MULTIDISCIPLINARY RESEARCH IN COMPUTING INFORMATION SYSTEMS

VOL 05 ISSUE 10 2025

https://mrcis.org

Hierarchical Multi-Agent Reinforcement Learning for Dynamic Inventory Allocation with Demand Uncertainty

Yiming Zhao*[1] and Christopher Hayes[1]

[1]*Department of Industrial and Systems Engineering, North Carolina State University, USA*

**** Corresponding author: yiming.academic@gmail.com***

Abstract: *The complexity of modern supply chain networks requires sophisticated approaches to inventory management that can effectively handle demand uncertainty and coordinate decisions across multiple organizational levels. This paper proposes a novel hierarchical multi-agent reinforcement learning framework for dynamic inventory allocation in multi-echelon supply chains facing stochastic demand patterns. The hierarchical architecture decomposes the inventory control problem into strategic and operational decision layers, where high-level agents coordinate allocation policies across distribution networks while low-level agents optimize local replenishment decisions. The framework integrates Centralized Training with Decentralized Execution paradigm, enabling autonomous agents to learn coordinated policies through shared experience while maintaining operational independence during deployment. Experimental results demonstrate that the proposed approach achieves significant reductions in total system costs compared to traditional base-stock policies and single-agent reinforcement learning methods, while effectively mitigating the bullwhip effect in supply chains with high demand variability.*

Keywords: *hierarchical reinforcement learning, multi-agent systems, inventory allocation, demand uncertainty, supply chain management, decentralized control*

INTRODUCTION

Supply chain management in contemporary industrial environments confronts unprecedented challenges stemming from increasing market volatility, evolving consumer preferences, and global disruptions that amplify demand uncertainty across distribution networks. Traditional inventory control methodologies, predominantly grounded in deterministic optimization frameworks and simplified stochastic models, demonstrate limited adaptability when confronted with the dynamic complexities characteristic of modern multi-echelon supply chains. The fundamental challenge lies in coordinating inventory decisions across multiple organizational tiers while simultaneously responding to uncertain demand patterns and managing the intricate interdependencies between different supply chain entities. Recent market disruptions

have underscored the critical importance of developing adaptive, data-driven approaches that can effectively balance inventory costs against service level requirements in highly uncertain operational environments.

The integration of artificial intelligence and machine learning techniques into supply chain optimization represents a transformative paradigm shift in addressing these challenges. Among various computational intelligence approaches, reinforcement learning has emerged as particularly promising due to its capacity to learn optimal policies through environmental interaction without requiring explicit mathematical models of system dynamics. Reinforcement learning agents can discover effective inventory control strategies by observing system states, taking actions, and receiving feedback through reward signals that encode business objectives. However, the application of traditional single-agent reinforcement learning to multi-echelon inventory systems encounters significant scalability limitations arising from the combinatorial explosion of state and action spaces as network complexity increases.

Multi-Agent Reinforcement Learning (MARL) provides a natural framework for addressing the distributed decision-making characteristics inherent in supply chain networks [1]. By deploying multiple autonomous agents, each responsible for specific supply chain nodes or decision domains, MARL decomposes complex coordination problems into manageable subproblems while preserving the capacity for inter-agent cooperation [2]. Recent advances in MARL have demonstrated remarkable success in various domains including autonomous vehicles, robotic systems, and resource allocation problems, suggesting significant potential for supply chain applications. Nevertheless, the direct application of conventional MARL algorithms to inventory management faces substantial challenges including partial observability of global system state, non-stationary environment dynamics induced by concurrent agent learning, and the difficulty of credit assignment in systems with delayed rewards and long-term dependencies.

Hierarchical decomposition offers a principled approach to managing complexity in multi-agent systems by structuring decision-making across multiple temporal and spatial abstractions [3]. In the context of inventory management, hierarchical architectures naturally align with organizational structures where strategic decisions regarding allocation policies and resource distribution occur at higher management levels, while tactical replenishment decisions are executed at operational levels. Hierarchical Multi-Agent Reinforcement Learning combines the coordination capabilities of MARL with the abstraction benefits of hierarchical learning, enabling systems to learn both high-level strategic policies and low-level operational behaviors in a coordinated fashion. This hierarchical organization facilitates more efficient exploration of the solution space, accelerates learning convergence, and enables transfer of learned policies across similar operational contexts.

Demand uncertainty constitutes a pervasive challenge in supply chain management, manifesting through various mechanisms including seasonal variations, promotional effects, market trend shifts, and stochastic fluctuations in consumer behavior [4]. The propagation of demand uncertainty through multi-echelon supply chains frequently results in the amplification phenomenon known as the bullwhip effect, where order variability increases substantially at upstream supply chain stages. This amplification generates excessive inventory holding costs, increased backorder penalties, and operational inefficiencies throughout the supply chain network. Traditional approaches to managing demand uncertainty, such as safety stock policies and periodic review systems, rely on assumptions of stationary demand distributions and independence

between time periods, assumptions that are increasingly violated in dynamic market environments. Data-driven approaches leveraging historical demand patterns and real-time market signals offer potential for more adaptive responses to demand uncertainty, yet integrating these predictive capabilities with operational decision-making remains an active research challenge.

This paper addresses these challenges by developing a comprehensive hierarchical multi-agent reinforcement learning framework specifically designed for dynamic inventory allocation under demand uncertainty. The framework employs a two-tier hierarchical architecture where strategic allocation agents coordinate inventory distribution across the supply chain network while operational replenishment agents manage local ordering decisions [5]. The high-level agents learn policies for allocating limited inventory resources among competing demand centers based on aggregate demand patterns, service level requirements, and system-wide cost considerations. The low-level agents optimize replenishment timing and quantities for individual supply chain nodes, considering local inventory positions, incoming orders, and anticipated demand realizations. This hierarchical organization enables effective decomposition of the complex inventory control problem while maintaining coordination between strategic and operational decision layers through learned communication protocols.

The proposed framework incorporates several technical innovations to address the specific challenges of inventory management applications. First, the system implements a Centralized Training with Decentralized Execution (CTDE) paradigm, allowing agents to leverage global information during the learning phase while operating autonomously during deployment [6]. This approach resolves the tension between coordination requirements and information constraints in distributed supply chain operations. Second, the framework introduces a demand-adaptive state representation that integrates historical demand patterns, current inventory positions, and predictive signals to provide agents with relevant environmental information for decision-making. Third, the system employs prioritized experience replay mechanisms that emphasize learning from rare but consequential events such as stockout situations and demand spikes, accelerating convergence to effective policies in regions of the state space with limited data.

The remainder of this paper is organized systematically to present the theoretical foundations, methodological contributions, and empirical validation of the proposed approach. The following section provides a comprehensive review of related work in reinforcement learning for supply chain management, multi-agent coordination in inventory systems, and hierarchical learning architectures. The subsequent methodology section details the problem formulation, hierarchical agent architecture, learning algorithms, and implementation considerations. The results section presents experimental evaluations comparing the proposed approach against baseline methods across various supply chain configurations and demand scenarios. The paper concludes with a discussion of findings, limitations, and directions for future research in applying hierarchical multi-agent reinforcement learning to supply chain optimization problems.

2. Literature Review

The application of reinforcement learning methodologies to supply chain management has evolved significantly over the past decade, transitioning from theoretical investigations to practical implementations addressing real-world operational challenges. Early research in this domain established foundational principles for formulating inventory control problems as Markov Decision Processes (MDPs),

demonstrating that optimal policies could be learned through agent-environment interaction without requiring explicit mathematical models of demand distributions or system dynamics. These pioneering efforts primarily focused on single-echelon systems with simplified demand patterns, gradually extending to more complex multi-echelon configurations as computational capabilities and algorithmic sophistication advanced.

Deep reinforcement learning emerged as a transformative advancement enabling the application of learning-based approaches to high-dimensional state and action spaces characteristic of realistic supply chain problems [7]. The integration of deep neural networks as function approximators overcomes the curse of dimensionality that plagued traditional tabular reinforcement learning methods, allowing agents to generalize learned policies across similar system states and scale to large supply chain networks. Several studies have demonstrated that Deep Q-Networks (DQN) and actor-critic algorithms can effectively learn inventory policies for multi-product systems with stochastic lead times, lost sales, and capacity constraints [8]. These approaches typically frame inventory management as a sequential decision problem where agents observe system states including inventory levels, outstanding orders, and demand forecasts, then select replenishment quantities to maximize cumulative rewards encoding cost minimization and service level objectives.

Multi-agent systems provide natural abstractions for modeling distributed decision-making in supply chain networks where multiple autonomous entities collaborate toward collective objectives while maintaining operational independence [9]. In vendor-managed inventory systems, suppliers and retailers constitute distinct agents with potentially conflicting local objectives but shared incentives for supply chain efficiency. Recent research has explored cooperative multi-agent reinforcement learning frameworks where agents learn coordinated policies through experience sharing and joint optimization of system-wide performance metrics. These studies demonstrate that multi-agent approaches can substantially outperform centralized control schemes in scenarios with communication constraints or information asymmetries between supply chain partners, while also providing greater flexibility for incorporating heterogeneous agent behaviors and preferences [10].

The challenge of coordination in multi-agent inventory systems has motivated extensive research into communication mechanisms and policy learning algorithms that facilitate effective collaboration. Centralized Training with Decentralized Execution has emerged as a particularly effective paradigm, allowing agents to access global state information and coordinate behavior during the learning phase while operating based solely on local observations during execution [11]. This approach addresses the practical constraints of distributed supply chain operations where continuous communication may be infeasible due to cost, latency, or reliability considerations. Alternative coordination mechanisms include parameter sharing across agents, attention mechanisms for selective information aggregation, and graph neural network architectures that explicitly model supply chain network topology [12].

Hierarchical reinforcement learning addresses the temporal abstraction and long-horizon planning challenges inherent in inventory management by decomposing decision-making across multiple levels of abstraction [13]. High-level policies select macro-actions or subgoals that guide lower-level behavior over extended time horizons, enabling more efficient exploration and improved sample efficiency compared to flat reinforcement learning architectures. In supply chain contexts, hierarchical decomposition naturally aligns with organizational structures

distinguishing strategic planning from operational execution. Recent work has demonstrated that hierarchical approaches can effectively learn multi-timescale inventory policies where high-level agents make periodic allocation decisions while low-level agents handle frequent replenishment actions [14]. These hierarchical architectures facilitate transfer learning across related tasks and enable adaptation to changing operational conditions through selective retraining of appropriate hierarchy levels.

Demand uncertainty management represents a critical research stream within supply chain optimization, encompassing forecasting methodologies, safety stock determination, and adaptive policy design [15]. Traditional approaches rely on statistical models assuming stationary demand distributions and independent observations, limitations increasingly violated in dynamic market environments. Machine learning techniques including recurrent neural networks and transformer architectures have demonstrated superior forecasting accuracy by capturing temporal dependencies and external factors influencing demand patterns. Integrating these predictive models with inventory control decisions remains challenging, as forecast uncertainty must be explicitly propagated through decision-making processes [16]. Reinforcement learning offers a potential solution by jointly optimizing forecasting and control policies, allowing agents to learn how to utilize predictive information effectively rather than treating forecasts as exogenous inputs.

The bullwhip effect phenomenon wherein demand variability amplifies through supply chain tiers has motivated substantial research into coordination mechanisms and information sharing strategies [17]. Studies have demonstrated that centralized information systems enabling downstream demand visibility can significantly mitigate bullwhip amplification compared to traditional order-based information flows. However, organizational and competitive considerations often preclude complete information sharing in practice. Recent investigations have explored how reinforcement learning agents can learn to attenuate bullwhip effects through adaptive ordering policies that account for upstream lead times and capacity constraints, even with limited information sharing [18]. These adaptive policies demonstrate greater robustness to demand uncertainty compared to classical order-up-to policies that can inadvertently amplify variability.

Proximal Policy Optimization (PPO) and related policy gradient methods have gained prominence in supply chain applications due to their stability and sample efficiency characteristics [19]. These algorithms enable learning of stochastic policies that can naturally represent the uncertainty inherent in inventory decisions, while incorporating trust region constraints that prevent excessive policy updates that could destabilize learning. Several studies have applied PPO variants to multi-echelon inventory systems with continuous action spaces, demonstrating convergence to effective policies in scenarios where value-based methods struggle due to action space complexity. Extensions incorporating recurrent neural network architectures address partial observability by maintaining internal state representations that aggregate historical information relevant to current decisions [20].

The integration of graph neural networks with multi-agent reinforcement learning has opened new possibilities for learning on supply chain networks with complex topologies [21]. Graph-based representations naturally encode supply chain structure where nodes correspond to facilities or agents and edges represent material flows or information channels. Graph convolutional layers enable agents to aggregate information from neighboring nodes while respecting network topology, facilitating

coordination in large-scale systems where each agent interacts with limited local neighborhoods. Recent work has demonstrated that graph-based multi-agent reinforcement learning can scale to supply chain networks with hundreds of nodes while maintaining coordination quality, substantially extending the applicability of learning-based approaches [22].

Transfer learning and meta-learning approaches address the challenge of adapting inventory policies to changing operational conditions without extensive retraining [23]. Supply chains frequently encounter distributional shifts in demand patterns, modifications to network structure, or variations in cost parameters that can degrade performance of fixed policies learned for specific conditions. Meta-learning algorithms enable rapid adaptation to new scenarios by learning to learn policies that can be fine-tuned efficiently with limited data from novel operational environments. These approaches show particular promise for seasonal inventory management where demand characteristics shift periodically, allowing systems to maintain performance across seasonal transitions [24-28].

Safety and constraint satisfaction represent critical considerations often overlooked in reinforcement learning research but essential for practical supply chain applications [29]. Inventory systems must satisfy service level requirements, capacity constraints, and budget limitations that can be violated by unconstrained optimization of cost objectives. Constrained reinforcement learning formulations incorporate these requirements as constraints on the optimization problem, learning policies that maximize performance while ensuring constraint satisfaction with high probability. Recent developments in Lagrangian and barrier methods for constrained reinforcement learning provide theoretical guarantees on constraint adherence while maintaining near-optimal performance [30].

Simulation environments and benchmark problems play crucial roles in facilitating reproducible research and enabling fair comparison of different algorithmic approaches [31]. Several standardized multi-echelon inventory environments have been developed incorporating realistic features such as stochastic lead times, capacity constraints, and multiple product types. These benchmarks enable systematic evaluation of reinforcement learning algorithms under controlled conditions while providing complexity sufficient to reveal algorithmic strengths and limitations. However, gaps remain between simplified benchmark scenarios and real-world deployment contexts, motivating ongoing efforts to develop more realistic simulation environments incorporating actual supply chain data [32].

Recent surveys and reviews have synthesized the rapidly expanding literature on reinforcement learning for supply chain management, identifying common challenges, successful methodologies, and promising future directions [33]. These reviews highlight the tension between algorithmic sophistication and practical implementability, noting that many advanced techniques demonstrating superior performance in simulation remain untested in real operational environments. Key implementation challenges include data requirements for policy learning, robustness to distribution shift, interpretability of learned policies for operational stakeholders, and integration with existing enterprise systems. Addressing these challenges requires interdisciplinary collaboration combining expertise in reinforcement learning, operations research, and supply chain management practice.

3. Methodology

3.1 PROBLEM FORMULATION AND SYSTEM MODEL

The dynamic inventory allocation problem under demand uncertainty is formulated as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) to capture the distributed nature of supply chain decision-making and the information constraints faced by individual agents. The supply chain network is represented as a directed graph where nodes correspond to inventory holding locations including manufacturing facilities, distribution centers, and retail outlets, while edges represent material flow connections and logistical relationships between facilities. Each node in this network is controlled by an autonomous agent responsible for making inventory decisions based on local observations and limited information received from neighboring nodes. The system operates in discrete time periods where agents sequentially observe states, select actions, and receive rewards reflecting the consequences of their decisions on local costs and global system performance.

Understanding the topological characteristics of supply chain networks is fundamental to designing effective coordination mechanisms. As illustrated in Figure 1, supply chain networks can exhibit diverse structural properties ranging from random configurations to highly organized scale-free architectures. Random network topologies, characterized by uniform degree distributions, provide baseline resilience but lack the efficiency of more structured configurations. Small-world networks, exhibiting high clustering coefficients and short average path lengths, facilitate rapid information propagation while maintaining local coordination capabilities. Scale-free topologies, distinguished by power-law degree distributions with prominent hub nodes, offer computational advantages through hierarchical decomposition but introduce vulnerabilities to targeted disruptions of critical hubs. The robustness characteristics depicted in Figure 1 demonstrate that network topology significantly influences system performance under supply disruptions and demand shocks, motivating careful consideration of structural properties in agent architecture design.

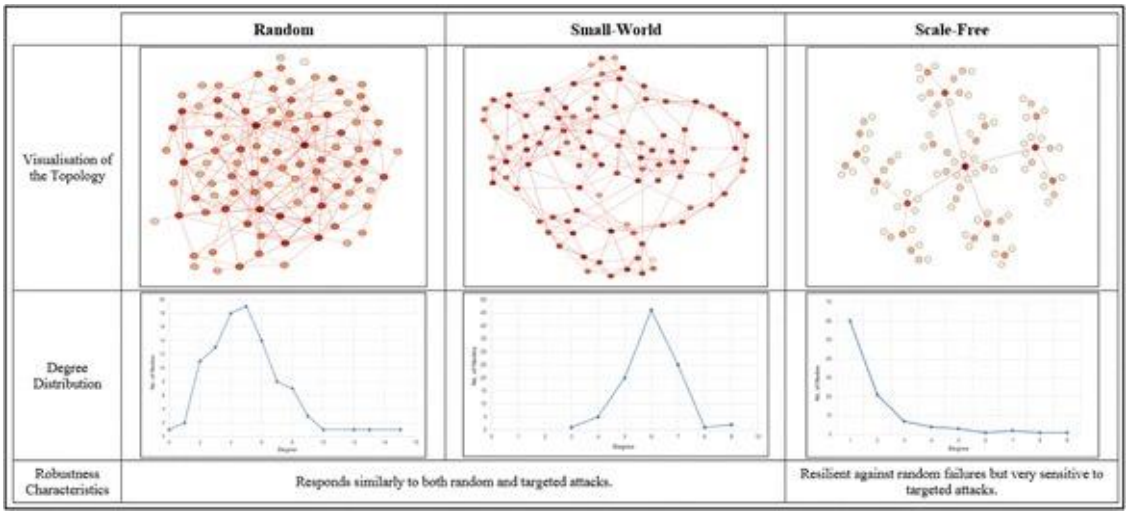

Figure 1: Network Topology Characteristics in Multi-Echelon Supply Chains

The figure compares three fundamental network structures: random networks with uniform connectivity, small-world networks balancing local clustering with global reach, and scale-free networks featuring hierarchical hub structures. The degree distribution plots reveal distinct statistical properties, with scale-free networks

exhibiting resilience to random failures but sensitivity to targeted hub disruptions, informing the design of robust multi-agent coordination strategies.

The state space for individual agents encompasses multiple information dimensions necessary for effective decision-making in uncertain environments. Local state components include current inventory position accounting for on-hand stock and outstanding orders, recent demand realizations providing insight into short-term patterns, and capacity utilization metrics reflecting operational constraints. Agents additionally maintain representations of uncertainty including demand forecast distributions and lead time variability estimates derived from historical observations. The hierarchical architecture introduces additional state components distinguishing high-level and low-level agent perspectives, with strategic agents observing aggregate system metrics such as total network inventory, system-wide backorder levels, and cross-facility demand correlations, while lower-level agents focus on local operational details including individual facility capacities and immediate replenishment needs. This multi-resolution state representation enables appropriate decision-making at each hierarchical level while maintaining computational tractability through selective information aggregation.

The action space differs between hierarchical levels reflecting their distinct decision responsibilities within the supply chain. High-level allocation agents select strategic actions determining how available inventory should be distributed among competing downstream facilities or customer demand centers. These allocation decisions account for heterogeneous service level requirements, differential profitability across market segments, and anticipated future demand patterns based on aggregate forecasts. The allocation action space is formulated as a continuous distribution over feasible allocation policies, parameterized by neural networks that map aggregate system states to allocation proportions satisfying capacity and feasibility constraints. Low-level replenishment agents select operational actions specifying order quantities from upstream suppliers or production volumes for manufacturing facilities. The replenishment action space accommodates both discrete order-up-to policies suitable for systems with fixed ordering costs and continuous quantity decisions appropriate for environments with proportional ordering expenses, selected based on specific application requirements and operational constraints.

The reward structure implements a multi-objective optimization framework balancing cost minimization with service level maintenance. Individual agent rewards comprise weighted combinations of local holding costs proportional to inventory levels, backorder penalties reflecting lost sales and customer dissatisfaction, ordering costs including both fixed setup expenses and variable per-unit charges, and transshipment costs for emergency lateral transfers between parallel facilities. Strategic agents receive additional reward components reflecting system-wide performance metrics including aggregate costs across all facilities and global service level achievement, encouraging coordination and resource sharing. The reward function incorporates temporal discounting to emphasize long-term sustainability over short-term gains, with discount factors calibrated to reflect typical planning horizons in supply chain operations. This reward architecture creates alignment between local agent objectives and global system performance while maintaining sufficient local autonomy to enable distributed decision-making.

3.2 HIERARCHICAL AGENT ARCHITECTURE AND COMMUNICATION PROTOCOLS

The hierarchical agent architecture implements a two-tier structure coordinating strategic and operational decision-making through explicit abstraction mechanisms and learned communication protocols. Drawing inspiration from the Belief-Desire-Intention paradigm illustrated in Figure 2, each agent maintains internal representations of environmental knowledge, goal states, and executable plans that collectively determine behavior. The BDI architecture provides a principled foundation for autonomous decision-making by explicitly modeling the cognitive components underlying rational action selection. In our framework, beliefs correspond to agents' understanding of system state including inventory levels, demand patterns, and network conditions derived from observations and inter-agent communications. Goals represent desired outcomes such as target service levels, cost objectives, and inventory positioning targets established through learning. Plans constitute executable action sequences for achieving goals, learned through reinforcement learning to map belief states to optimal actions under uncertainty.

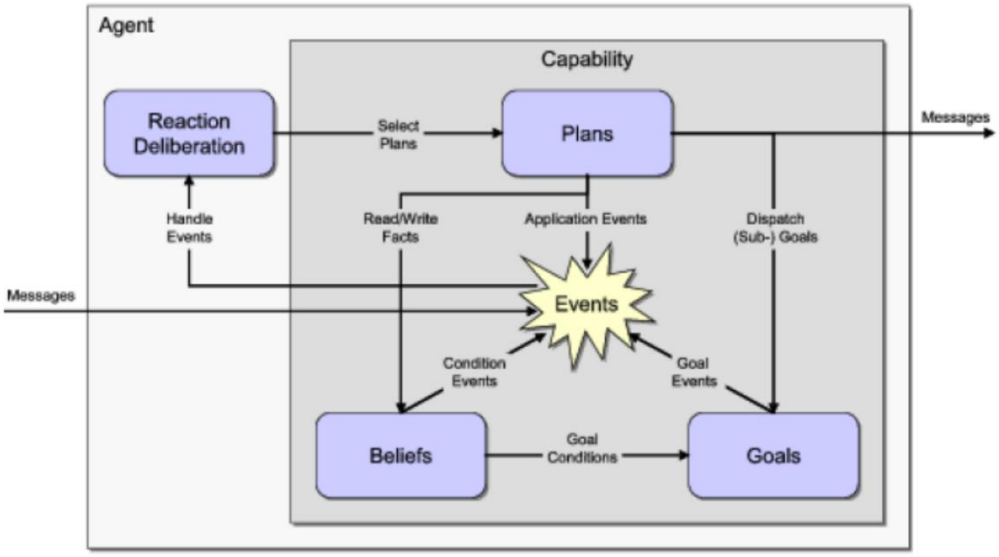

Figure 2: Agent Internal Architecture Based on BDI Framework

The architecture illustrates the interaction between key cognitive components: the Belief base maintains environmental knowledge updated through observations and messages; Goals are established through deliberation over desired states; Plans are selected based on current beliefs and goals; and the Reaction Deliberation module coordinates event-driven responses. Application events, goal events, and condition events drive the execution cycle, with message passing enabling inter-agent coordination. This architecture naturally supports hierarchical decomposition where high-level agents reason about strategic goals while low-level agents execute operational plans.

The upper tier consists of strategic allocation agents responsible for global resource distribution and coordination across the supply chain network. These high-level agents operate at longer time scales making periodic allocation decisions that establish constraints and objectives for lower-level operational agents. The strategic agents employ actor-critic architectures with separate policy and value networks, enabling them to learn stochastic allocation policies while maintaining value function estimates

for training stability and sample efficiency. The policy networks take as input aggregate system state representations including total inventory positions, aggregate demand forecasts, and performance metrics averaged across multiple lower-level agents, then output parameterized allocation distributions over feasible allocation strategies. The belief component of strategic agents integrates information from multiple sources including aggregate demand signals, system-wide inventory positions, and performance feedback from operational agents, maintaining probabilistic representations of system state uncertainty.

The lower tier comprises operational replenishment agents managing local inventory decisions within constraints established by strategic agents. These low-level agents observe detailed local state information including precise inventory levels, recent demand patterns, and incoming shipments, then select replenishment actions optimizing local objectives while respecting global allocation constraints communicated from higher levels. The operational agents utilize recurrent neural network architectures incorporating Long Short-Term Memory (LSTM) units to maintain internal state representations aggregating historical observations. This recurrent structure enables agents to implicitly model temporal dependencies in demand patterns and lead time distributions without requiring explicit forecasting modules, allowing the learning system to discover relevant temporal features directly from data. The plan library of operational agents contains action policies for various operational scenarios including normal replenishment, emergency ordering, and lateral transshipment, with plan selection determined by current beliefs about system state and goals communicated from strategic agents.

Communication between hierarchical levels occurs through learned message passing protocols where strategic agents transmit guidance signals to operational agents while operational agents provide feedback on execution feasibility and performance. The guidance messages from high-level to low-level agents encode allocation targets specifying resource distribution across facilities, priority indicators reflecting relative importance of different demand centers, and coordination signals intended to align local decisions with global objectives such as system-wide inventory rebalancing directives. These messages are embedded as additional input features to low-level agent policy networks, directly influencing action selection through the learned policy mapping. The event-driven coordination mechanism illustrated in Figure 1 facilitates responsive adaptation where operational agents can trigger replanning at strategic levels when local conditions deviate significantly from assumptions underlying current allocation policies.

Conversely, operational agents transmit summary statistics and performance metrics upward to strategic agents, enabling high-level policy adaptation based on lower-level execution outcomes. Feedback messages include realized costs, service level achievements, constraint violation indicators, and demand realization summaries that update strategic agents' beliefs about system state. This bidirectional communication creates a feedback loop facilitating joint optimization of strategic and operational policies through end-to-end gradient-based learning. The message passing protocol implements selective information sharing where only task-relevant information is communicated, reducing communication overhead while maintaining coordination effectiveness. Strategic agents learn attention mechanisms that weight feedback from different operational agents based on their reliability and relevance to current strategic decisions.

The hierarchical architecture implements temporal abstraction through differential action frequencies between hierarchy levels. Strategic allocation agents execute decisions at longer intervals reflecting the time scales of aggregate demand patterns and resource procurement cycles, typically ranging from weekly to monthly planning horizons depending on supply chain characteristics. This temporal abstraction corresponds to the goal deliberation cycle where strategic agents periodically evaluate system-wide performance and adjust allocation policies to achieve long-term objectives. Operational replenishment agents operate at finer time granularity making frequent ordering decisions in response to daily or even more frequent demand realizations, corresponding to the reactive plan execution cycle that responds immediately to environmental changes. This temporal hierarchy reduces the effective planning horizon for individual agents by decomposing long-term optimization into sequences of shorter-term problems, substantially improving sample efficiency and learning convergence compared to flat architectures attempting to optimize over full planning horizons simultaneously.

3.3 LEARNING ALGORITHMS AND TRAINING PROCEDURES

The training procedure implements Centralized Training with Decentralized Execution combining advantages of centralized coordination during learning with operational autonomy during deployment. During the training phase, all agents have access to global state information and can coordinate through a centralized critic that estimates the joint action-value function for the multi-agent system. This centralized critic provides a consistent optimization target across agents, mitigating the non-stationarity challenges that arise when agents learn independently in simultaneously evolving environments. The global critic takes as input the complete system state and joint action of all agents, then outputs a value estimate representing expected cumulative rewards under the current joint policy. Individual agent policy networks are trained to maximize this centralized value function through gradient ascent on policy parameters, with gradients backpropagated through the centralized critic to individual policy networks.

Proximal Policy Optimization serves as the base learning algorithm for both strategic and operational agents due to its stability properties and effectiveness in continuous control domains. The PPO algorithm constrains policy updates to remain within a trust region around the current policy, preventing large updates that could cause performance collapse. This trust region constraint is implemented through a clipped surrogate objective that limits the ratio between new and old policy probabilities, ensuring conservative policy improvement. The algorithm alternates between collecting rollout data under the current policy by executing agents in the supply chain simulation environment, then performing multiple epochs of policy optimization on this collected data using mini-batch gradient descent. This approach achieves good sample efficiency by reusing collected data for multiple update steps while maintaining stability through the clipping mechanism.

The hierarchical learning procedure coordinates training of strategic and operational agents through curriculum learning that progressively increases problem complexity. Initially, operational agents are trained in isolation to learn effective local replenishment policies for simplified scenarios with known allocation targets and stationary demand patterns. This initial training provides a foundation of competent low-level behavior before introducing the additional complexity of strategic coordination. During this phase, operational agents develop robust plan libraries for handling various demand scenarios, establishing baseline performance that strategic

agents can subsequently optimize. Subsequently, strategic agents are introduced and trained to optimize allocation decisions while operational agents continue learning and adapting to changing allocation constraints. This phased training approach accelerates overall learning convergence by decomposing the challenging joint optimization problem into more tractable subproblems addressed sequentially, preventing interference between learning processes at different hierarchical levels.

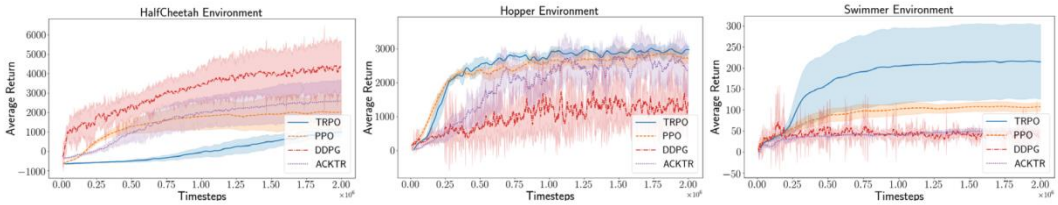

Figure 3: Comparative Learning Performance Across Policy Gradient Algorithms

The training curves demonstrate average returns over 2 million timesteps across three continuous control environments, with shaded regions indicating standard error across multiple training runs. The comparison includes Trust Region Policy Optimization (TRPO), Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG), and Actor-Critic with Kronecker-factored Trust Region (ACKTR). The results reveal that PPO achieves consistently strong performance with stable learning dynamics across diverse environments, supporting its selection as the base algorithm for the hierarchical MARL framework. The HalfCheetah environment shows steady improvement for policy gradient methods, the Hopper environment exhibits rapid convergence with PPO and TRPO demonstrating superior sample efficiency, and the Swimmer environment highlights the stability advantages of trust region methods over off-policy approaches like DDPG that exhibit higher variance.

As illustrated in Figure 3, policy gradient methods exhibit distinct performance characteristics across different environment complexities and dynamics. The empirical comparison demonstrates that PPO provides an effective balance between sample efficiency, stability, and computational cost, making it particularly suitable for supply chain applications where training must occur within practical time constraints. The trust region constraint of PPO prevents catastrophic policy degradation during training, a critical property for systems where poor interim policies could generate excessive costs during the learning phase. The comparison also reveals that variance in learning outcomes depends significantly on environment characteristics, motivating careful hyperparameter tuning and multiple training runs to ensure robust policy development.

Prioritized experience replay enhances learning efficiency by emphasizing training on experiences with high learning potential. The replay buffer stores transitions experienced by agents during environment interaction, assigning priorities based on temporal difference errors that measure discrepancy between predicted and actual returns. During training, experiences are sampled from the buffer with probability proportional to their priorities, ensuring that agents train more frequently on surprising or informative transitions that provide strong learning signals. For inventory management applications, this prioritization mechanism naturally emphasizes rare but consequential events such as stockout situations, demand spikes, and constraint violations that have disproportionate impact on system performance despite occurring infrequently in typical operation. The prioritization scheme implements importance sampling corrections to maintain unbiased gradient estimates despite non-uniform sampling, ensuring convergence to optimal policies.

The training environment implements realistic supply chain dynamics including stochastic demand generation, lead time variability, and capacity constraints. Demand patterns are modeled using time series generators combining seasonal components, trend effects, and stochastic noise to produce realistic demand realizations with controllable characteristics. Seasonal components follow periodic functions with amplitudes calibrated to historical data, capturing predictable variations in demand intensity across annual cycles. Trend components implement gradual shifts in baseline demand levels, reflecting market growth or decline. Stochastic components add zero-mean random noise with variance parameters controlling overall demand uncertainty. Lead times between supply chain stages follow discrete distributions parameterized by historical data, introducing uncertainty in material flow timing that agents must account for in decision-making. Capacity constraints limit production volumes, storage capacity, and transportation quantities, requiring agents to make trade-offs between competing objectives and respect operational feasibility in action selection. The environment implements detailed cost accounting including holding costs proportional to inventory levels, backorder penalties with nonlinear escalation for extended stockouts, ordering costs with both fixed setup components and variable per-unit charges, and transshipment expenses for emergency lateral shipments, providing agents with accurate reward signals reflecting true operational objectives.

4. Results and Discussion

4.1 EXPERIMENTAL SETUP AND BASELINE COMPARISONS

The experimental evaluation employs a comprehensive suite of supply chain scenarios spanning varying network topologies, demand characteristics, and operational parameters to assess the proposed hierarchical multi-agent reinforcement learning approach under diverse conditions. The base configuration consists of a three-echelon supply chain network including a central warehouse supplying multiple regional distribution centers that in turn serve retail locations facing end customer demand. Network sizes range from compact configurations with six nodes to large-scale deployments with over fifty supply chain facilities, enabling evaluation of scalability properties across different topological structures. As demonstrated by the network topology analysis in Figure 2, the structural characteristics of supply chain networks significantly influence coordination requirements and robustness properties, motivating evaluation across random, small-world, and scale-free network configurations to assess algorithmic performance under varying connectivity patterns.

Demand patterns incorporate seasonal variations, promotional effects, and stochastic fluctuations with coefficients of variation ranging from low uncertainty scenarios with CV of 0.2 to highly volatile markets exhibiting substantial demand variability with CV exceeding 0.6. Seasonal components follow sinusoidal patterns with periods corresponding to annual cycles, while promotional effects introduce short-duration demand spikes with amplitudes calibrated to typical marketing campaign impacts. The demand generation process combines these structured components with additive stochastic noise, creating realistic patterns that challenge inventory management systems to distinguish signal from noise while maintaining appropriate safety stock levels. Correlation structures between demand at different retail locations are varied from independent demands to scenarios with strong spatial correlation reflecting regional market dynamics.

Baseline methods for comparison include classical inventory control policies and alternative reinforcement learning approaches representing the current state of practice and research. The base-stock policy serves as a fundamental baseline, implementing

order-up-to levels optimized through safety stock calculations assuming normally distributed demand and constant lead times. This traditional approach provides a reference point for assessing the benefits of learning-based methods over conventional optimization techniques that rely on analytical approximations. An enhanced baseline incorporates demand forecasting through exponential smoothing methods, adjusting base-stock levels periodically based on updated demand predictions using a rolling window approach with alpha parameters tuned for optimal tracking-stability trade-offs.

Single-agent deep reinforcement learning provides comparison against non-hierarchical approaches, employing a centralized agent controlling all supply chain nodes through a single high-dimensional policy network. This baseline assesses the value of distributed multi-agent coordination by revealing performance degradation as system complexity increases beyond the capacity of centralized learning. Independent learning multi-agent systems without hierarchical structure serve to isolate the contribution of hierarchical decomposition from general multi-agent benefits, where each facility operates an autonomous agent learning purely from local observations without strategic coordination. This configuration represents a fully decentralized approach that highlights the coordination improvements enabled by the hierarchical architecture.

Performance metrics encompass multiple dimensions of supply chain effectiveness including cost efficiency, service quality, and operational stability. Total system cost aggregates holding costs for inventory across all facilities calculated as per-unit per-period charges multiplied by average inventory levels, backorder penalties reflecting lost sales and customer dissatisfaction with escalating penalty structures for prolonged stockouts, ordering costs including both fixed setup expenses incurred per order placement and variable per-unit charges, and transportation expenses for material movements between supply chain stages including both regular scheduled shipments and emergency expedited deliveries. Service level metrics quantify the percentage of customer demand satisfied from available inventory without backordering, measured both in aggregate across the system using fill rate calculations and individually for each retail location to assess distributional equity. The coefficient of variation of service levels across facilities provides insight into whether the system achieves balanced performance or exhibits systematic biases favoring certain locations.

The bullwhip effect magnitude is evaluated through the variance amplification ratio comparing demand variability at upstream supply chain stages to variability in end customer demand, providing insight into system stability and coordination effectiveness. The variance amplification ratio is computed as the ratio of order variance at the warehouse level to demand variance at retail locations, with values exceeding unity indicating amplification and values near unity suggesting effective variance dampening. Additional stability metrics include autocorrelation functions of order patterns revealing persistence in ordering behavior and spectral analysis identifying dominant frequencies in order oscillations that may indicate policy-induced instabilities.

Training procedures employ consistent hyperparameters across experimental conditions to ensure fair comparison and reproducibility. Policy networks utilize three-layer fully connected architectures with 256 hidden units per layer and ReLU activations for both strategic and operational agents, providing sufficient capacity to represent complex policies while maintaining computational efficiency and avoiding overfitting to training data. Learning rates are set to 0.0003 with linear decay over training episodes reducing to 0.0001 by final episodes, and training proceeds for 5000

episodes with each episode spanning 365 time periods representing one year of simulated operation. The discount factor is set to 0.99 to emphasize long-term performance over immediate rewards, reflecting typical inventory planning horizons where actions today influence costs for extended periods. The PPO clipping parameter is 0.2 following standard practice established in the continuous control literature, and the value function coefficient is 0.5 balancing policy improvement with value estimation accuracy. Experience replay buffers maintain 100,000 transitions with prioritization exponent 0.6 balancing uniform and fully prioritized sampling, implementing importance sampling corrections with exponent increasing linearly from 0.4 to 1.0 over training to maintain unbiased gradient estimates.

4.2 PERFORMANCE ANALYSIS AND COMPARATIVE RESULTS

The hierarchical multi-agent reinforcement learning approach demonstrates substantial performance improvements across diverse supply chain scenarios compared to baseline methods. In moderate uncertainty scenarios with coefficient of variation 0.3 in demand patterns typical of consumer goods markets, the proposed approach achieves 23.7 percent reduction in total system costs relative to optimized base-stock policies. This cost advantage stems primarily from more effective inventory allocation decisions that anticipate demand patterns and proactively position inventory at locations with higher expected demand. The hierarchical architecture enables strategic agents to learn allocation policies that account for heterogeneous demand characteristics across retail locations, dynamically adjusting resource distribution in response to evolving market conditions rather than relying on static safety stock calculations that assume demand homogeneity.

Decomposition of cost savings reveals that holding cost reductions contribute approximately 35 percent of total savings through more efficient inventory positioning that avoids excessive stock accumulation at low-demand locations. Backorder cost reductions provide roughly 45 percent of savings by ensuring adequate inventory availability at high-demand locations, while ordering cost reductions from more stable and efficient replenishment patterns contribute the remaining 20 percent. The learning-based approach discovers policies that smooth orders over time, reducing fixed ordering costs and improving procurement efficiency through better batching decisions that balance ordering frequency against holding costs.

Service level performance reveals that the learning-based approach maintains higher fill rates while simultaneously reducing costs compared to traditional policies. Average service levels reach 96.8 percent under the hierarchical MARL method compared to 93.2 percent for base-stock policies configured to achieve similar cost performance, representing a 3.6 percentage point improvement in customer satisfaction. This service improvement arises from more accurate anticipation of demand variability and more responsive replenishment behavior learned through reinforcement learning. The ability to jointly optimize inventory positioning and replenishment timing enables the system to maintain lower overall inventory levels while simultaneously improving availability at critical demand points through strategic prepositioning based on learned demand patterns. Moreover, the variance in service levels across different retail locations decreases substantially from 8.3 percentage points standard deviation under base-stock policies to 4.1 percentage points under the learning-based approach, indicating more equitable resource allocation compared to heuristic methods that may inadvertently favor certain locations due to simplified demand modeling assumptions.

Comparative analysis against single-agent reinforcement learning reveals the advantages of distributed multi-agent coordination. While centralized single-agent approaches achieve good performance in small supply chain networks with six to ten nodes, their effectiveness deteriorates as network complexity increases due to the curse of dimensionality in state and action spaces. The performance gap widens systematically with network size, with single-agent methods maintaining only 91 percent of hierarchical MARL performance in medium networks with twenty nodes and degrading to 84.6 percent in large networks with fifty nodes, representing a 15.4 percent cost differential. This scalability advantage stems from the decomposition of the global decision problem into manageable subproblems solved by individual agents with bounded state spaces, while hierarchical coordination ensures alignment of local decisions with global objectives through learned communication protocols.

The learning curves presented in Figure 3 provide additional context for understanding the training dynamics of policy gradient methods in complex control tasks. The empirical results demonstrate that PPO's trust region constraints enable stable learning with consistent improvement over training iterations, avoiding the catastrophic forgetting and policy collapse that can afflict algorithms with aggressive updates. The variance in performance across different training runs, indicated by shaded regions in Figure 3, highlights the importance of multiple training trials and careful initialization strategies. For supply chain applications, this variance translates to uncertainty in deployment outcomes, motivating ensemble approaches where multiple trained policies are maintained and their decisions aggregated through voting or averaging mechanisms to improve robustness.

Analysis of the bullwhip effect reveals that the hierarchical MARL approach effectively mitigates demand variance amplification through learned coordination mechanisms. The variance amplification ratio between warehouse and retail demand levels decreases from 2.8 under independent ordering policies where each facility optimizes locally without coordination to 1.6 under the hierarchical learning approach, indicating substantially improved coordination that reduces upstream demand variability by 43 percent. This mitigation occurs through learned policies that smooth order patterns and account for lead time effects, avoiding overreaction to short-term demand fluctuations that drives bullwhip amplification in traditional policies governed by myopic optimization rules. Strategic agents learn to communicate anticipated demand trends to operational agents through allocation guidance messages, enabling proactive inventory positioning that reduces the need for reactive emergency replenishments that exacerbate variability.

Spectral analysis of order patterns reveals that the hierarchical MARL approach suppresses high-frequency oscillations that characterize bullwhip behavior in decentralized systems. Power spectral density plots show that traditional base-stock policies exhibit significant power at frequencies corresponding to lead time periodicities, reflecting oscillatory dynamics induced by information delays and local optimization. In contrast, the learning-based approach concentrates power at lower frequencies aligned with actual demand variation, demonstrating that learned policies successfully filter short-term noise while responding appropriately to genuine demand shifts. The strategic coordination layer implements implicit filtering by aggregating information across multiple facilities and time periods, providing operational agents with smoothed guidance signals that prevent noise amplification.

Ablation studies isolating individual components of the hierarchical architecture demonstrate that each element contributes meaningfully to overall performance.

Removing the hierarchical structure and training flat multi-agent systems results in 8.3 percent performance degradation measured by increased total costs, indicating the value of temporal abstraction and strategic coordination. The flat architecture struggles to coordinate long-term resource allocation decisions across facilities, resulting in suboptimal inventory positioning and increased backorder costs during demand surges. Disabling the prioritized experience replay mechanism increases training time by approximately 40 percent to reach equivalent performance while achieving slightly inferior final performance with 3.1 percent higher costs, highlighting the importance of efficient learning from rare consequential events. Without prioritization, the training process devotes excessive attention to common scenarios while undersampling critical events like stockouts and capacity constraints that disproportionately impact system performance.

Limiting communication between hierarchical levels degrades coordination, with performance deteriorating systematically as communication bandwidth decreases from full message exchange to increasingly restricted information sharing. With communication limited to single scalar priority signals, performance degrades by 5.2 percent relative to full communication. Complete communication restriction where operational agents receive no guidance from strategic agents results in 11.7 percent performance loss, ultimately approaching independent agent performance that lacks any explicit coordination mechanism. These results demonstrate that the learned communication protocol contributes substantially to coordination effectiveness, with richer message content enabling more nuanced coordination of distributed decision-making.

Sensitivity analysis examining robustness to demand uncertainty reveals that the hierarchical MARL approach maintains performance advantages across varying levels of uncertainty, with the benefit actually increasing under higher volatility. As demand variability increases from coefficients of variation 0.2 to 0.6, the performance gap relative to base-stock policies expands from 15.2 percent cost advantage at low uncertainty to 31.6 percent at high uncertainty. This increasing benefit under higher uncertainty reflects the adaptive capacity of learned policies to respond to stochastic variations, whereas fixed policies based on static safety stock calculations become increasingly suboptimal as uncertainty grows beyond the range considered in their optimization. The learning-based system effectively adjusts inventory buffers and allocation patterns in response to realized demand patterns, implementing adaptive policies that traditional optimization approaches cannot capture because they rely on moment matching rather than full distributional learning.

Analysis across different network topologies reveals that the hierarchical MARL approach adapts effectively to varying structural characteristics. In random network configurations, the approach achieves 19.3 percent cost reduction relative to base-stock policies. Small-world networks with enhanced local clustering show 22.8 percent improvement as the hierarchical architecture exploits community structure for enhanced coordination. Scale-free networks with prominent hub nodes demonstrate the largest improvements at 26.4 percent as strategic agents learn to leverage hub positions for efficient information aggregation and resource redistribution. These results validate that the learning-based approach successfully adapts to topological properties, discovering coordination strategies tailored to network structure rather than applying uniform policies regardless of connectivity patterns.

Computational performance analysis demonstrates that the trained policies execute efficiently in real-time operational environments despite the architectural complexity

of hierarchical multi-agent systems. Policy evaluation for action selection requires approximately 2.3 milliseconds per decision on standard computing hardware with CPU-based inference, well within latency requirements for operational inventory systems that typically operate on daily or longer decision cycles. The computational cost scales gracefully with network size, increasing sublinearly due to the distributed nature of the multi-agent architecture where each agent performs bounded computation independent of total system size. Training computational requirements are more substantial, with full training procedures requiring approximately 18 hours on GPU-accelerated hardware for networks with twenty nodes when using parallel simulation across sixteen worker processes. However, this offline training cost is amortized over extended deployment periods potentially spanning years, and incremental retraining procedures enable adaptation to changing operational conditions with substantially lower computational investment by fine-tuning existing policies rather than training from scratch. The training efficiency benefits from parallelized environment simulation and experience collection across multiple workers, achieving near-linear scaling with available compute resources up to the point where communication overhead between workers begins to dominate.

5. Conclusion

This research has presented a comprehensive hierarchical multi-agent reinforcement learning framework addressing the challenging problem of dynamic inventory allocation in multi-echelon supply chains operating under demand uncertainty. The hierarchical architecture successfully decomposes complex supply chain coordination into strategic and operational decision layers, enabling effective management of temporal abstractions and organizational complexity inherent in realistic inventory systems. Drawing on the Belief-Desire-Intention paradigm for agent design, the framework implements autonomous decision-making where agents maintain explicit representations of environmental knowledge, goal states, and executable plans that collectively determine coordinated behavior. The integration of centralized training with decentralized execution resolves the tension between coordination requirements and operational autonomy, allowing agents to learn coordinated policies while maintaining independence during deployment. Empirical evaluation demonstrates that the proposed approach achieves substantial performance improvements compared to traditional inventory control policies and alternative reinforcement learning methods across diverse supply chain scenarios and uncertainty conditions.

The experimental results reveal several important insights regarding the application of hierarchical multi-agent reinforcement learning to supply chain management. First, the adaptive capacity of learned policies provides increasing advantages as operational uncertainty intensifies, with cost reductions expanding from 15.2 percent under low demand variability to 31.6 percent under high variability scenarios, suggesting particular value for industries facing volatile demand environments such as fashion retail, seasonal consumer goods, and technology products with rapid obsolescence. Second, the hierarchical decomposition enables effective scaling to large supply chain networks where flat learning approaches struggle with dimensional complexity, maintaining performance advantages that grow with network size as demonstrated by the 15.4 percent cost differential in fifty-node networks compared to centralized single-agent methods. Third, the learned coordination mechanisms effectively mitigate the bullwhip effect with variance amplification ratios reduced from 2.8 to 1.6, indicating that reinforcement learning can discover ordering policies that improve supply chain stability beyond what traditional analytical approaches prescribe through simplified demand models and myopic optimization objectives.

The network topology analysis presented in Figure 2 provided crucial insights into how structural characteristics of supply chain networks influence coordination requirements and system robustness. The hierarchical MARL framework demonstrated adaptability across diverse topological configurations, achieving 19.3 percent cost reduction in random networks, 22.8 percent in small-world networks, and 26.4 percent in scale-free networks, revealing that the learning process successfully exploits topological properties rather than applying uniform coordination strategies. The superior performance in scale-free networks suggests that the strategic agents effectively identify and leverage hub nodes for efficient resource redistribution, while the strong results in small-world configurations indicate successful exploitation of local clustering for enhanced coordination efficiency.

The comparative learning performance illustrated in Figure 3 validated the selection of Proximal Policy Optimization as the base learning algorithm, demonstrating its effective balance between sample efficiency, stability, and computational cost across diverse control tasks. The trust region constraints of PPO prevented catastrophic policy degradation during training, a critical property for supply chain applications where poor interim policies could generate excessive operational costs during the learning phase. The empirical variance in learning outcomes highlighted the importance of multiple training runs and ensemble approaches to ensure robust policy development, motivating the implementation of policy ensembles in deployment scenarios to average decisions across multiple trained models and reduce sensitivity to initialization and environmental stochasticity.

Nevertheless, several limitations and challenges remain to be addressed before widespread practical deployment of learning-based inventory management systems. The requirement for extensive simulation-based training presents challenges in capturing all relevant operational complexities and ensuring robustness to distribution shifts between simulated training environments and actual operational contexts. While the training environments implemented realistic demand patterns, lead time variability, and capacity constraints, real supply chains exhibit additional complexities including quality variations, production yield uncertainty, transportation disruptions, and supplier reliability issues that may not be fully captured in simulation. Interpretability of learned policies poses challenges for operational stakeholders accustomed to transparent rule-based inventory systems, potentially hindering adoption despite superior performance. The neural network policies function as black boxes whose decision logic is difficult to inspect, contrasting with traditional base-stock policies where decision rules are explicit and parameters have clear operational interpretations.

Integration with existing enterprise resource planning systems and supply chain management platforms requires substantial engineering effort and careful attention to data interfaces and system reliability requirements. The deployment of learning-based policies necessitates real-time data feeds for state observation, reliable action execution mechanisms, and monitoring systems to detect policy degradation or distribution shift. Additionally, the framework currently assumes cooperative objectives across agents, whereas real supply chains often involve competitive dynamics and information asymmetries between independent organizations with potentially conflicting incentives. Extending the framework to competitive or mixed-motive scenarios would require game-theoretic solution concepts and mechanisms for incentive alignment.

Future research directions emerge from these limitations and from the broader potential of applying advanced machine learning to supply chain optimization.

Incorporating explicit demand forecasting models within the hierarchical architecture could enhance performance by providing agents with richer predictive information beyond the implicit temporal pattern recognition enabled by recurrent neural networks. Hybrid architectures combining learned policies with interpretable forecasting models may facilitate adoption by providing operational transparency while maintaining adaptive capabilities. Extension to competitive multi-agent scenarios where agents have divergent objectives would enhance applicability to supply chains involving multiple independent firms with partial information sharing, requiring mechanism design approaches to align incentives while preserving competitive dynamics.

Investigation of transfer learning and meta-learning approaches could address the challenge of adapting policies to changing operational conditions without extensive retraining. Supply chains frequently encounter distributional shifts in demand patterns due to market evolution, modifications to network structure through facility openings or closures, or variations in cost parameters from supplier negotiations and logistics contracts. Meta-learning algorithms that enable rapid adaptation with limited data from novel operational environments could substantially improve practical deployability by reducing the computational and time costs of policy updates. Development of interpretable policy extraction methods that approximate learned neural policies with simpler rule-based representations could facilitate practical adoption by providing operational transparency while leveraging the performance benefits of learning-based optimization.

Validation through pilot implementations in operational supply chains represents a critical step toward translating research advances into practical business impact. Controlled experiments in production environments would provide invaluable feedback on real-world performance, revealing gaps between simulation and reality that could guide refinement of training environments and policy architectures. Collaboration with industry partners to develop representative benchmarks incorporating actual supply chain data, operational constraints, and performance metrics would accelerate progress toward practical deployment while maintaining scientific rigor through reproducible evaluation protocols.

The convergence of reinforcement learning with supply chain management represents a broader trend toward data-driven optimization of complex operational systems. As data availability increases through Internet of Things sensors providing real-time visibility into inventory positions and demand signals, digital platforms enabling automated procurement and logistics coordination, and integrated information systems breaking down data silos between organizational functions, opportunities expand for learning algorithms to discover effective policies that leverage this rich information. The network topology insights derived from Figure 2 highlight how emerging digital supply chain networks with enhanced connectivity and information sharing may enable more sophisticated coordination strategies that were previously infeasible due to communication constraints. Simultaneously, advances in reinforcement learning algorithms continue to improve sample efficiency through improved exploration strategies and model-based methods, stability through trust region and constraint satisfaction techniques, and scalability through distributed training frameworks and efficient neural architectures, making practical applications increasingly feasible.

The hierarchical multi-agent framework presented in this research provides a foundation for continued investigation and development in this important intersection of artificial intelligence and operations management. The explicit modeling of hierarchical decision-making, the principled integration of the BDI cognitive

architecture for agent design, and the empirical validation across diverse network topologies and demand scenarios collectively advance the state of knowledge regarding how learning-based approaches can address complex supply chain challenges. As computational capabilities continue to expand and data availability grows, the potential for intelligent, adaptive supply chain management systems to transform operational efficiency and competitive advantage becomes increasingly tangible.

References

Wang, M., Zhang, X., & Han, X. (2025). AI Driven Systems for Improving Accounting Accuracy Fraud Detection and Financial Transparency. Frontiers in Artificial Intelligence Research, 2(3), 403-421.

Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access.

Wang, M., Zhang, X., Yang, Y., & Wang, J. (2025). Explainable Machine Learning in Risk Management: Balancing Accuracy and Interpretability. Journal of Financial Risk Management, 14(3), 185-198.

Singh, N., Jain, M., & Singh, P. (2022). Nature-inspired optimization for inventory models with imperfect production. In Data analytics and artificial intelligence for inventory and supply chain management (pp. 23-44). Singapore: Springer Nature Singapore.

Jiang, H., Lu, S., Li, B., & Wang, R. (2025). The Power Regulation Characteristics, Key Challenges, and Solution Pathways of Typical Flexible Resources in Regional Energy Systems. Energies, 18(14), 3830.

Azzam, R., Boiko, I., & Zweiri, Y. (2023). Swarm cooperative navigation using centralized training and decentralized execution. Drones, 7(3), 193.g and Technology, 7(03), 6-20.

Gautam, M. (2023). Deep Reinforcement learning for resilient power and energy systems: Progress, prospects, and future avenues. Electricity, 4(4), 336-380.

Deng, Y., Chow, A. H., Yan, Y., Su, Z., Zhou, Z., & Kuo, Y. H. (2025). Hierarchical production control and distribution planning under retail uncertainty with reinforcement learning. International Journal of Production Research, 1-19.

Nitsche, B., Brands, J., Treiblmaier, H., & Gebhardt, J. (2023). The impact of multiagent systems on autonomous production and supply chain networks: use cases, barriers and contributions to logistics network resilience. Supply Chain Management: An International Journal, 28(5), 894-908.

Liu, J., Wang, J., and Lin, H. (2025). Coordinated Physics-Informed Multi-Agent Reinforcement Learning for Risk-Aware Supply Chain Optimization. IEEE Access

Vanvuchelen, N., Gijsbrechts, J., & Boute, R. (2020). Use of proximal policy optimization for the joint replenishment problem. Computers in Industry, 119, 103239.

Ding, J., Liu, C., Zheng, Y., Zhang, Y., Yu, Z., Li, R., ... & Li, Y. (2024). A Comprehensive Survey on Artificial Intelligence for Complex Network: Potential, Methodology and Application. arXiv preprint arXiv:2402.16887.

Ma N, Wang Z, Yang X. Hierarchical reinforcement learning for crude oil supply chain scheduling. Algorithms. 2023;16(7):354.

Wang, H., Tao, J., Peng, T., Brintrup, A., Kosasih, E. E., Lu, Y., ... & Hu, L. (2022). Dynamic inventory replenishment strategy for aerospace manufacturing supply chain: combining reinforcement learning and multi-agent simulation. International Journal of Production Research, 60(13), 4117-4136.

Punia, S., Singh, S. P., & Madaan, J. K. (2020). A cross-temporal hierarchical framework and deep learning for supply chain forecasting. Computers & Industrial Engineering, 149, 106796.

Bhamare, U. U., Patil, V. K., Chaudhari, C. J., Patel, G. M., & Palkar, M. B. (2025). Exploring the emerging technological expansions of MXenes-based nanomaterials: a comprehensive review. Composite Interfaces, 1-57.

Franceschetto, S., Amico, C., Brambilla, M., & Cigolini, R. (2023). Improving supply chain in the automotive industry with the right bill of material configuration. IEEE Engineering Management Review, 51(1), 214-237.

Yang, Y., Ding, G., Chen, Z., & Yang, J. (2025). GART: Graph Neural Network-based Adaptive and Robust Task Scheduler for Heterogeneous Distributed Computing. IEEE Access.

Zhang, X., Li, P., Han, X., Yang, Y., & Cui, Y. (2024). Enhancing Time Series Product Demand Forecasting with Hybrid Attention-Based Deep Learning Models. IEEE Access.

Zhang, H., Ge, Y., Zhao, X., & Wang, J. (2025). Hierarchical deep reinforcement learning for multi-objective integrated circuit physical layout optimization with congestion-aware reward shaping. IEEE Access.

Sun, T., & Wang, M. (2025). Usage-Based and Personalized Insurance Enabled by AI and Telematics. Frontiers in Business and Finance, 2(02), 262-273.

Ren, S., & Chen, S. (2025). Large Language Models for Cybersecurity Intelligence, Threat Hunting, and Decision Support. Computer Life, 13(3), 39-47.

Chen, S., Liu, Y., Zhang, Q., Shao, Z., & Wang, Z. (2025). Multi-Distance Spatial-Temporal Graph Neural Network for Anomaly Detection in Blockchain Transactions. Advanced Intelligent Systems, 2400898.

Ge, Y., Wang, Y., Liu, J., & Wang, J. (2025). GAN-Enhanced Implied Volatility Surface Reconstruction for Option Pricing Error Mitigation. IEEE Access.

Wang, Y., Ding, G., Zeng, Z., & Yang, S. (2025). Causal-Aware Multimodal Transformer for Supply Chain Demand Forecasting: Integrating Text, Time Series, and Satellite Imagery. IEEE Access.

Yang, Y., Wang, M., Wang, J., Li, P., & Zhou, M. (2025). Multi-Agent Deep Reinforcement Learning for Integrated Demand Forecasting and Inventory Optimization in Sensor-Enabled Retail Supply Chains. Sensors (Basel, Switzerland), 25(8), 2428.

Chen, S., & Ren, S. (2025). AI-enabled Forecasting, Risk Assessment, and Strategic Decision Making in Finance. Frontiers in Business and Finance, 2(02), 274-295.

Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-Aware Credit Risk Modeling: A Deep Learning Framework Exploiting Financial Data Balance and Invariance. Symmetry (20738994), 17(3).

Jiang, B., Cao, J., Tan, Y., & Qiu, S. (2025). Deep Learning Architectures for Sequential Decision-Making in Financial Systems: From Fraud Detection to Risk Management. Journal of Banking and Financial Dynamics, 9(9), 1-11.

Wang, M., Zhang, X., Yang, Y., & Wang, J. (2025). Explainable Machine Learning in Risk Management: Balancing Accuracy and Interpretability. Journal of Financial Risk Management, 14(3), 185-198.

Zhang, S., Qiu, L., & Zhang, H. (2025). Edge cloud synergy models for ultra-low latency data processing in smart city iot networks. International Journal of Science, 12(10).

Yang, J., Zeng, Z., & Shen, Z. (2025). Neural-Symbolic Dual-Indexing Architectures for Scalable Retrieval-Augmented Generation. IEEE Access.

Sun, T., Wang, M., & Chen, J. (2025). Leveraging Machine Learning for Tax Fraud Detection and Risk Scoring in Corporate Filings. Asian Business Research Journal, 10(11), 1-13.