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.

Deep Reinforcement Learning for Dynamic Inventory Control in Supply Chain

Ching Shih TSOU

cstsou@ntub.edu.tw

https://orcid.org/0000-0001-8937-1729

Yun-Shan Lee

National Taipei University of Business

Research Article

Keywords: Deep Reinforcement Learning, Inventory Optimization, Supply Chain Management, Multi-Echelon Systems

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11761558/v1

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

Deep Reinforcement Learning for Dynamic Inventory Control in Supply Chain

Ching-Shih Tsou cstsou@ntub.edu.tw and Yun-Shan Lee deby308698@gmail.com

Institute of Information and Decision Sciences, National Taipei University of Business, Taipei 10051, Taiwan

ARTICLE HISTORY
Compiled November 16, 2026

ABSTRACT
Inventory control in modern supply chains involves complex sequential decision-making under demand uncertainty, replenishment delays, and service-level constraints. Traditional analytical models often rely on restrictive assumptions that limit their applicability in high-dimensional and dynamic environments.

This study investigates the application of Deep Reinforcement Learning (DRL) for supply chain inventory optimization through a two-phase experimental framework. In Phase I, value-based and policy-gradient DRL algorithms are evaluated in a single-SKU single-echelon inventory system under different demand fulfillment mechanisms, including complete backordering, lost sales, and partial backordering. In Phase II, the environment is extended to a multi-SKU two-echelon supply chain, where replenishment and allocation decisions must be jointly optimized in a continuous action space.

The experimental results demonstrate that DRL-based inventory policies can effectively reduce total costs and improve service performance compared with traditional rule-based benchmarks. Value-based methods exhibit strong learning stability in discrete decision settings, while policy-gradient approaches provide superior robustness in continuous and high-dimensional environments. The findings highlight the potential of DRL as a scalable and data-driven approach for adaptive inventory control in modern supply chains.

KEYWORDS
Deep Reinforcement Learning; Inventory Optimization; Supply Chain Management; Multi-Echelon Systems

Practitioner Summary

Managing inventory in modern supply chains is increasingly difficult due to fluctuating demand and supply delays. Traditional inventory rules often struggle to adapt to these dynamic conditions. This study explores how artificial intelligence—specifically Deep Reinforcement Learning (DRL)—can automate and optimize inventory decisions. We tested several DRL algorithms using real-world supply chain data, moving from a simple single-warehouse setup to a complex network with multiple products and locations.

Our findings offer a practical guide for supply chain managers: when maintaining high service levels (avoiding stockouts) is the top priority, the DDQN algorithm per-

forms best. However, for complex networks requiring continuous coordination of both ordering and allocating stock across multiple locations, the PPO algorithm is highly effective at reducing overall costs and improving inventory turnover. Furthermore, combining DRL with short-term demand forecasting allows the system to proactively adjust to market changes. Ultimately, this research provides supply chain practitioners with a scalable, data-driven approach to replace rigid traditional rules, enabling more resilient, cost-effective, and automated inventory management.

1. Introduction

1.1. Research Background and Motivation

Inventory optimization is a fundamental component of supply chain operations, where firms must balance holding costs, shortage risks, and service-level requirements under uncertain market conditions. Classical inventory control models, such as the Economic Order Quantity (EOQ) framework and *(s, S)*-type policies, have provided analytically tractable solutions for decades and remain widely taught and applied in practice. For more complex situations, the seminal work of Clark and Scarf established optimal policies for multi-echelon inventory systems and laid the theoretical foundation for coordinated replenishment across supply chain tiers [1].

Despite their theoretical elegance, classical inventory models typically rely on restrictive assumptions, including stationary demand distributions, known lead times, and simplified system dynamics. As modern supply chains become increasingly complex, these assumptions are often violated in practice. Demand patterns may exhibit non-stationarity, abrupt fluctuations, and seasonality, while replenishment lead times are subject to upstream disruptions and capacity constraints. Under such conditions, exact analytical solutions quickly become intractable as the dimensionality of the system increases, limiting the effectiveness of traditional rule-based inventory policies.

Recent advances in Deep Reinforcement Learning (DRL) have provided a promising alternative to address sequential decision-making problems in complex and uncertain environments. By combining reinforcement learning with deep neural networks, DRL enables agents to learn effective decision policies directly through interaction with the environment, without requiring explicit probabilistic demand models or simplified structural assumptions. A comprehensive roadmap by Boute et al. highlighted the growing interest in applying DRL to inventory control problems and identified scalability, stability, and evaluation as key research challenges [2].

Early empirical evidence suggests that DRL-based inventory policies can outperform traditional heuristics in dynamic settings. For example, Oroojlooyjadid et al. demonstrated the effectiveness of Deep Q-Networks (DQN) in the classic beer game, showing improved cost and service performance compared to standard ordering rules [3]. Subsequent studies further explored DRL applications in lost-sales conditions, dual-sourcing environments, and multi-echelon structures, reinforcing the potential of data-driven inventory control approaches [4].

1.2. Research Objectives and Problem Definition

Although prior studies have demonstrated the feasibility of DRL for inventory control, several limitations remain. Many existing works focus on stylized environments, single-SKU systems, or synthetically generated demand data, limiting their applicability to

real-world supply chain operations. Moreover, systematic comparisons across different DRL algorithm families—such as value-based and policy-gradient methods—under varying demand-handling mechanisms and system complexities are still relatively scarce.

To address these limitations, this study investigates the application of DRL methods for inventory optimization in increasingly realistic supply chain environments. The primary objectives of this research are as follows:

(1) Develop a simulation-based inventory environment capable of capturing realistic demand uncertainty, replenishment delays, and cost structures.
(2) Implement and compare multiple DRL algorithms, including **Deep Q-Network (DQN)**, **Proximal Policy Optimization (PPO)**, and **Deep Deterministic Policy Gradient (DDPG)**, to evaluate learning efficiency, convergence behavior, and policy stability.
(3) Assess the performance of learned inventory policies using multiple operational metrics, such as total cost, service level, and inventory efficiency, rather than relying on a single criterion.
(4) Examine the impact of increasing system complexity by extending the analysis from a single-SKU single-echelon system to a multi-SKU two-echelon supply chain setting.

Through these objectives, this study aims to bridge the gap between recent methodological advances in DRL and the practical requirements of supply chain inventory management.

1.3. Research Methodology and Framework Overview

This study adopts a two-phase experimental research framework. Phase I focuses on a single-SKU single-echelon inventory system to validate the fundamental learning behavior of DRL algorithms under different demand fulfillment mechanisms, including complete backordering, lost sales, and partial backordering. These settings allow for controlled analysis of algorithmic stability and policy characteristics under varying shortage-handling rules.

Phase II extends the experimental environment to a multi-SKU two-echelon supply chain, where a central warehouse must coordinate replenishment and inventory allocation decisions across multiple downstream nodes. This setting introduces a high-dimensional continuous action space, motivating the use of policy-gradient and actor–critic DRL algorithms. The experimental data are derived from a dynamic real-world inventory dataset, enabling empirical evaluation under realistic demand variability [5].

All environments are implemented in Python following the modular architecture of the OpenAI Gym framework, which supports standardized reinforcement learning training and evaluation. This design ensures reproducibility and facilitates systematic comparison across algorithms and system configurations.

The remainder of this paper is organized as follows. Section 2 reviews related literature on inventory control and DRL applications. Section 3 presents the research methodology and experimental design. Section 4 discusses the experimental results and analysis. Finally, Section 6 concludes the paper and outlines future research directions.

2. Literature Review

2.1. Classical Inventory Control and Multi-Echelon Systems

Inventory control has long been a core research area in production and operations management, with early studies focusing on analytically tractable policies under clearly defined assumptions. Among the most influential contributions, Clark and Scarf developed optimal policies for multi-echelon inventory systems and introduced the concept of echelon stock, providing a theoretical foundation for coordinated replenishment decisions across supply chain tiers [1]. This framework has shaped subsequent research on hierarchical inventory systems and remains a benchmark for evaluating new inventory control approaches.

Although classical inventory models offer valuable structural insight, they generally assume stationary demand, known lead times, and limited system dimensionality. As supply chains expand in scope and complexity, these assumptions become increasingly restrictive, reducing the practical applicability of closed-form analytical solutions in dynamic and high-dimensional environments.

2.2. Deep Reinforcement Learning for Inventory Control

Recent advances in Deep Reinforcement Learning (DRL) have enabled new approaches to inventory decision-making that move beyond traditional analytical models. By integrating reinforcement learning with deep neural networks, DRL allows agents to approximate optimal policies in complex state spaces through direct interaction with the environment. Boute et al. provided a comprehensive roadmap of DRL applications in inventory control, highlighting methodological challenges related to reward design, stability, scalability, and benchmarking [2].

One of the earliest and most influential DRL applications in inventory management is the classic beer game. Oroojlooyjadid et al. demonstrated that Deep Q-Networks (DQN) can effectively learn ordering policies that outperform traditional heuristics under demand uncertainty and delayed information feedback [3]. Their findings provided strong empirical evidence supporting the feasibility of value-based DRL methods for inventory optimization.

2.3. Lost Sales, Demand Uncertainty, and Non-Stationary Environments

Beyond backordering settings, several studies have examined DRL performance in more challenging inventory environments characterized by lost sales and uncertain demand. Gijsbrechts et al. evaluated DRL algorithms across lost-sales systems, dual-sourcing strategies, and multi-echelon problems, showing that DRL-based policies can outperform classical benchmarks when demand variability is high [4]. Their results indicate that DRL can adapt to operational conditions where traditional safety-stock-based approaches struggle.

More recently, Dehaybe et al. explicitly considered non-stationary and uncertain demand patterns, demonstrating that DRL-based inventory policies maintain robust performance even when demand distributions evolve over time [6]. These findings further highlight the suitability of DRL for real-world supply chain environments, where demand dynamics are rarely stable.

2.4. Multi-Echelon and Warehouse–Retailer Inventory Systems

The application of DRL to multi-echelon and warehouse–retailer inventory systems has attracted increasing research attention. Kaynov et al. studied a one-warehouse multi-retailer system and showed that DRL can effectively coordinate replenishment and allocation decisions across multiple downstream nodes [7]. Their results suggest that DRL-based policies can exploit information sharing and network structure to improve overall system performance.

Similarly, Stranieri and Stella compared multiple DRL algorithms in two-echelon supply chains and analyzed their convergence behavior, stability, and final performance [8]. Their study highlights the trade-offs between value-based and policy-gradient methods in high-dimensional action spaces, providing important insights for algorithm selection in multi-echelon inventory environments.

2.5. DRL Applications in Industrial Supply Chains

Beyond theoretical and stylized settings, DRL has also been applied to practical industrial supply chain problems. Chong et al. demonstrated the effectiveness of DRL in optimizing apparel supply chains, achieving improvements in cost efficiency and service performance compared with conventional planning methods [9]. Bermúdez et al. proposed a distributionally constrained reinforcement learning framework that explicitly accounts for uncertainty and risk in supply chain optimization, further extending the applicability of DRL in industrial contexts [10].

These studies illustrate the growing interest in deploying DRL-based decision-making systems in real operational environments, while also highlighting the need for systematic evaluation under realistic data conditions.

2.6. Research Gaps and Positioning of This Study

Despite rapid progress, several research gaps remain. First, many DRL-based inventory studies rely on synthetic demand data or simplified network structures, limiting empirical validation under realistic operating conditions. Second, relatively few studies provide systematic comparisons across different DRL algorithm families under varying demand-handling mechanisms and system complexities. Third, integrated evaluation frameworks that jointly consider cost efficiency, service level, and inventory stability remain limited.

To address these gaps, this study develops a two-phase DRL-based inventory optimization framework that integrates real-world dynamic inventory data [5]. By evaluating multiple DRL algorithms across single-echelon and two-echelon environments, this research provides empirical insights into algorithm suitability under different decision spaces and demand conditions. The study further extends existing literature by incorporating multiple performance metrics and demand fulfillment mechanisms, offering a more comprehensive assessment of DRL-based inventory control strategies.

3. DRL for Inventory Control in Supply Chains

This study adopts a two-phase experimental research framework to systematically investigate the applicability and performance of Deep Reinforcement Learning (DRL) in supply chain inventory optimization. The framework is designed to evaluate DRL

algorithms under increasing levels of system complexity and decision dimensionality, following recommendations from prior DRL inventory research roadmaps [2].

Phase I focuses on a single-SKU single-echelon inventory system to validate the fundamental learning behavior of DRL algorithms under different demand fulfillment mechanisms. Phase II extends the environment to a multi-SKU two-echelon supply chain, where replenishment and allocation decisions must be jointly optimized. This phased design enables controlled comparison of algorithm performance while progressively introducing realistic supply chain complexities.

3.1. Dataset Description

The dataset used in this study originates from the Kaggle platform's Dynamic Inventory Dataset (Kaizen Analytics, n.d.) [5]. The raw transaction records were aggregated into weekly time-series data to standardize the simulation horizon. The entire analysis relies solely on this dataset to ensure interpretability and reproducibility.

Prior to simulation, demand series are analyzed to assess variability and temporal dependence. Rather than imposing parametric demand assumptions, the study directly utilizes empirical demand patterns to preserve non-stationary characteristics observed in real supply chains. This data-driven design addresses limitations noted in previous studies that rely primarily on synthetic or stationary demand processes [6].

3.2. Phase I: Single-SKU Single-Echelon Inventory Environment

Phase I models a retailer replenishing inventory from an upstream supplier subject to deterministic or stochastic lead times. At each decision epoch, the DRL agent observes the system state, which includes current on-hand inventory, outstanding replenishment orders, and recent demand history. The agent then selects an order quantity from a discretized action space, enabling the application of value-based DRL algorithms.

Three demand fulfillment mechanisms are examined to capture different operational settings:

(1) **Complete Backordering**: All unmet demand is backlogged and fulfilled when inventory becomes available.
(2) **Lost Sales**: Unmet demand is permanently lost and incurs a penalty cost.
(3) **Partial Backordering**: A fixed proportion of unmet demand is backlogged, while the remainder is treated as lost sales.

The immediate reward is defined as the negative total cost incurred in each period, including holding costs and shortage-related penalties. This formulation aligns with standard cost-based evaluation approaches in inventory control while allowing DRL agents to learn cost-minimizing policies through interaction. In this phase, Deep Q-Network (DQN), Double DQN (DDQN), and Proximal Policy Optimization (PPO) are implemented and evaluated [3,4].

3.3. Phase II: Multi-SKU Two-Echelon Supply Chain Environment

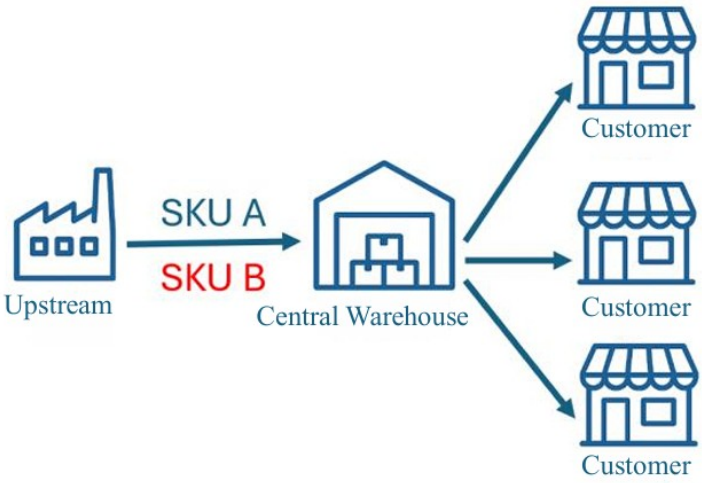

Figure 1. Structure of the multi-SKU two-echelon supply chain in Phase II, where a central warehouse coordinates replenishment from an upstream supplier and allocates inventory to multiple downstream customer nodes.

Phase II extends the experimental setting to a multi-SKU two-echelon supply chain consisting of a central warehouse and multiple downstream customer nodes. The warehouse places replenishment orders to an upstream supplier and allocates available inventory to customers. This structure reflects common warehouse–retailer systems studied in the inventory literature (Figure 1) [1,7].

The state space incorporates warehouse-level and customer-level information, including on-hand inventory, in-transit orders, and historical demand windows for each SKU. The action space consists of continuous-valued replenishment and allocation decisions across SKUs and nodes, resulting in a high-dimensional control problem. Due to the continuous and multi-dimensional nature of the action space, policy-gradient and actor–critic DRL algorithms—specifically PPO and Deep Deterministic Policy Gradient (DDPG)—are employed in this phase [8].

To avoid backlog propagation across multiple nodes, only the lost-sales demand fulfillment mechanism is considered in Phase II. This design choice follows prior multi-echelon DRL studies that emphasize stability and tractability in complex network settings [4].

3.4. Experimental Environment and Modeling

The state–action framework is constructed as follows:

- **State Space**: Current inventory level, on-order inventory, historic demand, and lead time.
- **Action Space**: Replenishment quantity (discrete or continuous decision variable).
- **Reward Function**: Negative total cost per period, composed of ordering cost, holding cost, and shortage penalty.

The single-echelon model simulates a retailer facing end-customer demand, while the two-echelon model introduces a manufacturer node to simulate upstream coordination and supply delays. All environments are implemented in Python and interfaced

through OpenAI Gym for modular simulation control [11].

3.5. DRL Algorithms and Training Configuration

Three algorithms—DQN, DDQN, and PPO—are implemented using the Tensor-Flow library. The key hyperparameters are summarized below:

- **Learning rate**: 0.001.
- **Discount factor (γ)**: 0.99.
- **Replay buffer size**: 10,000.
- **Mini-batch size**: 64.
- **Exploration rate interval**: [0.1, 1.0] linearly decayed.
- **Number of episodes**: 1,000.

The training process follows the episodic learning framework, with each episode representing one simulated year (52 decision periods). Model performance is monitored through cumulative reward convergence and cost minimization stability.

3.6. Demand Forecasting Auxiliary Module

To enhance decision-making accuracy, an auxiliary Long Short-Term Memory (LSTM) forecasting module is developed to predict future demand for the next three periods. The predicted values are integrated into the DRL agent's state vector to create a hybrid LSTM-DRL model, which helps the agent anticipate near-term demand fluctuations. This approach improves inventory responsiveness and reduces shortage risk in volatile environments.

The LSTM model is pretrained on a related SKU and generates short-horizon demand predictions. The forecasting model is trained separately from the DRL agent to prevent data leakage and ensure realistic decision timing. Similar hybrid forecasting–control architectures have been adopted in recent DRL-based supply chain studies to improve robustness under demand uncertainty [6].

3.7. Performance Evaluation Metrics

Learned inventory policies are evaluated using multiple performance metrics to provide a comprehensive assessment of operational effectiveness. These metrics include:

- **Total Cost**: The cumulative sum of holding costs and shortage or lost-sales penalties over the evaluation horizon.
- **Service Level**: The proportion of customer demand satisfied without shortage.
- **Inventory Efficiency**: Measured through average inventory levels and inventory turnover rates.
- **Learning Stability and Convergence**: Assessed using reward trajectories, variance across training runs, and convergence speed.

Traditional inventory policies, including *(s, S)*, *(s, Q)*, Economic Order Quantity (EOQ), and periodic-review *(R, S)* policies, are implemented as benchmark strategies to contextualize the performance of DRL-based approaches. This benchmarking design follows evaluation practices commonly adopted in the inventory control literature [2].

3.8. Summary of Experimental Flow

The overall research process follows a progressive two-stage approach:

- Start with a simple single-SKU, single-echelon environment to validate algorithmic feasibility.
- Expand to multi-SKU, multi-echelon environments to test model generalizability and scalability.

This sequential design ensures methodological rigor, enabling a structured evaluation of DRL algorithms under varying levels of supply-chain complexity.

4. Results and Discussion

4.1. Phase I: Single-SKU Single-Echelon Environment

This section presents the experimental results obtained from Phase I, which evaluates the performance of Deep Reinforcement Learning (DRL) algorithms in a single-SKU single-echelon inventory environment. The objective of this phase is to examine fundamental learning behavior, convergence characteristics, and policy performance under different demand fulfillment mechanisms before introducing additional system complexity.

4.1.1. Learning Convergence and Stability

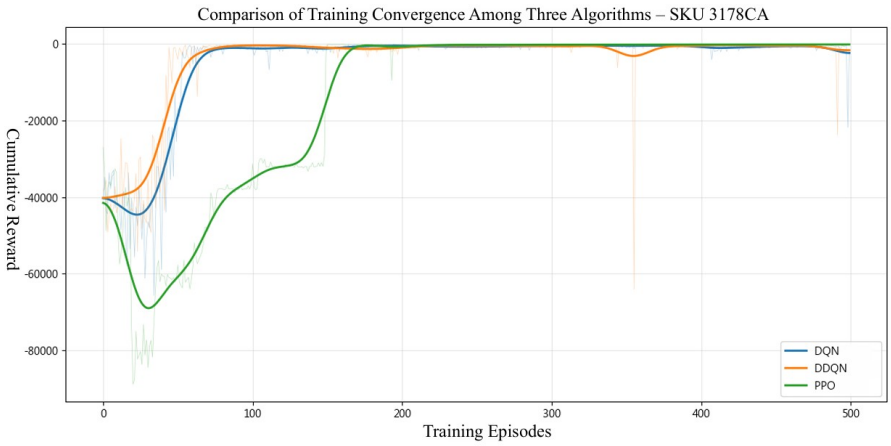

Figure 2. Comparison of training convergence among DQN, DDQN, and PPO in terms of cumulative reward (SKU 3178CA).

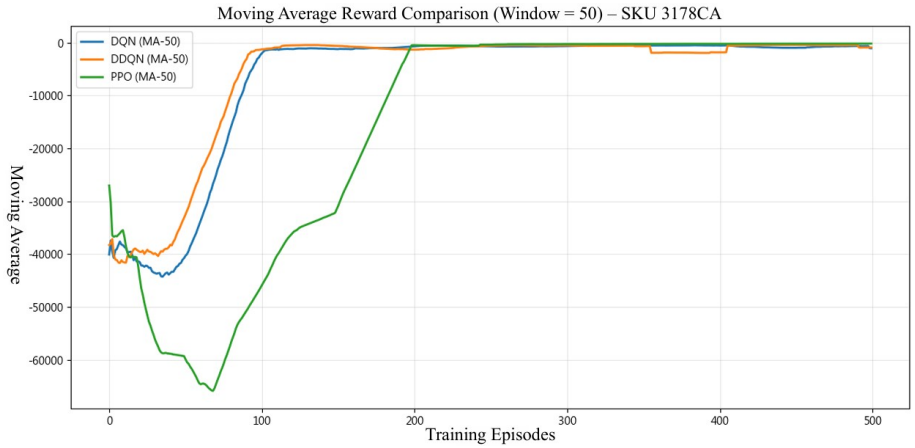

Figure 3. 50-episode moving average reward comparison during training (SKU 3178CA).

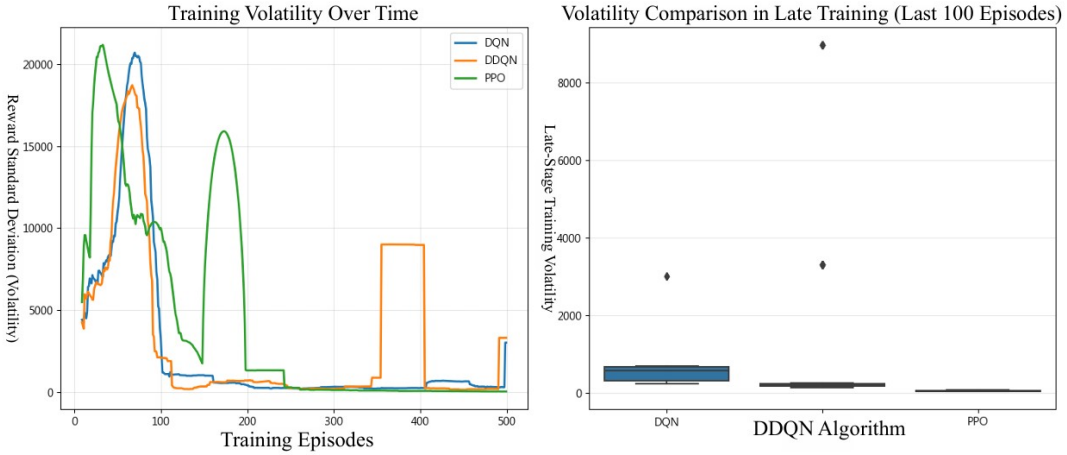

Figure 4. Training volatility over time and late-stage volatility comparison among DRL algorithms (SKU 3178CA).

Three DRL algorithms—DQN, DDQN, and PPO—were applied under different demand processing modes and lead-time settings. Figures 2 and 3 illustrate the cumulative reward trajectories. The DDQN model achieved the fastest convergence within approximately 100 episodes, exhibiting stable reward growth and minimal oscillation. DQN converged moderately (around 180 episodes), while PPO required longer training (over 300 episodes) due to its continuous policy update mechanism.

Although DQN is able to learn a feasible replenishment policy, its Q-value estimates show higher variance, leading to oscillation in the early training stage (Figure 4). This behavior is consistent with prior findings that policy-gradient methods typically require more samples due to stochastic policy updates but can offer improved robustness under certain cost structures [2,3]. Overall, these results confirm that value-based DRL methods are well suited for discrete inventory replenishment decisions with moderate state dimensionality.

In the left panel of Figure 4, volatility is measured by the variance of rewards across episodes. The volatility curve of PPO rapidly approaches zero after episode 200. Although DDQN exhibits lower volatility than DQN, its volatility remains higher than that of PPO. The right panel of Figure 4 further compares the reward distributions during the final stage of training (the last 100 episodes). PPO exhibits the narrowest interquartile range and the fewest extreme values, indicating the highest stability among the three algorithms. DDQN ranks second, whereas DQN exhibits more outliers and a wider distribution, suggesting that its converged policy is more susceptible to perturbations from occasional samples.

4.1.2. Inventory Performance Analysis and Effects of Demand Fulfillment Modes

To further investigate the practical performance of different algorithms in single-SKU replenishment decision-making, this section compares their performance across several key inventory management performance indicators (KPIs), including service level, demand fulfillment rate, average inventory level, inventory turnover rate, ordering frequency, and excess inventory rate. Figure 5 presents a radar chart of these indicators.

The results reveal distinct inventory management priorities among the algorithms: DDQN emphasizes service quality, PPO prioritizes inventory efficiency, and DQN achieves a balance between cost and service performance. These findings suggest that algorithm selection should depend on decision-makers' priorities regarding service levels and inventory costs. Figure 6 provides bar chart comparisons of service performance and inventory efficiency. For applications where minimizing stockout risk is a primary concern, DDQN may be the preferred choice. In contrast, PPO offers advantages when inventory turnover and capital efficiency are the main considerations. These observations highlight that the choice of DRL algorithm should be aligned with operational risk preferences.

The impact of demand fulfillment mechanisms on inventory performance is furthermore examined using three settings: complete backordering, lost sales, and partial backordering. Figures 7 and 8 summarize the trade-offs among cost efficiency, service level, and average inventory. Under the **complete backordering** mechanism, all algorithms achieve high service levels in the long run. However, DDQN consistently achieves lower total costs by maintaining balanced inventory levels and avoiding excessive backlog accumulation. In contrast, DQN tends to place more frequent orders with smaller quantities, leading to higher holding costs.

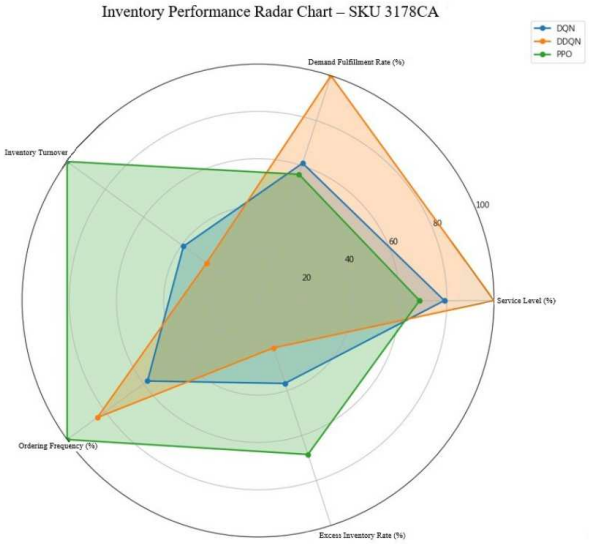

Figure 5. Radar chart comparing inventory performance indicators across DRL algorithms (SKU 3178CA).

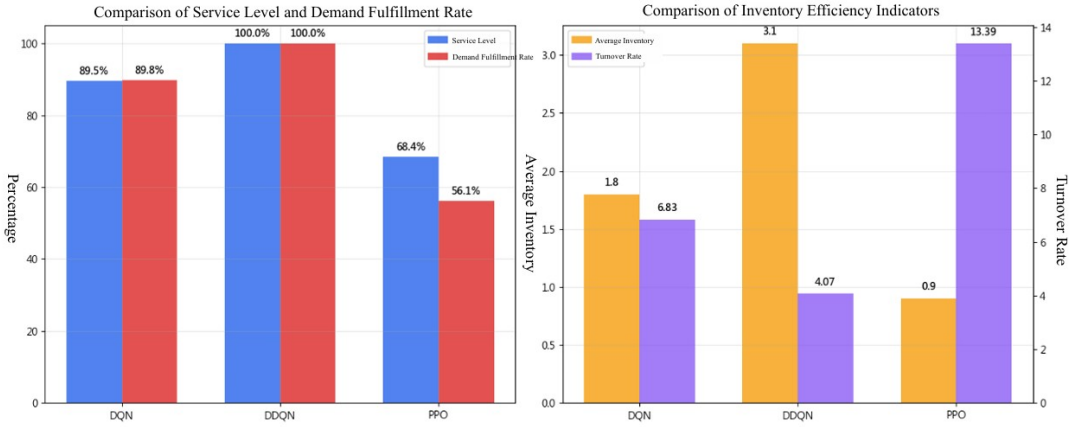

Figure 6. Comparison of service level, demand fulfillment rate, and inventory efficiency indicators.

In the **lost sales** setting, shortage penalties are significantly amplified, and conservative replenishment behavior becomes more advantageous. PPO outperforms DQN and DDQN by proactively increasing order quantities in anticipation of demand fluctuations, resulting in higher service levels and lower penalty costs. This behavior reflects PPO's ability to incorporate longer-term cost considerations through smoother policy updates. The **partial backordering** mechanism provides an intermediate setting in which both DDQN and PPO perform well. Allowing limited backlog flexibility stabilizes learning and reduces sensitivity to demand spikes, leading to improved cost–service trade-offs.

4.1.3. Ordering Behavior and Risk Analysis

The upper-left and upper-right panels of Figure 8 show that under the Complete Backorder mode, the ordering behaviors of all algorithms are relatively stable, with low variance. In contrast, variability increases substantially under the Lost Sales mode, particularly for DQN and DDQN, indicating that their ordering policies become more volatile when faced with demand that cannot be recovered. Under the Lost Sales mode, PPO places no orders at all, with both its average order quantity and variance equal to zero. This suggests that its policy tends to forgo satisfying some demand in exchange for greater cost stability.

The lower-left and lower-right panels of Figure 8 further examine the risk levels associated with each policy. Under the Complete Backorder mode, although PPO maintains a high average inventory level, it exhibits the largest coefficient of variation, indicating that its replenishment policy is highly unstable. Regarding the risk of stockouts and delayed deliveries, PPO achieves a stockout rate as high as 42.1% under the Lost Sales mode, suggesting that it sacrifices demand fulfillment to reduce holding costs when demand cannot be backordered. DDQN effectively controls the stockout rate under both modes, demonstrating the most robust performance.

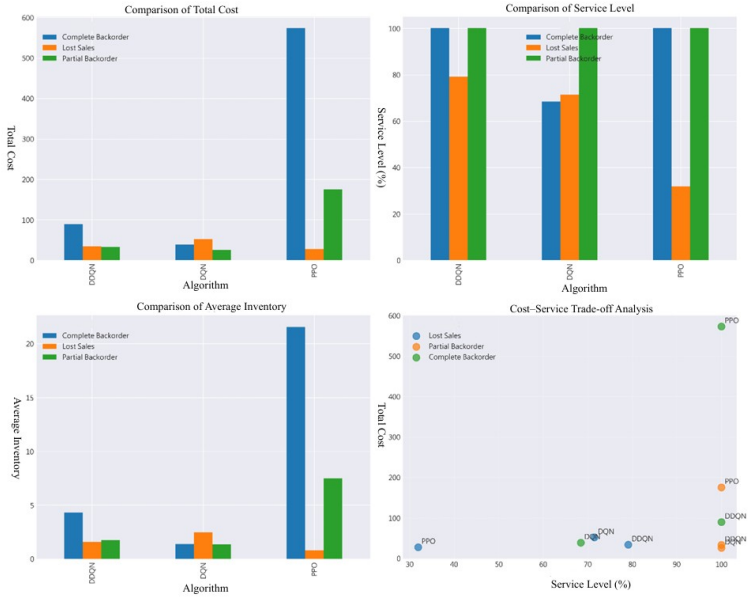

Figure 7. Comparison of total cost, service level, average inventory, and cost-service trade-offs under different demand fulfillment mechanisms.

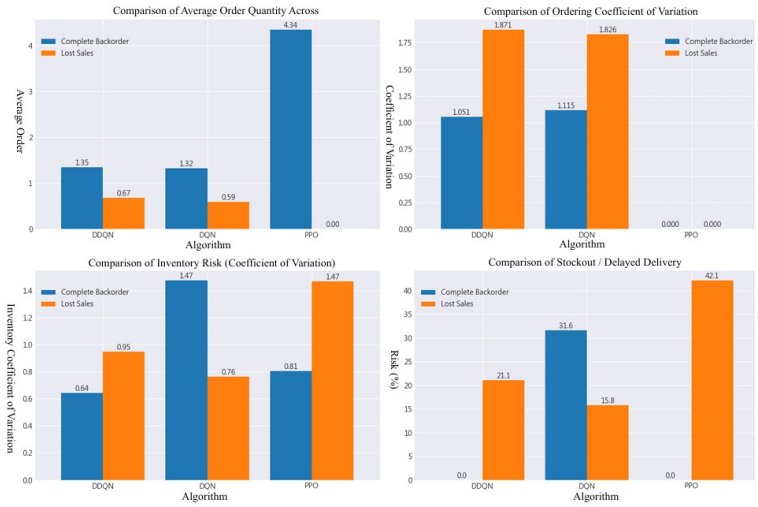

Figure 8. Comparison of ordering behavior, inventory risk, and stockout characteristics across algorithms under different demand fulfillment mechanisms.

4.1.4. Impact of Lead-Time Settings

This section investigates the effects of different lead time settings (fixed lead time and dynamic lead time) on policy performance. The analysis is conducted based on cumulative rewards, inventory levels and ordering behaviors. The dynamic lead time setting is designed primarily to account for the uncertainty in transportation time encountered in real-world operations. By using random sampling to represent fluctuations in actual supply chain transportation processes, this setting enables a more realistic simulation of real-world scenarios.

As shown in Figure 9, the training convergence curves and 50-episode moving average reward curves indicate that the learning curve under the fixed lead time setting exhibits considerable fluctuations during the initial stage but converges to a stable state after approximately 150 episodes. In contrast, the policy under the dynamic lead time setting converges to a stable state more rapidly. Although the dynamic lead time setting initially yields more negative rewards, indicating higher costs, its faster convergence suggests that the policy has a stronger adaptive capability and can quickly adjust to unstable transportation times in practical scenarios.

Figure 10 further compares the differences between the two lead time settings in terms of inventory levels, order quantities, stockout quantities, and cumulative costs. The inventory level comparison shows that the dynamic lead time policy maintains lower inventory levels for most of the time, reflecting its preference for lower inventory levels to reduce holding costs. In contrast, inventory levels under the fixed lead time setting fluctuate more substantially, suggesting that the policy maintains a larger inventory buffer when lead times are constant. Regarding order quantities, the fixed lead time policy consistently places relatively small and stable replenishment orders in each period, whereas the dynamic lead time policy places substantially smaller orders, with quantities approaching zero, indicating a more conservative replenishment strategy.

The stockout quantity comparison shows that stockouts occur more frequently and severely under the dynamic lead time policy. In particular, at time steps with substantial demand fluctuations, such as time steps 24 and 30, stockout quantities are significantly higher than those under the fixed lead time policy. This indicates that

while the dynamic lead time policy reduces inventory costs, it also entails a higher risk of stockouts. The cumulative cost comparison, however, shows that despite the more severe stockout conditions under the dynamic lead time policy, the overall cumulative costs of the two policies remain very similar. This suggests that the savings in inventory holding costs achieved by the dynamic lead time policy effectively offset the increase in stockout costs.

The KPIs presented in Figure 11 further validate these observations. The dynamic lead time policy yields a total stockout rate and total stockout quantity of 42.1 and 8.5, respectively, both substantially higher than the corresponding values of 31.6 and 5.5 under the fixed lead time policy, highlighting the greater stockout risk associated with dynamic lead times. However, the two policies exhibit very similar performance in terms of average inventory, inventory standard deviation, and average cost. These results indicate that although the dynamic lead time policy performs worse on stockout-related metrics, its overall inventory management costs remain competitive.

In summary, different lead time settings have a significant impact on reinforcement learning policy performance. The fixed lead time policy performs better in terms of service levels and inventory security, whereas the dynamic lead time policy offers clear advantages in inventory cost control by rapidly adapting to transportation uncertainty in real-world scenarios, albeit at the expense of a higher stockout risk. In practice, enterprises should weigh these trade-offs according to their operational requirements and cost structures to determine the most suitable lead time management strategy.

4.1.5. Comparison with Traditional Policies

Compared with the baseline *(s, S)* policy, all DRL algorithms achieved lower total costs (6–12% reduction) and higher service levels in dynamic conditions. The improvement was most notable under the Complete Backordering mode, confirming DRL's ability to optimize trade-offs dynamically without explicit modeling assumptions.

4.1.6. Discussion

Table 2 summarizes the performance outcomes. DDQN attained the highest service level (94.74%) but maintained a relatively high average inventory. PPO achieved the best inventory turnover efficiency (8.2 turns per period) but with a lower service level (73.68%). DQN balanced both objectives, reaching a service level of 88.4% with a moderate cost level.

Table 1. Performance Summary under Single-SKU Setting

Algorithm	Service Level (%)	Inventory Turnover	Convergence (Episodes)
DQN	88.4	6.9	180
DDQN	94.7	5.4	100
PPO	73.7	8.2	320

The Phase I results yield several important insights. First, DRL algorithms can autonomously learn effective replenishment strategies without relying on explicit demand models. Second, algorithm selection plays a critical role in determining learning stability and operational performance. Third, demand fulfillment mechanisms substantially influence policy behavior and cost structure, underscoring the importance of aligning

DRL design with practical inventory settings.

The three demand mechanisms-Complete Backordering, Lost Sales, and Partial Backordering-showed substantial performance differences. Complete Backordering resulted in the lowest total cost due to demand recoverability, while Lost Sales created severe penalties that led to policy instability, particularly for PPO. Under Partial Backordering, policies maintained intermediate stability and adaptability.

Overall, Phase I establishes a strong foundation for extending DRL-based inventory control to more complex supply chain environments, motivating the multi-SKU two-echelon analysis presented in Phase II.

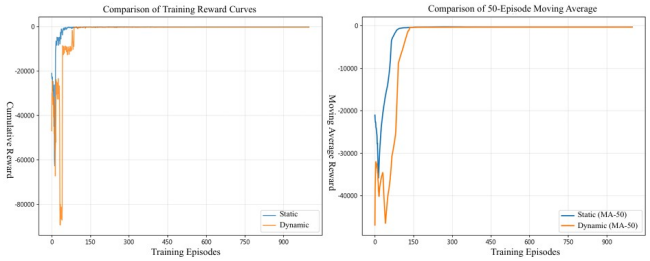

Figure 9. Comparison of training reward convergence and 50-episode moving average between static and dynamic lead time.

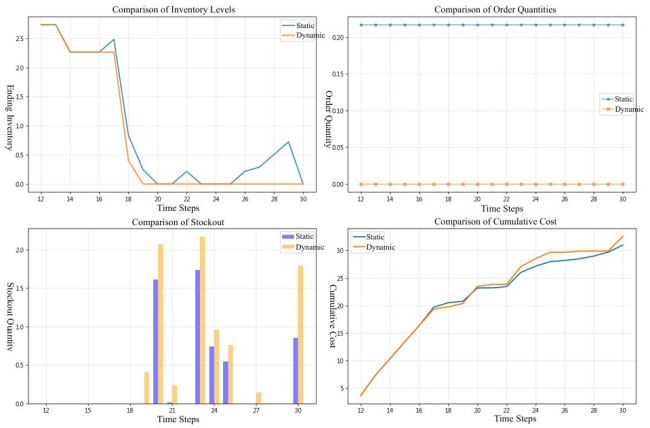

Figure 10. Comparison of inventory levels, order quantities, stockouts, and cumulative cost under static and dynamic lead time.

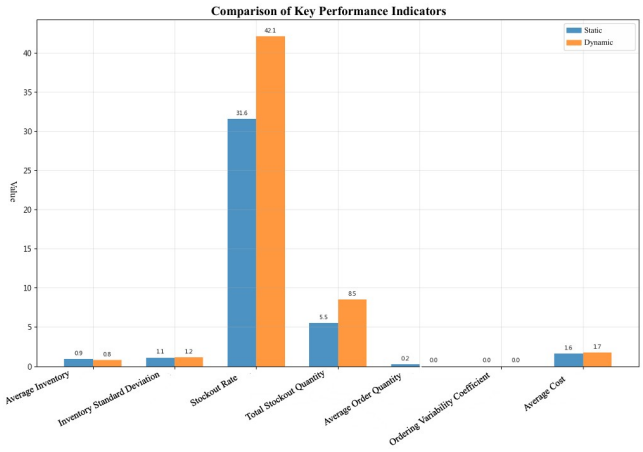

Figure 11. Comparison of key performance indicators between static and dynamic lead time.

4.2. Phase II: Multi-SKU Two-Echelon Environment

Phase II extends the analysis to a multi-SKU two-echelon supply chain environment, where a central warehouse coordinates replenishment from an upstream supplier and allocates inventory to multiple downstream customer nodes. This setting introduces a action space and requires joint optimization of replenishment and allocation decisions, representing a more complex decision problem than Phase I. Figure 12 demonstrates several indicators during training process. Figure 13 complements the learning stability and performance between PPO and DDPG.

4.2.1. Learning Behavior in Continuous and Multi-Dimensional Action Spaces

This section analyzes the convergence performance and training dynamics of PPO and DDPG in a two-echelon, multi-SKU supply chain environment with a high-dimensional continuous action space. Figure 12 presents the evolution of six key metrics throughout training, while Figure 13 further examines long-term training stability and the distribution of final performance.

(1) Trends in Cumulative Reward and Average Service Level

As shown in the upper-left and upper-middle panels of Figure 12, both PPO and DDPG gradually converge after approximately the first 600 episodes. PPO exhibits a greater increase in average cumulative reward and ultimately achieves slightly better performance than DDPG. PPO reaches the 95% average service level threshold at approximately episode 548, earlier than DDPG, which reaches it at episode 646. This indicates that PPO offers greater learning efficiency and stability during the early stages of training.

(2) Stable Decline in Policy and Value Losses

The upper-right and lower-left panels of Figure 12 show decreasing trends in both policy loss (policy loss/actor loss) and value loss (value loss/critic loss) for both algorithms. However, DDPG exhibits greater fluctuations in actor loss, suggesting slightly lower stability in policy learning. Overall, the loss curves of both algorithms exhibit declining trends consistent with the expected training dynamics of reinforcement learning models.

(3) Trends in Exploration Behavior

The lower-middle panel of Figure 12 shows that the entropy of PPO gradually

decreases, indicating that its policy becomes increasingly deterministic as training progresses. Similarly, the exploration level of DDPG, as reflected by its noise scale, decreases over time. However, the noise scale declines too rapidly during the early stages, potentially resulting in insufficient exploration and an increased risk of premature convergence, which may adversely affect final policy performance.

(4) Comparison of Convergence Rates and Time to Reach Target Performance

The lower-right panel of Figure 12 presents the convergence rate evaluation, measuring the proportions of training episodes that achieve service levels of 90%, 95%, and 100%. The results indicate that both algorithms ultimately reach the 95% threshold, while PPO demonstrates better overall convergence speed and greater stability in reaching the target performance levels. Its advantage is particularly evident in the higher proportion of episodes achieving the 100% service level threshold.

(5) Comparison of Stability and Sample Efficiency

The upper-middle panel of Figure 13 reveals that the coefficient of variation of PPO's rewards gradually decreases throughout training, indicating greater long-term performance stability than DDPG. The comparison of learning speed in the upper-right panel shows no substantial difference during the early stages; however, PPO achieves more stable convergence in the later stages. The sample efficiency comparison in the lower-left panel further highlights that PPO obtains a higher average reward per 1,000 interactions, indicating more effective utilization of training data.

(6) Evaluation of Final Performance and Overall Effectiveness

The lower-middle panel of Figure 13, which presents the performance distribution over the final 200 episodes, shows that PPO achieves a more concentrated distribution and higher average performance than DDPG. Furthermore, the radar chart in the lower-right panel indicates that PPO slightly outperforms DDPG across all five evaluation metrics: stability, final performance, sample efficiency, exploration capability, and convergence speed.

In summary, PPO demonstrates superior overall convergence stability and learning efficiency in continuous action spaces, particularly in terms of the speed at which the target service level is reached and the efficiency of sample utilization. In comparison, DDPG exhibits relatively stable performance during the initial convergence stage but is more susceptible to policy fluctuations caused by noise decay. Consequently, its overall convergence performance is slightly inferior to that of PPO. The suitability of DDPG for a given application therefore depends on achieving an appropriate balance between its exploration mechanism and sample efficiency and stability.

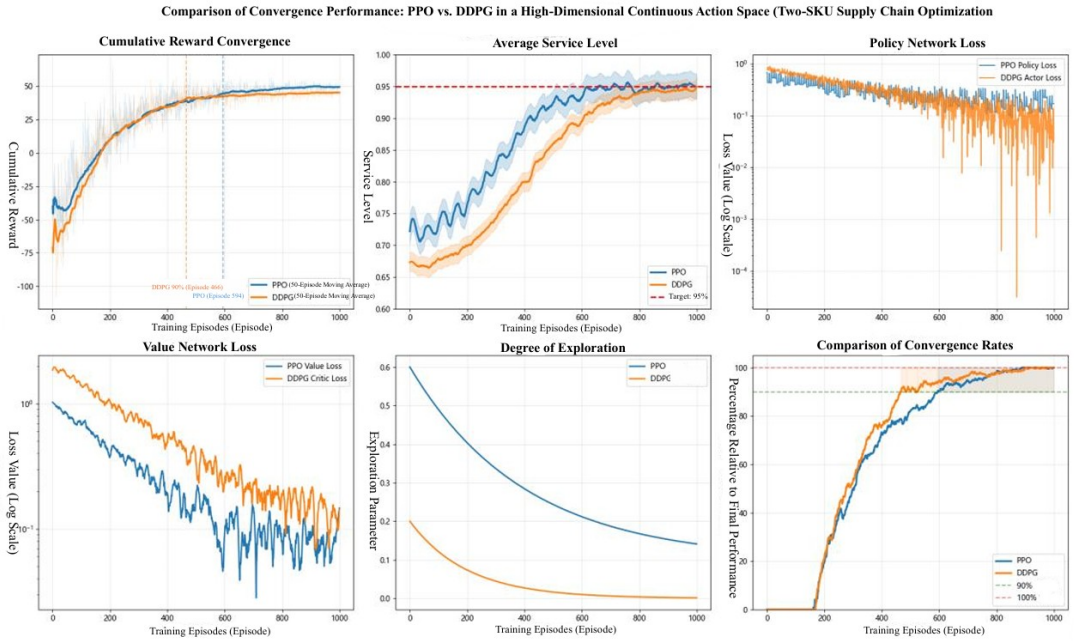

Figure 12. Comparison of convergence behavior, service level evolution, loss dynamics, exploration decay, and convergence rates of PPO and DDPG in a high-dimensional continuous action space.

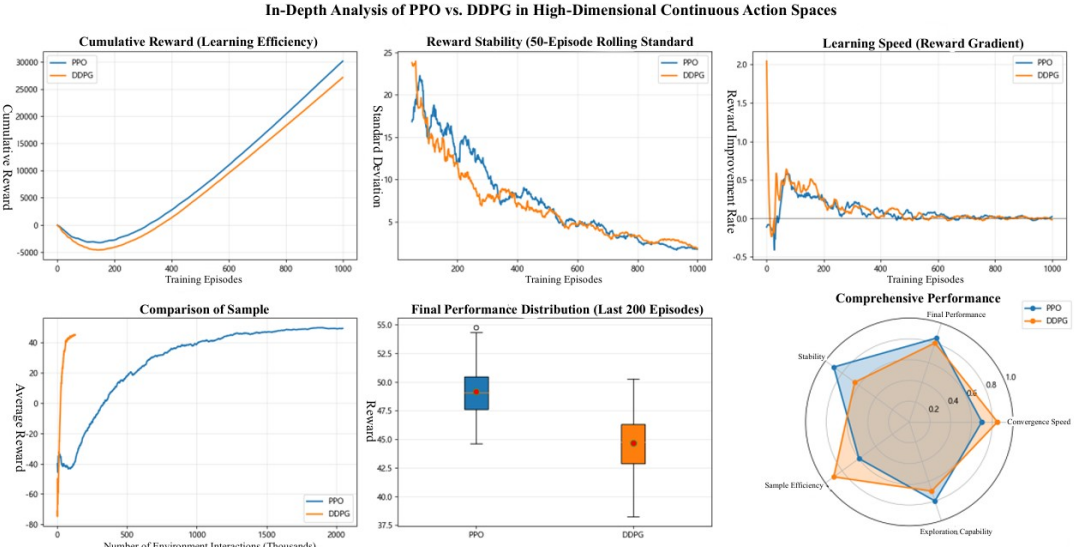

Figure 13. Comparison of learning efficiency, reward stability, sample efficiency, and final performance distribution between PPO and DDPG.

4.2.2. Service Level and Inventory Efficiency across Nodes

Inventory turnover analysis indicates that PPO maintains leaner inventory levels while preserving service performance, whereas DDPG tends to accumulate excess safety stock during early learning stages. Although DDPG gradually improves allocation efficiency with extended training, its higher fluctuation may pose challenges for real-world deployment, where operational stability is often prioritized.

Figure 14 illustrates the evolution of overall service level across training episodes. Both PPO and DDPG exhibit a steady upward trend and ultimately reach the target threshold of 95%. PPO converges slightly faster and achieves marginally better service performance than DDPG. Overall, both algorithms ultimately outperform the acceptable benchmark of 90%, demonstrating their ability to learn effective service-level policies.

Figure 15 further presents the final service levels at each node, defined as a combination of SKU and channel type, averaged over the last 100 periods. The results reveal that:

(1) PPO achieves or approaches the 95% target across all six nodes, whereas DDPG falls slightly below the target at three nodes, with service levels below 93% at nodes such as 2391-Expo and 2391-Whol.

(2) At nodes with higher demand or greater turnover, such as 2418-Dist, both PPO and DDPG perform well, indicating that both algorithms can adapt effectively to high-demand SKUs.

To examine the stability and variability of service levels, Figure 16 presents box plots of service-level distributions for each node over the final 200 periods. PPO exhibits narrower interquartile ranges and higher medians across most nodes, indicating greater service-level stability and lower variability. In contrast, DDPG displays skewed distributions at certain nodes, with narrower ranges between the upper and lower bounds, suggesting somewhat inferior stability during the learning process.

The performance dashboard in Figure 17 further shows that PPO slightly outperforms DDPG across three service-related indicators: overall service level, service stability, and churn control. PPO also achieves a marginally higher final performance score, with 95 points compared with 94 points for DDPG. However, the two algorithms perform comparably in terms of convergence speed, with only a minor difference between them.

In summary, PPO slightly outperforms DDPG in terms of average service level while also demonstrating greater stability and consistency. Furthermore, PPO is more effective at controlling churn risk and delivers better service performance in a complex, multi-node environment.

4.2.3. Discussion

The Phase II results highlight several important implications for DRL-based inventory control in multi-echelon supply chains. First, continuous-action DRL methods are necessary for joint replenishment and allocation problems involving multiple SKUs and nodes. Second, PPO provides superior stability and coordination performance, making it more suitable for practical deployment in complex supply chain environments. Third, while DDPG can achieve competitive long-term performance, it requires careful tuning and extended training to mitigate instability.

Overall, Phase II demonstrates that DRL-based inventory policies can scale beyond simplified single-echelon settings and effectively address realistic multi-echelon supply chain decision problems. These findings complement the Phase I analysis and further

support the applicability of DRL for adaptive inventory management in modern supply chains.

4.3. Cross-Phase Comparative Discussion

Table 2 compares key insights across both experimental phases. The results verify that algorithm performance depends strongly on environmental complexity and operational goals:

(1) DDQN excels when service reliability is prioritized.

(2) PPO is ideal for efficiency-oriented objectives under dynamic, continuous control.

(3) DQN offers a balanced approach for hybrid scenarios with moderate variability.

Table 2. Summary of Algorithmic Performance Across Phases

Algorithm	Best Scenario	Key Advantage	Limitation
DQN	Mixed demand	Balanced performance	Moderate convergence
DDQN	Stable demand	High service level	Larger inventory
PPO	Stochastic demand	Fast learning, efficient use of data	Lower service rate in instability
DDPG	Multi-echelon, continuous	Coordination capability	Slower convergence

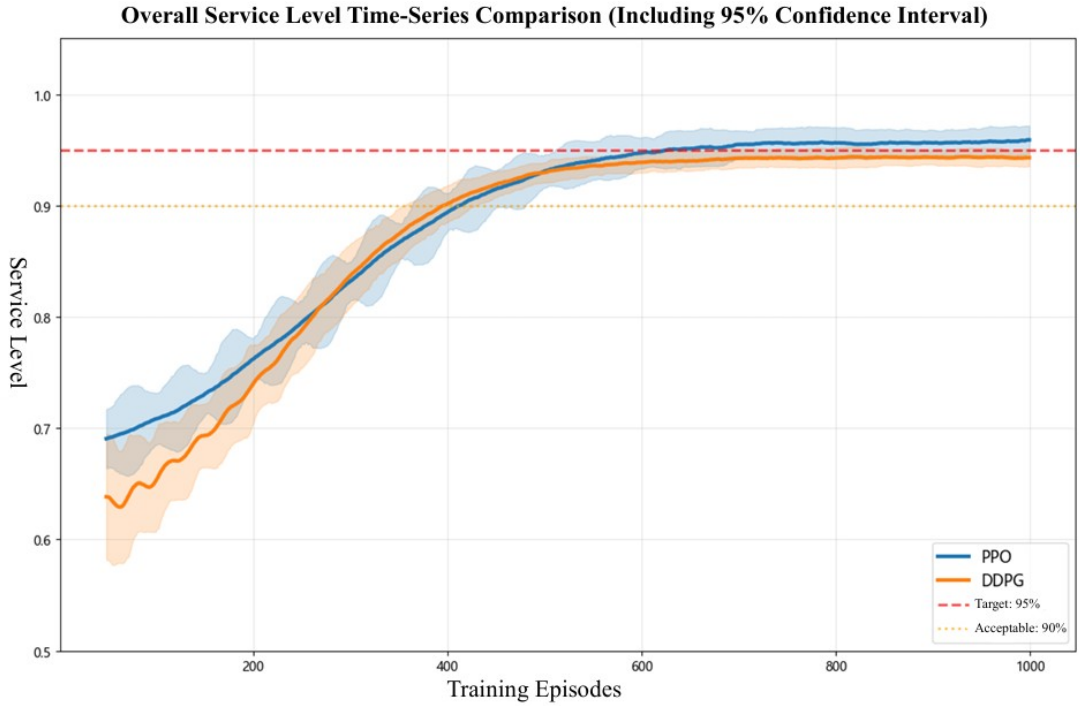

Figure 14. Overall service level during training with 95% confidence intervals under PPO and DDPG in the multi-SKU two-echelon environment.

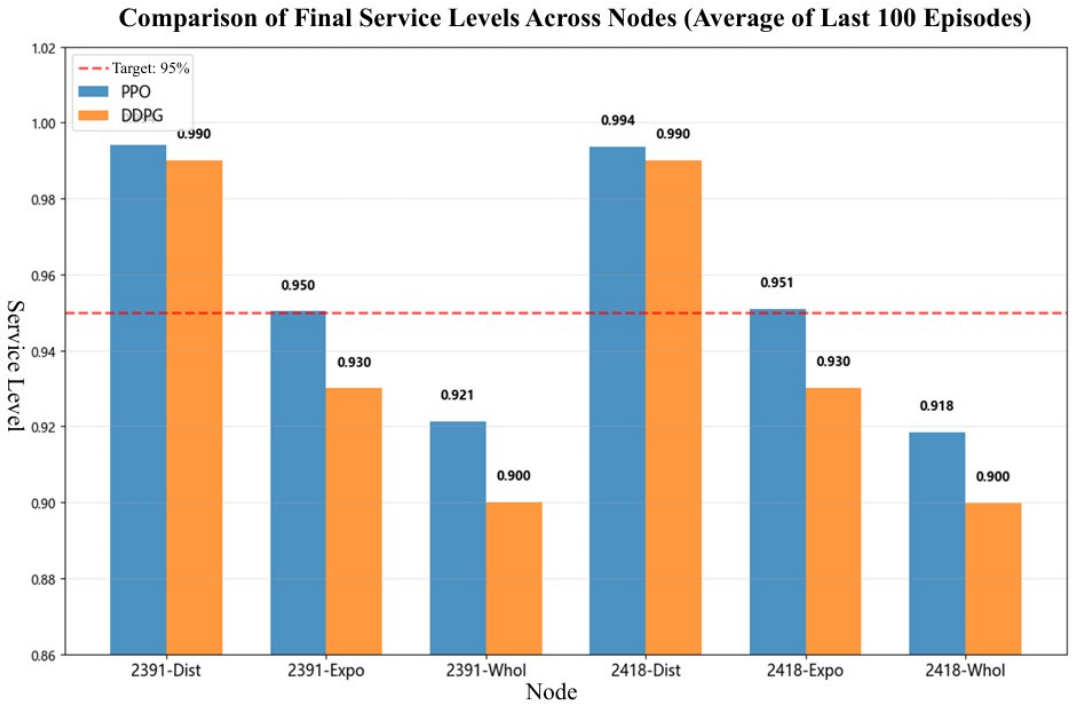

Figure 15. Comparison of final service levels across downstream nodes, averaged over the last 100 training episodes.

5. Key Findings and Implications

Overall, this study demonstrates that DRL models are capable of achieving effective inventory control in dynamic supply chain environments. The experimental findings highlight three major insights:

(1) Algorithm choice should align with operational priorities. DDQN is preferable for service-critical settings, while PPO provides higher efficiency in resource utilization.

(2) Environmental complexity directly impacts learning stability. Multi-echelon coordination introduces delay propagation effects, requiring algorithms with strong temporal credit assignment (e.g., actor-critic structures).

(3) Hybrid models combining DRL with forecasting (e.g., LSTM-DRL) improve responsiveness to demand fluctuations and reduce shortage risks.

These findings provide a practical foundation for applying DRL to intelligent inventory management and supply chain digital transformation.

6. Conclusion and Future Work

6.1. Conclusions

This research investigated the application of DRL in supply chain inventory management through a progressive two-phase experimental framework. The findings from the first phase, which focused on a single-SKU and single-echelon environment, demonstrated that DRL algorithms can effectively learn replenishment strategies that outperform traditional rule-based methods. Among the algorithms tested, the DDQN

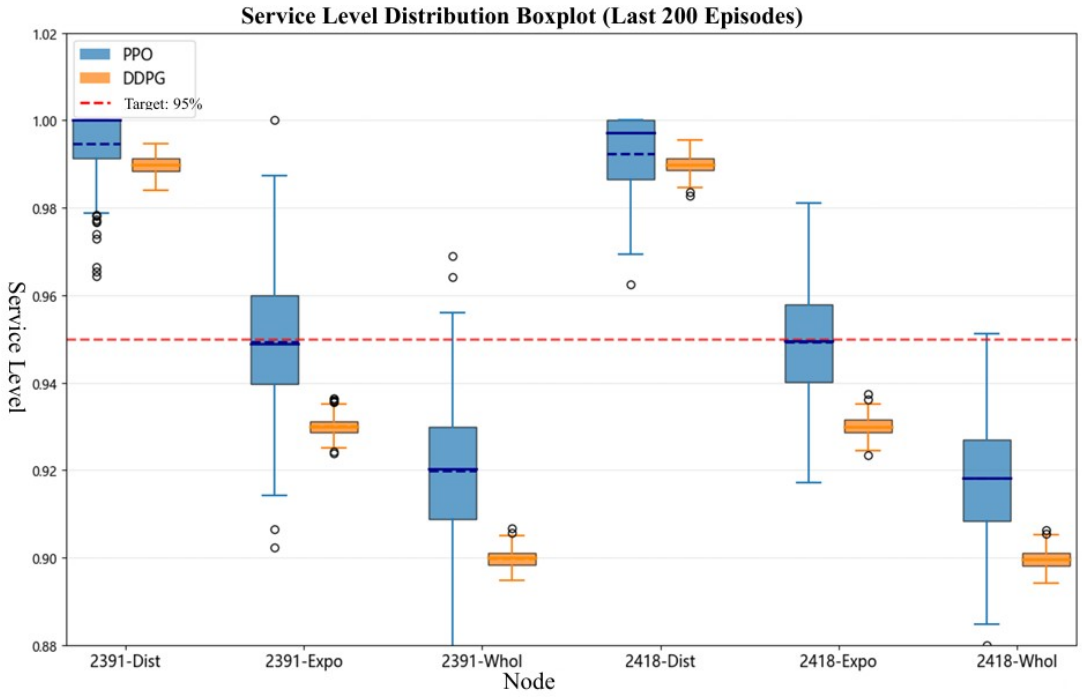

Figure 16. Overall service level boxplot under PPO and DDPG in the multi-SKU two-echelon environment.

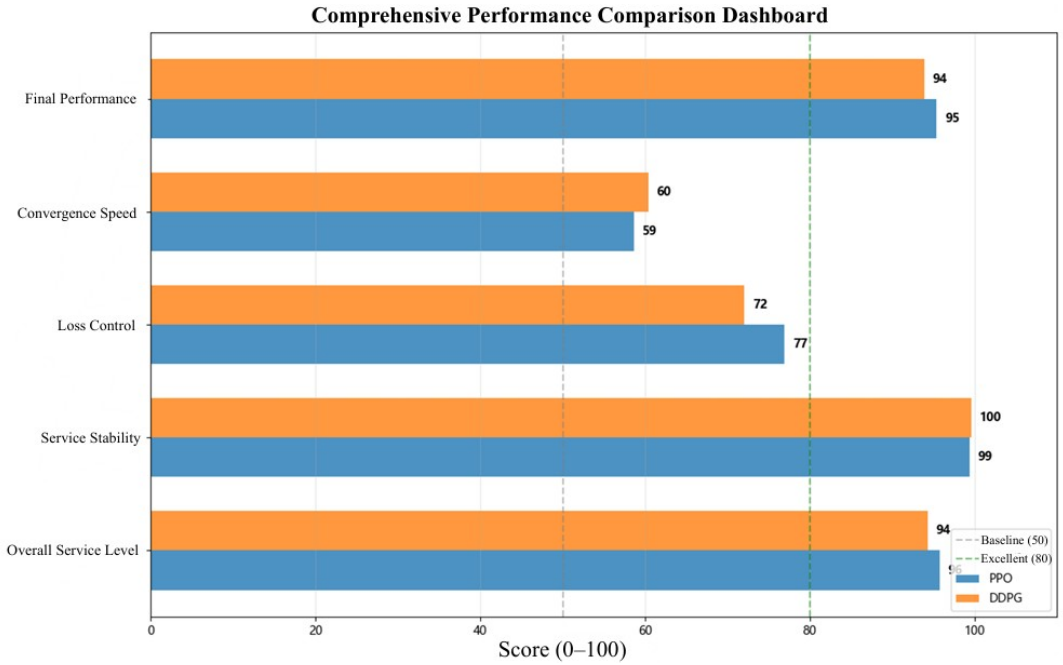

Figure 17. Performance dashboard for PPO and DDPG in the multi-SKU two-echelon environment, averaged over the last 100 training episodes.

achieved the most stable convergence and the highest service level, while PPO provided the best inventory turnover efficiency. The second phase extended the study to a multi-SKU and two-echelon supply chain environment, confirming the scalability and adaptability of DRL approaches under higher complexity. The PPO and DDPG algorithms successfully coordinated replenishment between upstream and downstream nodes, showing improved responsiveness to demand fluctuations. These results collectively suggest that DRL can serve as a promising foundation for intelligent, data-driven decision systems in dynamic supply chain operations.

6.2. Research Contributions

This study provides several key contributions to the literature and practical implementation of supply chain analytics:

- It demonstrates the feasibility of integrating DRL into inventory management and empirically verifies its advantages over conventional models such as EOQ and (s, S) policies.
- It establishes a two-phase simulation framework that enables systematic evaluation of algorithmic performance under varying complexity levels.
- It introduces a forecasting-assisted hybrid model (LSTM–DRL) that enhances predictive capability and mitigates stockout risks.
- It contributes to bridging the gap between theoretical reinforcement learning models and their operational deployment in real-world logistics environments.
- The proposed simulation framework is generalizable and can be extended to other supply chain decision-making problems, such as transportation planning and multi-agent coordination.

6.3. Future Research Directions

Building upon the findings of this study, several future research extensions are suggested:

(1) Multi-agent coordination: Future studies may extend DRL frameworks to multi-agent environments where different supply chain entities cooperate or compete for shared resources.

(2) Integration with real-time data: Incorporating IoT and sensor data can improve decision timeliness and enable adaptive learning in dynamic environments.

(3) Multi-objective optimization: Expanding DRL reward functions to consider sustainability, carbon footprint, or service equity among suppliers could align decision-making with ESG goals.

(4) Explainable DRL models: Future work may explore interpretability frameworks that enhance managerial trust in AI-driven decisions, such as policy visualization or sensitivity analysis.

In conclusion, this study verifies that Deep Reinforcement Learning holds strong potential to transform supply chain inventory management from a rule-based process into a self-optimizing, intelligent system. With further development and integration into real-time decision platforms, DRL-based methods are expected to play an essential role in the digital transformation of modern supply chains.

Disclosure of interest

The authors report there are no competing interests to declare.

Acknowledgements

This work was supported by the National Science and Technology Council (NSTC), Taiwan, under Grant No. NSC115-2221-E-141-010.

Data Availability Statement

The data supporting the findings of this study are publicly available from the Kaggle repository, Dynamic Inventory Dataset – Kaizen Analytics, at https://www.kaggle.com/datasets/andrewniko/dynamic-inventory-dataset-kaizen-analytics/data. The dataset was used as the source of data for the analyses presented in this paper.

References

[1] Clark AJ, Scarf H. Optimal policies for a multi-echelon inventory problem. Management Science. 1960;6(4):475–490.

[2] Boute RN, Gijsbrechts J, Van Jaarsveld W, et al. Deep reinforcement learning for inventory control: A roadmap. European Journal of Operational Research. 2022;298(2):401–412.

[3] Oroojlooyjadid A, Nazari M, Snyder LV, et al. A deep q-network for the beer game: Deep reinforcement learning for inventory optimization. Manufacturing & Service Operations Management. 2022;24(1):285–304.

[4] Gijsbrechts J, Boute RN, Van Mieghem JA, et al. Can deep reinforcement learning improve inventory management? performance on lost sales, dual-sourcing, and multi-echelon problems. Manufacturing & Service Operations Management. 2022;24(3):1349–1368.

[5] Dynamic Inventory Dataset_Kaizen Analytics. Dynamic inventory dataset [https://www.kaggle.com/datasets/andrewniko/dynamic-inventory-dataset-kaizen-analytics/data]; n.d. [Dataset].

[6] Dehaybe H, Catanzaro D, Chevalier P. Deep reinforcement learning for inventory optimization with non-stationary uncertain demand. European Journal of Operational Research. 2024;314(2):433–445.

[7] Kaynov I, Van Knippenberg M, Menkovski V, et al. Deep reinforcement learning for one-warehouse multi-retailer inventory management. International Journal of Production Economics. 2024;267:109088.

[8] Stranieri F, Stella F. Comparing deep reinforcement learning algorithms in two-echelon supply chains. In: Meo R, Silvestri F, editors. Machine learning and principles and practice of knowledge discovery in databases. Vol. 2136. Springer Nature Switzerland; 2025. p. 454–469.

[9] Chong JW, Kim W, Hong J. Optimization of apparel supply chain using deep reinforcement learning. IEEE Access. 2022;10:100367–100375.

[10] Bermúdez JS, Del Rio Chanona A, Tsay C. Distributional constrained reinforcement learning for supply chain optimization. In: Computer aided chemical engineering. Vol. 52. Elsevier; 2023. p. 1649–1654.

[11] Brockman G, Cheung V, Pettersson L, et al. OpenAI Gym ; 2016. Available from: https://arxiv.org/abs/1606.01540.