stable parameters and perform well in steady environments (Scarf, 1960). Although these models provide deep theoretical insight into inventory optimization problems, they are incapable of modeling stochastic fluctuations, multi-supplier interactions, and inter-period dependencies (Thompson et al., 2009). Dynamic programming and the Wagner-Whitin algorithm were subsequently developed to improve decision-making accuracy (Wagelmans et al., 1992; Wagner & Whitin, 1958, 2004). While these methods yield optimal solutions, they require complete knowledge of the environment and face computational limitations in large-scale problems (Van Roy et al., 1997).

In many real-world supply chain management problems, decision-makers encounter limited information regarding future conditions (Powell et al., 2002). In other words, only current-period data are available and future values are inherently stochastic. This issue is exacerbated in environments characterized by price volatility, structural changes in the supply network, and environmental uncertainties (Zhao & Sun, 2010).
The Markov Decision Process (MDP) provides a suitable framework for modeling multi-period decision-making problems under uncertainty (Azaron et al., 2009). In this framework, the ordering problem is formulated as an MDP in which the system state encompasses inventory levels and cost conditions; upon observing the state, the decision-maker determines order quantities and supplier allocations (Maulidi et al., 2022). The literature indicates that traditional scheduling methods are predominantly offline and require frequent re-solution when confronting uncertainty; reinforcement learning, by modeling MDP for online decision-making, offers a more efficient solution to bridge this gap (Liu et al., 2025). Nevertheless, implementing this framework requires full knowledge of transition probability matrices, which is unavailable in many real-world settings.
Reinforcement learning, as a model-free approach, enables the learning of decision-making policies through direct interaction with the environment (Sutton & Barto, 1998). This characteristic renders it a suitable option for complex and information-constrained environments (Geevers et al., 2024), particularly those with high dimensionality and uncertainty (Darezereshki & Derhami, 2024). Employing Markov Decision Processes (MDP) and dynamic programming for multi-stage ordering in multi-supplier systems, while accounting for price fluctuations, led to a reduction in supply chain costs(Hosseini et al., 2026). Recent studies have demonstrated the effectiveness of reinforcement learning in multi-echelon inventory management, supplier selection, and operational cost reduction (Piao et al., 2023; Zhou et al., 2024). Furthermore, the integration of reinforcement learning with robust optimization has contributed to mitigating market volatility risk (Kharvi & Pakkala, 2022; Luo et al., 2021). The literature also reveals that data-driven methods, particularly machine learning, have received insufficient attention in supplier selection, with existing studies predominantly focusing on time series forecasting and isolated models; accordingly, an integrated framework based on machine learning for the simultaneous prediction of delay risk and sustainability criteria (environmental and social) has been proposed (Ziari et al., 2026).
Despite significant advances in the literature, a fundamental gap persists. A systematic comparison between model-based approaches (e.g., dynamic programming) and model-free approaches (e.g., reinforcement learning) within a unified framework under multi-echelon, multi-supplier uncertainty has rarely been conducted (Zhang et al., 2024). Moreover, many existing models fail to integrally model simultaneous multi-supplier interactions, cost dynamics, and inter-period decision effects. These limitations have hindered the development of practical and reliable solutions for complex industrial environments. Additionally, the absence of adaptive selection mechanisms that dynamically choose among different decision-making methods in each period based on available information represents one of the most critical deficiencies in the current literature.
The primary objective of this research is to propose a hybrid decision-making framework for the multi-echelon ordering problem in multi-supplier and uncertain environments, capable of dynamically selecting between model-based and model-free approaches under information constraints to apply the optimal ordering policy. The main contributions of this research are threefold: First, the definition of a limited-information scenario in which only current-period costs are known and future states are predicted probabilistically. Second, the design of a policy selection mechanism that compares the outputs of dynamic programming and reinforcement learning to select the policy with the lowest predicted cost. Third, the development of an adaptive hybrid framework integrating model-based and model-free methods.
The results of this study are expected to provide clear insight into the relative performance of the methods while offering practical guidance for decision-makers in selecting ordering policies in dynamic and volatile environments. Accordingly, this paper contributes meaningfully to the supply chain management and decision-making-under-uncertainty literature both methodologically and practically. The remainder of this paper is organized as follows: Section 2 presents the literature review; Section 3 introduces the problem model and the proposed framework; Section 4 describes the solution methods; Section 5 presents the numerical results and analyses; and Section 6 discusses managerial implications and conclusions.

2. Methodology

2.1 Problem description and assumptions

This research investigates a multi-period ordering problem from multiple suppliers in a stochastic and dynamic environment, in which the decision-maker has access only to current-period information, including current ordering costs for each supplier, demand over all periods until the end of the planning horizon, and inventory holding costs for each period. Each supplier exhibits an independent cost characteristic, with ordering costs fluctuating over time according to a Markov chain with a finite number of cost levels. The transition probability from one cost level to another in each period is predetermined in the dynamic programming approach and computed from historical data, whereas in the reinforcement learning approach, transition probabilities are not pre-specified and the agent gradually discovers and optimizes them through repeated interaction with the environment. Current-period demand is assumed to be deterministic and known, which aligns with the reality of polymer and steel industries based on order-driven production planning. The number of suppliers is fixed and equal to S, and each supplier is assumed to be capable of fully meeting demand, with no capacity constraints considered for simplicity. The decision-making horizon is finite and equal to T periods, and at the end of the horizon the ending inventory must be zero to eliminate surplus inventory costs and enable a more precise evaluation of decision-making policies. In each period, the decision-maker must simultaneously answer two key questions: how much to order and how to allocate it among suppliers, such that total ordering and inventory holding costs over the planning horizon are minimized. Future supplier costs are unknown and can only be predicted through the probabilistic Markov chain model. This uncertainty highlights the fundamental difference between the two solution approaches employed in this research, namely dynamic programming and reinforcement learning; the former is based on the assumption of access to a complete probabilistic model, while the latter is based on experiential learning from environmental interaction.

2.2 Markov Decision Process (MDP) modeling

The ordering problem is formulated as a Markov Decision Process characterized by the tuple {S, A, P, R}. The components of this model are elaborated below.

2.2.1 State space

To fully represent the system condition at each point in time, a set of variables must be defined that encompasses all essential information required for decision-making. The Markov property requires that the future state of the system depend solely on the current state and the selected action, with past behavioral history having no influence. Accordingly, the state vector at period t consists of three components, as expressed in Equation (1):

$$\mathrm{S_t} = \left\{ c_i^{(\mathrm{w_i})}, \mathrm{I}_t, t \right\} \qquad t = 1,\ 2, 3, 4 \tag{1}$$

where:

- $c_i^{(\mathrm{w_i})}$i: denotes the ordering cost level for supplier s in period t,
- I_t: represents the inventory level at the beginning of period t,
- t: indicates the decision-making period.

2.2.2 Action space

In each period, the decision-maker must be able to answer two fundamental questions: How much to order? and How to allocate this quantity across suppliers? Accordingly, the action space is defined as an ordered pair, as presented in Equation (2):

$$A_t = (\lambda, X_t) \tag{2}$$

The order quantity allocated to supplier s in period t is given by Equation (3):

$$x_{st} = \lambda_{st} \times X_t \tag{3}$$

where λ_{st} denotes the allocation coefficient for supplier s. Allocation coefficients must satisfy the following constraint, as specified in Equation (4):

$$\sum_{s=1}^{S} \lambda_{st} = 1 \qquad \forall t \in T \tag{4}$$

such that x_{st} represents the order quantity from supplier s in period t. This dual-layer framework enables simultaneous decision-making regarding total order quantity and each supplier's share, allowing the model to respond optimally to cost fluctuations and inventory level changes (Azaron et al., 2009).

2.2.3 Transition function

The transition function specifies the probability with which the system transitions from the current state to the next state upon selecting a given action. To model cost variations for each supplier, an independent Markov chain is employed, as stated in Equation (5):

$$P_s = \left[p_{sij}\right] = \begin{pmatrix} p_{s11} & \cdots & p_{s1W_S} \\ \vdots & \ddots & \vdots \\ p_{sW_S1} & \cdots & p_{sW_SW_S} \end{pmatrix} = i, j, 1, \ldots, W_S \tag{5}$$

where, according to Equation (6), p_{sij} denotes the transition probability from cost level i to cost level j. The multi-period transition probability is computed based on Equation (7):

$$p_{sij} = \left(C_{s(t+1)} = C_s^j \middle| C_{st} = C_s^i\right) \tag{6}$$

$$p_{sij}{}^{u} = \left(C_{s(t+u)} = C_s^j \middle| C_{st} = C_s^i\right) \tag{7}$$

The stochastic property of the transition matrix requires that, according to Equation (8), the sum of elements in each row equals unity:

$$\sum_{j=1}^{W_S} p_{ij}^{s} = 1 \tag{8}$$

In the dynamic programming approach, these matrices are assumed to be known a priori. In contrast, in reinforcement learning, the agent implicitly learns these probabilities through repeated interaction with the environment and observation of outcomes.

2.2.4 Reward function

Since the ultimate objective of the problem is to minimize the sum of ordering and holding costs, the reward function is defined as the negative of the current-period cost, as expressed in Equation (9):

$$R(s_t, a_t) = -\left(\sum_{s=1}^{S} c_s^{(w_s)} x_{st} + h_t I_t\right) \tag{9}$$

where:

- $c_s^{(w_s)}$: denotes the unit ordering cost from supplier s at cost level w_s,
- x_{st}: represents the order quantity from supplier s in period t,
- h_t: is the unit holding cost in period t.

This negative reward definition aligns the maximization framework of reinforcement learning with the cost minimization problem. By repeatedly interacting with the environment and continuously updating action values (Q-values), the agent gradually converges to the optimal policy without requiring explicit knowledge of the transition function.

2.3 Solving the model with dynamic programming (optimal baseline under full information)

Dynamic programming is a model-based approach that guarantees the optimal policy under the assumption of complete model knowledge (transition and cost matrices). The value function, denoted $V_T(s)$, represents the minimum expected cost from state s at the beginning of period t until the end of horizon T, and is solved recursively using the Bellman equation backward from the end of the horizon, as expressed in Equations (10) and (11):

$$V_T\left(c_1^{(w_1)}, c_2^{(w_2)} \dots, c_s^{(w_s)}, I_j, j\right) = 0 \tag{10}$$

$$V_j\left(c_1^{(w_1)}, c_2^{(w_2)} \dots, c_s^{(w_s)}, I_j, j\right) = \min_{j<k\le T} \left\{ \begin{array}{l} \sum_{s=1}^{S} \lambda_s c_s^{(w_s)} \sum_{r=j+1}^{k} d_r \\ + \sum_{t=j+1}^{k-1} h_t \left(\sum_{r=t+1}^{k} d_r \right) \quad (10) \\ + \sum_{n_1=1}^{W_1} \dots \sum_{n_s=1}^{W_s} (P_1)_{i n_1}^{(k-j)} \dots (P_2)_{l n_s}^{(k-j)} V_k\left(c_1^{(n_1)}, c_2^{(n_2)} \dots, c_s^{(n_s)}, I_k, k\right) \end{array} \right\} \tag{11}$$

$$i = 1,2,\dots,n\,,\, l = 1,2,\dots,m$$

,

In this formulation:

- λ_s: is the allocation share assigned to the s-th supplier,
- $c_s^{(w_s)}$: is the ordering cost for supplier s at cost level w_s,

- $(P_s)^{(k-j)}_{w_s n_s}$ is the transition probability from cost level w_s to level n_s for supplier s over the interval k–j,
- d_r: is the demand in period r,
- h_t : is the holding cost in period t.
- h_t is the holding cost in period t.

In the specialized Bellman equation for the multi-supplier model, the discount factor is set to one and a long-term perspective is considered. In practice, for each state in each period, all feasible actions are evaluated and the least-cost option (in terms of the sum of immediate cost and expected future cost) is selected as the optimal decision.

2.4 Solving the model with Q-learning

The Q-learning algorithm is a model-free method that learns the optimal policy through direct interaction with the environment without requiring knowledge of the transition function P. The cost function and reward function in period t are given by Equations (12) through (14):

$$(\text{Cost})_t = \left(\sum_{t=1}^{T} \sum_{s=1}^{S} c_{st}^{w_s} X_{st} + \sum_{t=1}^{T-1} h_t I_t\right) \quad (12)$$

$$R(s_t, a_t) = -(\text{Cost})_t \quad (13)$$

or

$$R(s_t, a_t) = -\left[\left[\lambda c_1^{(w_1)} + (1-\lambda) c_2^{(w_2)}\right] \times \sum_{r=j+1}^{k} d_r + \sum_{t=j+1}^{k-1} h_t \left(\sum_{r=t+1}^{k} d_r\right)\right] \quad (14)$$

The learning process is updated according to Equation (15):

$$Q(S_t, A_t) \rightarrow Q(S_t, A_t) + \alpha \left[R_t + \gamma . \max_{\acute{A}} (S_{t+1}, \acute{A}) - Q(S_t, A_t)\right] \quad (15)$$

The optimal policy is extracted as defined in Equation (16):

$$\pi^*(s) = \arg \max_{a \in A(s)} Q(s, a) \quad (16)$$

Learning parameters, including learning rate α, discount factor γ, and the number of episodes, were empirically set to 20,000. By implementing the ε-greedy policy, the value of ε started at 0.995 and was gradually reduced.

2.5 Proposed hybrid decision-making framework

The core and innovative contribution of this paper is the hybrid decision-making framework that integrates dynamic programming and reinforcement learning (Q-learning) through intelligent synthesis. The central idea of this framework rests on the premise that neither approach alone achieves optimal performance under all conditions. Dynamic programming is capable of delivering the optimal policy when the system model is fully known and deterministic; however, it is vulnerable and inefficient against structural uncertainty. Conversely, Q-learning exhibits adequate flexibility in unknown and variable environments, but provides no guarantee of convergence to the global optimum. The proposed framework leverages both approaches simultaneously, reinforcing their strengths and compensating for their limitations.

At each decision-making period t, the hybrid framework generates two distinct policies in parallel. The first policy is the dynamic programming policy, derived using known transition matrices and the Bellman equation solution. This policy provides a theoretical and analytical basis for decision-making. More precisely, if the dynamic programming policy is employed, at each period t and for each performance level, an action is selected such that for every possible state j, the probability of reaching that state is computed, future costs from that state are considered, and the sum of these equals the total expected cost. This approach evaluates the expected value of a decision based on the probabilistic distribution of future states and their long-term consequences. The optimal policy derived from the DP approach is given by Equation (17):

$$\hat{C}^{DP} = \sum_{\acute{s}} P(s_t, \pi^{DP}(s_t), \acute{s}) \cdot \hat{C}\left(\acute{s}, \hat{\hat{\pi}}^{DP}(\acute{s})\right) \quad (17)$$

The second policy is the reinforcement learning policy, which selects the best action from the current state based on the trained Q-table and agent-environment interaction. This policy reflects the agent's practical experience from the environment and indicates the expected cost if the reinforcement learning policy is executed in the real environment. The optimal policy derived from the RL approach is given by Equation (18):

$$\hat{C}^{RL} = \sum_{\acute{s}} P(s_t, \pi^{RL}(s_t), \acute{s}) \cdot \hat{C}\left(\acute{s}, \hat{\hat{\pi}}^{RL}(\acute{s})\right) \quad (18)$$

Following the generation of these two policies, the proposed framework employs a comparison-based selection mechanism that chooses the optimal policy from the two available options, as defined in Equation (19):

$$\pi^*(s_t) = \begin{cases} \pi^{DP(s_t)} & \hat{C}^{DP} \leq \hat{C}^{RL} \\ \pi^{RL}(s_t) & \hat{C}^{RL} > \hat{C}^{DP} \end{cases} \quad (19)$$

In other words, at each period and based on real-time information and probabilistic forecasts, a dynamic selection is made between the two decision-making approaches. This adaptive selection constitutes the primary advantage of the hybrid framework over the exclusive use of either method alone.

To forecast future conditions and support the selection mechanism, Markov chain transition matrices and the concept of multi-period probability distribution are employed. The u-period transition probability from level i to level j for the s-th supplier is derived from the equation $p_{sij}{}^{u}$. This multi-period forecast enables the selection mechanism to consider not only the current state but also the probable state of the environment in future periods within the decision-making process.

The hybrid mechanism optimizes performance from two perspectives. In periods where environmental information is stable and reliable and transition matrices exhibit high accuracy, the dynamic programming policy gains superiority and more precise, theory-grounded decisions are made. Conversely, in periods where the environment undergoes structural changes or unexpected fluctuations and probabilistic model estimates lose accuracy, the reinforcement learning policy—which learns directly from interaction with reality—delivers superior performance. Accordingly, the hybrid framework automatically and adaptively selects the approach appropriate to the actual environmental conditions, simultaneously benefiting from both analytical and experiential sources of knowledge.

2.6 Data and inflation adjustment

The dataset used in this research comprises 188 weekly observations spanning from April 2022 to October 2025. The core structure of this dataset encompasses three categories of key information: recorded demand quantity in each period, ordering costs associated with two distinct suppliers, and inventory holding costs. The data were sourced from an actual operating industrial unit, and its operational information was employed as the basis for the empirical analysis of this research.

One of the fundamental challenges in economic data analysis over recent years is the presence of significant inflationary fluctuations, which may distort modeling results. To address this issue, all cost figures were converted to constant values using the Consumer Price Index (CPI), with April 2022 selected as the base period. The objective of this adjustment was to eliminate monetary effects attributable to inflation and obtain comparable real figures over time. The adjustment process was performed based on Equation (20):

$$C_t^{Real} = C_t^{nominal} \times \frac{CPI_{Base}}{CPI_{Curent}} \tag{19}$$

For the sake of brevity in presentation, only a sample of the data(including nominal values and adjusted values) is presented in Table 1.

Table 1. Sample of Raw Data: Initial Information, CPI Index, Nominal and Adjusted Costs per Period

Period	CPI	Demand	Nominal Cost Supplier 1	Nominal Cost Supplier 2	Nominal Holding Cost	Real Cost Supplier 1	Real Cost Supplier 2	Real Holding Cost
1	108	350	30,700	30,500	350	30,700	30,500	350
2	108	350	31,200	31,200	350	31,200	31,200	350
3	108	350	31,200	31,100	350	31,200	31,100	350
4	108	350	31,400	31,300	350	31,400	31,300	350
...	...	...	...	...	...	...	...	...
185	371	650	87,900	88,000	650	24,569.03	24,596.98	181.68
186	385	650	88,500	88,400	650	24,736.74	24,708.79	181.68
187	385	650	88,300	88,100	650	24,680.84	24,624.93	181.68
188	385	650	88,600	88,700	650	24,764.69	24,792.64	181.68

The results of the data examination indicate that despite a sustained upward trend in nominal costs over the study period, real cost values experienced declines at certain time intervals. This divergent pattern between nominal and real costs reflects structural transformations and changes in supply conditions and market pricing. Notably, in the terminal periods, despite a more than twofold increase in nominal cost figures relative to the initial periods, the adjusted cost values remained at approximately similar levels. This evidence clearly confirms the necessity and importance of employing adjusted data in the modeling and decision-making process, demonstrating that sole reliance on nominal figures may lead to incorrect conclusions regarding actual cost trends and, consequently, the adoption of inappropriate decisions.

2.7 Cost discretization and representative value extraction

A fundamental prerequisite for employing Markov stochastic process-based modeling is the conversion of continuous variables into discrete and categorizable values. In this research, continuous costs derived from

adjusted data were partitioned into three levels (Low (L), Medium (M), and High (H)) using a three-level mapping. The basis for this discretization was the relative changes in costs over time, and the Median Absolute Deviation (MAD) criterion was employed to ensure model stability and reduce the impact of outliers. By applying the conversion factor of 1.4826 (equivalent to the standard deviation ratio for the normal distribution), the acceptable volatility threshold is defined according to Equations (21) through (23):

$$MAD = Median(|x_i - Median(x)| \quad (20)$$

$$\varepsilon = 1.4826 \times MAD \quad (21)$$

The threshold conversion factor is given by Equation (23):

$$\Phi^{-1}(0.75) = 1.4826 \quad (23)$$

Cost level labeling based on this threshold is performed according to Equation (24):

$$Level_t = \begin{cases} L & \Delta_t < -\varepsilon \\ M & -\varepsilon \leq \Delta_t \leq \varepsilon \\ H & \Delta_t > \varepsilon \end{cases} \quad (24)$$

This approach preserves the statistical structure of the data while enabling probabilistic modeling of cost levels.

Subsequently, corresponding representative values were computed for each of the three identified levels. The calculation method for these values is based on a four-period rolling window that employs the 0.33, 0.66 percentiles and the mean as reference points. The use of a rolling window ensures that representative values at each point in time reflect the relative position of costs with respect to their recent behavior, while eliminating the effect of long-term trends that might provide a distorted picture of the current situation. Details of these values are presented in Table 2.

Table 2: Sample of Representative Data for Supplier Cost Levels

Horizon	Period	Supplier 1	Supplier 2	L Representative	M Representative	H Representative	L Representative	M Representative	H Representative
1	1	M	M	30,933.3	31,050	31,166.6	30,500	30,500	30,766.6
1	2	H	H	30,933.3	31,050	31,166.6	31,200	31,200	30,766.6
1	3	M	M	30,933.3	31,050	31,166.6	31,100	31,100	30,766.6
1	4	M	M	30,933.3	31,050	31,166.6	31,300	31,300	30,766.6
...	...	...	...	...	...	...	...	...	...
47	18	L	L	24,859.0	25,004.0	25,149.1	88,000	24,596.9	24,858.3
47	18	M	M	24,859.0	25,004.0	25,149.1	88,400	24,708.7	24,858.3
47	18	M	M	24,859.0	25,004.0	25,149.1	88,100	24,624.9	24,858.3
47	18	M	M	24,859.0	25,004.0	25,149.1	88,700	24,792.6	24,858.3

2.8 Transition probability modeling

Cost level transitions for each supplier were modeled as a Markov chain. Transition matrices were computed using empirical transition frequencies between states and row-wise normalization, in accordance with Equation (25):

$$P_s(i,j) = \frac{N_s(i,j)}{\sum_j N_s(i,j)} \qquad \forall i,j \epsilon \{L,M,H\} \tag{25}$$

The transition probability matrices for the suppliers are presented in Equations (26) and (27):

$$P_1(i,j) = \begin{bmatrix} 0.02778 & 0.80555 & 0.16667 \\ 0.23729 & 0.55932 & 0.20339 \\ 0.21212 & 0.69697 & 0.09091 \end{bmatrix} \tag{26}$$

$$P_2(i,j) = \begin{bmatrix} 0.02941 & 0.85294 & 0.11765 \\ 0.23140 & 0.56199 & 0.20661 \\ 0.15625 & 0.75 & 0.09375 \end{bmatrix} \tag{27}$$

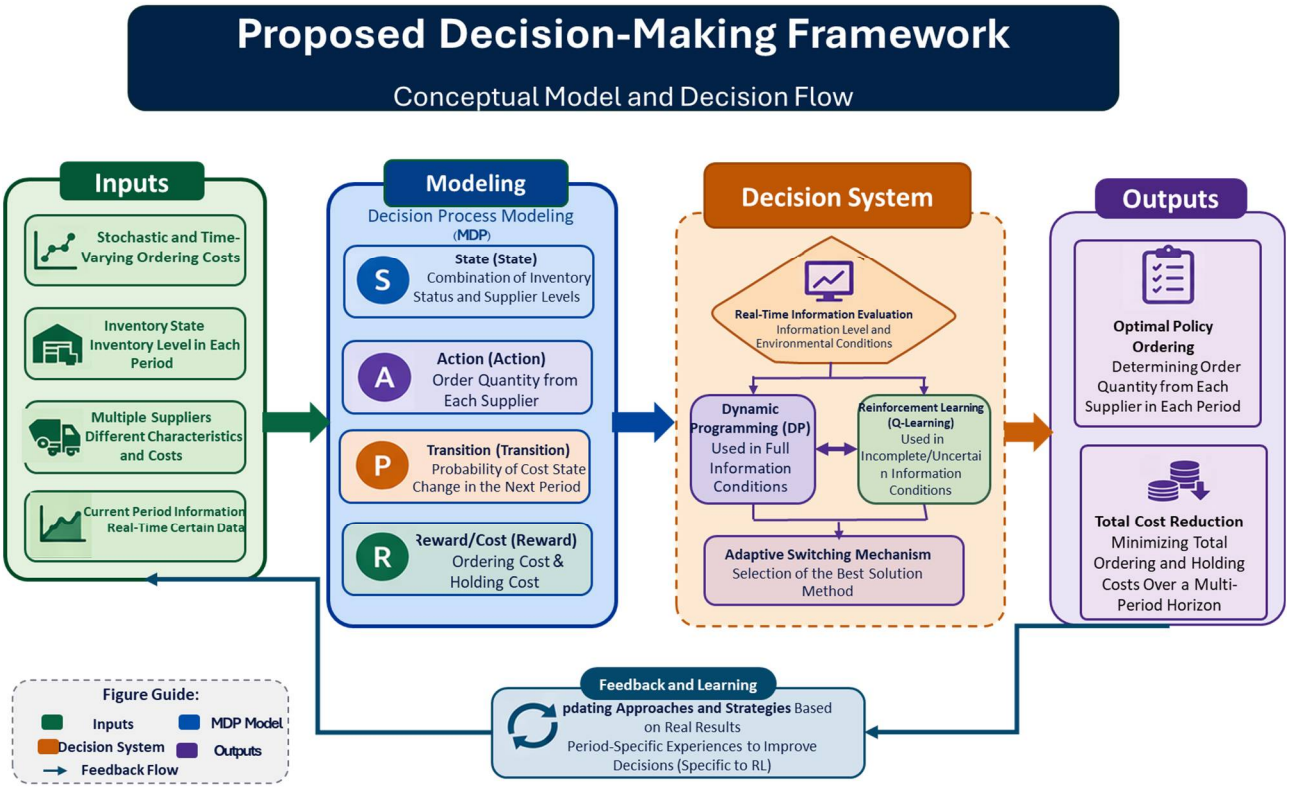


Figure 1 illustrates the conceptual model and decision-making flow

3. Results and discussion

To validate and assess the performance of the proposed framework, a set of decision-making policies was compared. These policies include the company's current policy as the baseline, the dynamic programming-based policy (DP), the reinforcement learning-based policy (QL), and finally, the hybrid policy derived from combining DP and QL. The analyses are based on a four-period planning horizon with two suppliers considered.

3.1 Total cost comparison of policies

Table 3: Sample of Total Cost Comparison for Three Policies Across 47 Decision-Making Horizons

Horizon	Total Cost (Current)	Total Cost (DP)	Total Cost (QL)	Total Cost (Hybrid)	Savings (Current-DP)	Savings (Current-QL)	Difference (QL-DP)
1	7,235,000	7,192,150	7,221,200	7,192,150	42,850	13,800	29,050
2	11,412,300	11,294,500	11,288,200	11,288,200	117,800	124,100	−6,300
3	17,656,700	17,604,600	17,614,000	17,604,600	52,100	42,700	9,400
4	7,235,000	7,192,150	7,221,200	7,192,150	42,850	13,800	29,050
...	...	...	...	...	...	...	...
46	50,162,100	50,138,100	50,138,100	50,138,100	24,000	24,000	0
47	49,282,400	49,247,000	49,247,000	49,247,000	35,400	35,400	0

The results of the total cost comparison for the four evaluated policies are presented in Table 3. Descriptive statistics of these results are summarized in Table 4 and Figure 2.

Table 4. Descriptive statistics of costs for the three policies

Table 4. Descriptive Statistics of Costs for the Three Policies

Policy	Mean Cost	Maximum Cost	Minimum Cost	Standard Deviation
Total Cost (Current)	31,802,070	50,162,100	7,235,000	9,824,365
Total Cost (DP)	31,728,920	50,138,100	7,192,150	9,823,718
Total Cost (QL)	31,735,479	50,138,100	7,221,200	9,824,304

The total cost comparison results indicate that across all 47 planning horizons, all three optimization approaches (DP, QL, and hybrid) recorded lower costs than the company's current policy. The hybrid policy, with a mean cost of 31,727,902 monetary units, demonstrated the best performance among all policies, achieving an average saving of 0.23% relative to the baseline policy.

Another notable finding is that in most horizons, particularly in the terminal periods of the dataset, the DP and QL policies achieved identical results. Furthermore, the hybrid policy consistently achieved at least as much saving as the DP policy, and in cases where QL performed inferior to DP (e.g., Horizons 2, 5, and 37), the hybrid policy prevented performance degradation through intelligent selection of the DP policy. This clearly demonstrates the primary advantage of the adaptive selection mechanism embedded in the hybrid framework.

The maximum saving in certain horizons reached 0.87% (Horizon 21). The standard deviation of costs across all four policies was approximately equal, at around 9.8 million units, indicating similar stability in terms of cost dispersion across different horizons. Therefore, both optimization methods (DP and Q-learning) outperform the current policy, and despite the seemingly modest percentage savings, given the magnitude of cost values, the absolute savings are substantial (averaging more than 73,150 monetary units per horizon).

Figure 1 illustrates the trend of total cost changes for the four policies across different horizons. As shown, all three optimization policies consistently incur lower costs than the baseline policy. The difference between DP and QL was negligible in most horizons, and the hybrid policy performed at or better than the best possible alternative in all cases.

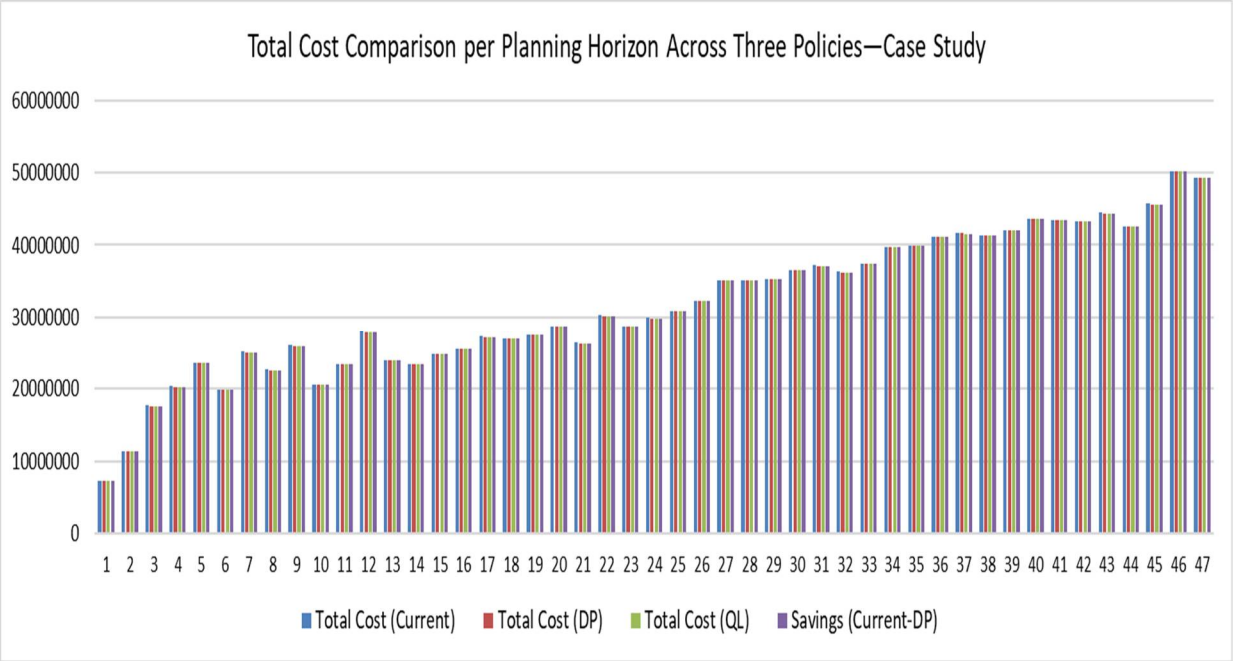


Figure 2: Comparative Chart of Policy Total Costs

3.2 Order allocation analysis across suppliers

One of the critical dimensions of policy performance evaluation is the allocation of orders between the two suppliers. As observed in Tables 5 and 6 and Figure 2, the four compared policies (baseline, dynamic programming, reinforcement learning, and hybrid) exhibit distinct behavioral patterns in order allocation.

Table 5: Order Quantity Allocation to Suppliers in Each Period

Horizon	Period	Demand	R_1	R_2	DP_1	DP_2	QL_1	QL_2	H_x1	H_x2
1	1	71	36	35	0	71	0	71	0	154
1	2	83	0	83	0	0	58	25	58	25
1	3	33	20	13	0	33	0	33	0	33
1	4	46	46	0	0	46	0	46	0	46
2	1	44	20	24	0	44	67	29	67	29
2	2	52	25	27	52	0	0	0	52	0
2	3	121	70	51	0	121	0	121	0	121
2	4	133	133	0	66	67	133	0	133	0
...	...	...	...	...	...	...	...	...	...	...
47	1	132	80	52	132	0	132	0	132	0
47	2	142	80	62	0	142	0	142	0	142
47	3	142	80	62	0	142	0	142	0	142
47	4	142	80	62	142	0	142	0	142	0

Table 6: Order Allocation Percentage to Suppliers in Each Period

Horizon	Period	Demand	R_1	R_2	DP_1	DP_2	QL_1	QL_2	H_x1	H_x2
1	1	71	51%	49%	0%	100%	0%	100%	0%	100%
1	2	83	0%	100%	0%	0%	70%	30%	70%	30%
1	3	33	61%	39%	0%	100%	0%	100%	0%	100%
1	4	46	100%	0%	0%	100%	0%	100%	0%	100%
2	1	44	45%	55%	0%	100%	70%	30%	70%	30%
2	2	52	48%	52%	100%	0%	0%	0%	100%	0%
2	3	121	58%	42%	0%	100%	0%	100%	0%	100%
2	4	133	100%	0%	50%	50%	100%	0%	100%	0%
...	...	...	...	..	...	...	...	...	...	...
47	1	132	61%	39%	100%	0%	100%	0%	100%	0%
47	2	142	56%	44%	0%	100%	0%	100%	0%	100%
47	3	142	56%	44%	0%	100%	0%	100%	0%	100%
47	4	142	56%	44%	100%	0%	100%	0%	100%	0%

Examination of allocation patterns across the four studied policies reveals significant structural differences. The baseline policy (current company practice) follows a relatively stable, predetermined approach in which order allocations are generally carried out in an approximately uniform manner based on fixed quota proportions. This approach exhibits lower flexibility in responding to cost level changes and predominantly follows a non-adaptive pattern.

In contrast, dynamic programming determines allocation in a state-dependent manner proportional to the cost conditions of each period; when one supplier exhibits a clear cost advantage, order concentration shifts notably toward that supplier. This approach demonstrates active exploitation of the relative cost advantage available in the supply market.

The reinforcement learning policy also reproduced allocation patterns similar to dynamic programming in many horizons, and occasionally exhibited a mixed allocation between the two suppliers. This indicates that the learning agent was able to extract optimal allocation patterns from environmental interaction.

The hybrid policy, as an intelligent integrator of the DP and QL approaches, demonstrated notable performance. The results indicate that in most cases, the hybrid policy adopted allocations similar to the best choice between DP and QL.

Overall, the behavior of the three data-driven policies (DP, QL, and hybrid) demonstrates the dependence of allocation decisions on the cost conditions of each period. These policies shift order concentration toward the supplier exhibiting a cost advantage when such an advantage is present, thereby exploiting relative cost benefits. Conversely, the current policy exhibits less inclination toward such concentration, and its allocation pattern remains relatively fixed and independent of cost conditions.

In summary, employing dynamic data-driven decision-making frameworks, particularly the proposed hybrid framework, can transform the allocation process from an experiential and intuitive procedure into a systematic, adaptive, and optimal decision rule. Figure 2 provides a visual comparison of allocation percentages to suppliers across all four policies.

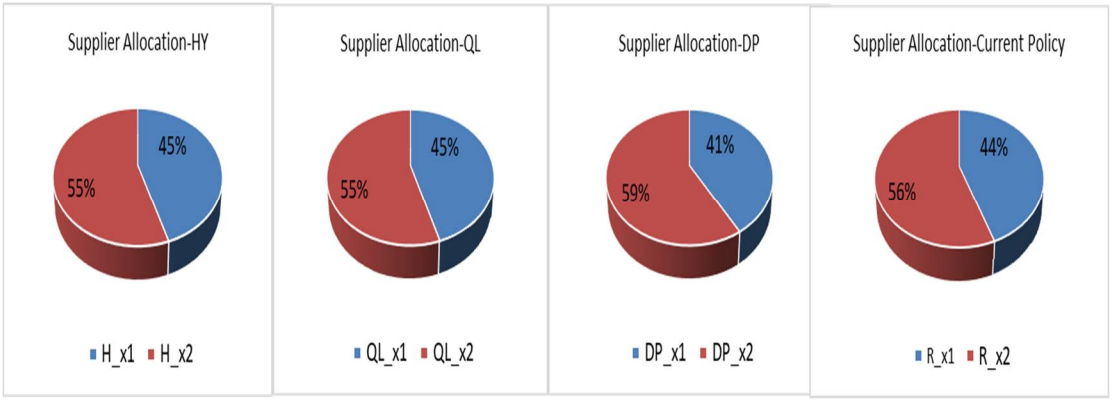


Figure 3: Supplier Allocation Percentage for Each Policy

To further validate the proposed framework under near-real-world conditions, the evaluation process was conducted sequentially across multiple consecutive planning horizons. Specifically, three new planning horizons were considered in a sequential manner, as presented in Table 7, and at each stage, the model decision was made based on limited information at the beginning of the horizon, after which the obtained results were incorporated into the existing dataset. This evaluation structure aligns with the logic of decision-making in dynamic operational environments and enables the examination of policy stability and generalizability over time.

Table 7. New horizons for sequential evaluation under limited information

Horizon	Demand (Periods 1–4)	Supplier 1 Ordering Costs	Supplier 2 Ordering Costs	Holding Cost
1	150	92000	93000	650
	150	-	-	650
	150	-	-	650
	148	-	-	650
2	145	102000	103000	650
	143	-	-	650
	145	-	-	650
	142	-	-	650
3	152	105000	105000	650
	153	-	-	650
	155	-	-	650
	150	-	-	650

In each of these horizons, available information included demand for all periods, holding cost, and first-period ordering costs. Future-period costs were estimated using the forecasting mechanisms of both the dynamic programming and Q-learning approaches. Subsequently, both methods independently derived their respective optimal policies, and within the hybrid decision-making framework, the best policy was selected based on the lowest expected cost.

The results from executing this process across three consecutive horizons demonstrated that in all cases, both methods converged to structurally similar policies, and the adopted decisions exhibited high stability and consistency. A comparative examination against actual data also indicated that the derived policies generally possessed appropriate accuracy and aligned with the real optimal choice in most cases. To provide a more in-depth analysis, the results pertaining to the first planning horizon are described below as a representative example, as presented in Table 8.

Table 8: Results from Various Policies in New Horizons

Horizon	Period	Demand	Dynamic Programming				Q-Learning				Hybrid Policies			
			Total Order	Allocation S_1	Allocation S_2	λ	Total Order	Allocation S_1	Allocation S_2	λ	Total Order	Allocation S_1	Ending Inventory	λ

New 1	1	150	150	150	0	1	150	150	0	1	150	0	0	1
	2	150	150	150	0	0	150	150	0	1	150	0	0	1
	3	150	150	150	0	1	150	150	0	1	150	0	0	1
	4	158	148	148	0	1	148	148	0	1	148	0	0	1

In the first horizon, initial information included demands of (150, 150, 150, 148 units) and first-period ordering costs for the two suppliers (92,000 and 93,000). Based on this information and future cost forecasts, both dynamic programming and Q-learning uniformly selected Supplier 1 as the optimal option, with λ set to 1. Additionally, order quantities in all periods were set exactly equal to the demand of that period, with ending inventory at zero, reflecting the adoption of a lot-for-lot ordering policy in the presence of positive holding costs.

Comparison of these results with the actual data presented in Table 9 indicates that in Period 1, selecting Supplier 1 was fully optimal, as its actual cost was lower than that of Supplier 2. In Period 2, with equal costs, selecting Supplier 1 remained economically acceptable. In Periods 3 and 4, the actual cost of Supplier 1 was also lower than that of Supplier 2, and the model's decisions aligned with the real optimal choice in all periods. Therefore, in this horizon, the derived policy was not only structurally logical but also fully accurate in terms of actual performance.

Table 9: Actual Ordering Costs in Reality

Horizon	Period	Demand (Units)	Supplier 1 Ordering Cost	Supplier 2 Ordering Cost	Holding Cost
New 1	1	150	92,000	93,000	650
	2	150	92,500	92,500	650
	3	150	96,000	97,000	650
	4	148	95,800	97,000	650
New 2	1	145	102,000	103,000	650
	2	143	10,500	103,800	650
	3	145	104,000	104,000	650
	4	142	104,800	104,000	650
New 3	1	15	105,000	105,000	650
	2	153	107,000	108,000	650
	3	155	110,000	109,000	650
	4	150	110,000	110,000	650

A notable observation is that in subsequent horizons, despite increased cost levels and relative changes between suppliers, the proposed framework was able to make stable, economically rational decisions without significant fluctuations. This demonstrates that even under conditions where forecasts are accompanied by uncertainty, the model's decision structure possesses sufficient robustness, minimizing the risk of incorrect decisions.

Overall, the results of this sequential evaluation demonstrate that the limited-information-based predictive framework, in combination with the adaptive decision-making mechanism, is capable of deriving stable, reliable, and near-optimal policies within a dynamic, multi-period process. This finding substantially confirms the practical applicability of this framework in real-world inventory management problems characterized by information uncertainty.

4. Conclusion

This research proposed a hybrid decision-making framework for the multi-period ordering problem from multiple suppliers in an environment with limited information. The principal contributions of this research can be summarized in four dimensions:

(a) Markovian modeling and inter-period dynamics: Formulating the ordering problem as a Markov decision process enabled the simultaneous consideration of temporal dynamics and inter-period decision dependencies. This characteristic addresses the gap in classical models based on static assumptions and provides a more realistic framework for analyzing system behavior.

(b) Integration of model-based and data-driven approaches: Combining dynamic programming (as a model-based approach) and Q-learning (as a model-free approach) created a novel paradigm in decision-making under uncertainty. This integration not only reduces the individual limitations of each approach but also employs analytical accuracy and learning flexibility simultaneously.

(c) Superior performance of the hybrid framework: The experimental results demonstrated that the adaptive selection accuracy rate of the hybrid framework was 100%, such that in all 26 horizons where DP and QL policies exhibited cost differences, the hybrid mechanism correctly identified and selected the lower-cost policy. The average saving of the hybrid framework relative to the company's current policy was 0.23%, reaching as high as 0.87% in certain horizons. Given the magnitude of cost values, the absolute savings are substantial and practically meaningful.

(d) Demonstration of Q-learning capability in information-poor environments: Despite its model-free nature and lack of initial knowledge of transition matrices, Q-learning-based reinforcement learning produced policies of quality very close to the optimal solutions derived from dynamic programming. This finding reinforces the validity of employing learning methods in complex problems with incomplete information.

Overall, by bridging classical optimization approaches and modern data-driven methods, this research contributes to the development of the dynamic decision-making and supply chain inventory management literature.

5. Recommendations

5.1 Research recommendations

Given the limitations identified in this research and the research opportunities ahead, several avenues for further development and refinement of the present findings are proposed. First, replacing the Q-learning algorithm with more advanced deep reinforcement learning approaches such as DQN, actor-critic, and PPO can enable more effective management of larger and continuous state spaces, substantially improving policy learning accuracy in high-dimensional problems. Second, extending the model to multi-product scenarios that incorporate synergistic interactions and interferences between different products represents a necessary step toward closer alignment with industrial realities. Third, adding more realistic constraints such as supplier capacity limits, varying lead times, and variable transportation costs will enhance the realism of the results. Fourth, integrating risk measures such as Value at Risk (VaR) and Conditional Value at Risk (CVaR) into the objective function will enable risk-aware decision-making in critical and volatile conditions. Fifth, designing a dynamic weighted allocation mechanism that, rather than exclusively selecting one policy, adopts a weighted combination of dynamic programming and reinforcement learning policies based on the reliability of each method in each decision period can elevate the hybrid framework's efficiency to a higher level. Finally, examining the framework's performance under structural changes such as the entry or exit of a new supplier, abrupt shifts in the cost structure, or supply chain disruptions is of significant importance in terms of model stability and generalizability in non-stationary environments.

5.2 Practical recommendations

Phased implementation: Companies can implement the hybrid framework initially on a pilot scale (limited number of suppliers and periods) and gradually expand it upon confirming performance.

Periodic update of transition matrices: As market conditions evolve, regular updates of Markov chain transition matrices are essential to maintain forecast accuracy.

Use of adjusted data: As the results demonstrated, reliance on nominal figures without inflation adjustment can lead to incorrect decision-making; therefore, the use of price adjustment indices is recommended.

Training of decision-making personnel: Familiarizing operational teams with the basic concepts of reinforcement learning and system dynamics will facilitate the successful adoption and implementation of the framework.

6. Summary

This research proposed a hybrid decision-making framework for the multi-period ordering problem from multiple suppliers in an environment with limited information. To this end, the problem was first formulated as a Markov decision process, and two solution approaches—including dynamic programming (as a model-based method) and the Q-learning algorithm (as a model-free method)—were applied in parallel. The primary contribution of this research was the design of an adaptive selection mechanism that, in each period and based on available information and probabilistic forecasts, selected the policy with the lowest expected cost from the two aforementioned alternatives. The numerical results from implementation on 188 weekly observations from an actual industrial unit demonstrated that the hybrid framework outperformed the company's current policy, with average savings of 0.23% (and a maximum of 0.87% in certain horizons). Furthermore, the adaptive selection accuracy rate was 100%, and the hybrid framework successfully identified the optimal policy in all 26 horizons where DP and QL policies exhibited cost differences. Additionally, the findings indicated that the optimized policies determined order concentration proportional to the suppliers' relative cost advantages and provided adaptive and dynamic allocation patterns; in contrast, the company's current policy exhibited an approximately fixed and non-adaptive pattern. Overall, this research demonstrated that the intelligent combination of model-based and data-driven methods provides an efficient and practically reliable approach for decision-making in complex and

uncertain supply chain environments, and can offer practical guidance for industrial decision-makers in real-world conditions.

Declarations

Data Availability Statement

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethical Approval and accordance

Not applicable. This study did not involve human participants, their data, or live vertebrates.

Consent to participate

Not applicable. No human subjects were involved in this research.

Consent to publish

Not applicable. This manuscript does not contain any individual person's data in any form.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Azaron, A., Tang, O., & Tavakkoli-Moghaddam, R. (2009). Dynamic lot sizing problem with continuous-time Markovian production cost. *Int. J. Prod. Econ.*, *120*(2), 607–612. https://doi.org/https://doi.org/10.1016/j.ijpe.2009.04.007

Chen, Z., & Rossi, R. (2021). A dynamic ordering policy for a stochastic inventory problem with cash constraints. *Omega*, *102*, 102378. https://doi.org/https://doi.org/10.1016/j.omega.2020.102378

Cheraghalipour, A., & Farsad, S. (2018). A bi-objective sustainable supplier selection and order allocation considering quantity discounts under disruption risks: A case study in plastic industry. *Computers & industrial engineering*, *118*, 237–250. https://doi.org/https://doi.org/10.1016/j.cie.2018.02.041

Darezereshki, F., & Derhami, V. (2024). A Reinforcement Learning Approach to Determine When and How Many Stocks to Buy in Stock Trading.

Geevers, K., Van Hezewijk, L., & Mes, M. R. (2024). Multi-echelon inventory optimization using deep reinforcement learning. *Central European Journal of Operations Research*, *32*(3), 653–683. https://doi.org/https://doi.org/10.1007/s10100-023-00872-2

Gholamazad, M., & Eydi, A. (2026). A hybrid data envelopment analysis and artificial intelligence framework for sustainable supplier selection: a case study in the petrochemical industry. *Ain Shams Engineering Journal*, *17*(1), 103871. https://doi.org/https://doi.org/10.1016/j.asej.2025.103871

Hosseini, L., Fallah Nezhad, M. S., Derhami, V., & Owlia, M. S. (2026). A multi-stage ordering model for multi-supplier systems via Markov decision process. *Communications in Statistics-Theory and Methods*, *55*(11), 3241–3255. https://doi.org/20.1001.1.20089198.1404.27.87.3.7

Hosseini, Z. S., Flapper, S. D., & Pirayesh, M. (2022). Sustainable supplier selection and order allocation under demand, supplier availability and supplier grading uncertainties. *Comput. Ind. Eng.*, *165*, 107811. https://doi.org/https://doi.org/10.1016/j.cie.2021.107811

Kawakami, K., Kobayashi, H., & Nakata, K. (2021). Seasonal inventory management model for raw materials in steel industry. *INFORMS Journal on Applied Analytics*, *51*(4), 312–324. https://doi.org/https://doi.org/10.1287/inte.2021.1073

Kharvi, S., & Pakkala, T. (2022). An optimal ordering inventory policy when the purchase price follows a discrete-time Markov chain. *Commun. Stat. Simul. Comput.*, 1–18. https://doi.org/https://doi.org/10.1080/03610918.2022.2158340

Li, Q., Wu, X., & Cheung, K. L. (2009). Optimal policies for inventory systems with separate delivery-request and order-quantity decisions. *Operations Research*, *57*(3), 626–636. https://doi.org/https://doi.org/10.1287/opre.1090.0696

Liu, X., Hu, M., Peng, Y., & Yang, Y. (2025). Multi-agent deep reinforcement learning for multi-echelon inventory management. *Production and Operations Management*, *34*(7), 1836–1856. https://doi.org/https://doi.org/10.1177/10591478241305863

Luo, S., Ahiska, S. S., Fang, S.-C., King, R. E., Warsing Jr, D. P., & Wu, S. (2021). An analysis of optimal ordering policies for a two-supplier system with disruption risk. *Omega*, *105*, 102517. https://doi.org/https://doi.org/10.1016/j.omega.2021.102517

Maulidi, I., Radhiah, R., Hayati, C., & Apriliani, V. (2022). Optimal Raw Material Inventory Analysis Using Markov Decision Process with Policy Iteration Method. *JTAM 6*(3), 638–650. https://doi.org/https://doi.org/10.31764/jtam.v6i3.8563

Piao, M., Zhang, D., Lu, H., & Li, R. (2023). A supply chain inventory management method for civil aircraft manufacturing based on multi-agent reinforcement learning. *Applied Sciences*, *13*(13), 7510. https://doi.org/https://doi.org/10.3390/app13137510

Powell, W. B., Shapiro, J. A., & Simão, H. P. (2002). An adaptive dynamic programming algorithm for the heterogeneous resource allocation problem. *Transportation Science*, *36*(2), 231–249. https://doi.org/https://doi.org/10.1287/trsc.36.2.231.561

Scarf, H. (1960). The optimality of (S, s) policies in the dynamic inventory problem. https://doi.org/https://doi.org/10.1057/9781137024381_1

Shidpour, H., Karimi, N., Baryannis, G., & Shidpour, M. (2025). A multi-phase analytics framework for supply chain supplier selection and order allocation with delay risks and Industry 4.0 readiness. *Supply Chain Analytics*, 100172. https://doi.org/https://doi.org/10.1016/j.sca.2025.100172

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT press Cambridge. https://doi.org/978-0-262-19398-6

Thompson, S., Nunez, M., Garfinkel, R., & Dean, M. D. (2009). OR practice—efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research*, *57*(2), 261–273. https://doi.org/http://dx.doi.org/10.1287/opre.1080.0584

Van Roy, B., Bertsekas, D. P., Lee, Y., & Tsitsiklis, J. N. (1997). A neuro-dynamic programming approach to retailer inventory management. Proceedings of the 36th IEEE Conference on Decision and Control,

Wagelmans, A., Van Hoesel, S., & Kolen, A. (1992). Economic lot sizing: An O (n log n) algorithm that runs in linear time in the Wagner-Whitin case. *Oper. Res.*, *40*(1-supplement-1), S145–S156. https://doi.org/https://doi.org/10.1287/opre.40.1.S145

Wagner, H. M., & Whitin, T. M. (1958). Dynamic version of the economic lot size model. *Manag. Sci.*, *5*(1), 89–96. https://doi.org/https://www.jstor.org/stable/30046142

Wagner, H. M., & Whitin, T. M. (2004). Dynamic version of the economic lot size model. *Manag. Sci.*, *50*(12_supplement), 1770–1774. https://doi.org/https://doi.org/10.1287/mnsc.5.1.89

Zhang, H., Li, N., & Lin, J. (2024). Modeling the Decision and Coordination Mechanism of Power Battery Closed-Loop Supply Chain Using Markov Decision Processes. *Sustainability*, *16*(11), 4329. https://doi.org/https://doi.org/10.3390/su16114329

Zhao, G., & Sun, R. (2010). Application of multi-agent reinforcement learning to supply chain ordering management. 2010 Sixth International Conference on Natural Computation,

Zhou, Y., Guo, K., Yu, C., & Zhang, Z. (2024). Optimization of multi-echelon spare parts inventory systems using multi-agent deep reinforcement learning. *Applied Mathematical Modelling*, *125*, 827–844. https://doi.org/https://doi.org/10.1016/j.apm.2023.10.039

Ziari, M., Soheyli, A., & Taleizadeh, A. A. (2026). An Intelligent Analytics Framework for Supplier Selection and Delay Risk Management in Sustainable Supply Chains. *Supply Chain Analytics*, 100208. https://doi.org/https://doi.org/10.1016/j.sca.2026.100208