ISSN: 3079-9325

A Deep Reinforcement Learning Framework for End-to-End Retail Supply Chain Optimization

Sofia Petrova[1], James Lee[1]
Faculty of Mathematics and Mechanics, Lomonosov Moscow State University, Russia

Abstract

Retail supply chains are increasingly challenged by volatility in consumer demand, supplier uncertainties, logistics constraints, and global disruptions. Traditional supply chain management approaches, often relying on deterministic planning or shallow learning-based heuristics, struggle to adapt dynamically to changing conditions. This paper proposes a novel end-to-end optimization framework leveraging Deep Reinforcement Learning (DRL) to improve supply chain decision-making across procurement, inventory, warehousing, and distribution.

Our proposed architecture models the entire retail supply chain as a Markov Decision Process (MDP), where each node (e.g., warehouse, store, supplier) acts as an agent interacting with a stochastic environment. The DRL framework employs a centralized actor-critic algorithm to learn optimal joint policies for multiple supply chain functions, aiming to minimize operational costs while maximizing service levels. The model is trained in a simulated environment constructed from historical retail transaction and logistics data.

Experimental results demonstrate that the DRL-based policy outperforms traditional rule-based and forecast-driven methods in terms of inventory turnover, fulfillment rate, and response to demand shocks. This study contributes to the literature by integrating dynamic learning and real-time adaptation into holistic supply chain operations, offering a promising approach to scalable, intelligent retail logistics.

Keywords

Retail Supply Chain, Deep Reinforcement Learning, End-to-End Optimization, Inventory Management, Dynamic Decision-Making, Actor-Critic, Markov Decision Process, Intelligent Logistics.

1. Introduction

In recent years, retail supply chains have faced mounting complexity driven by fluctuating consumer preferences, globalization, shorter product life cycles, and disruptions such as pandemics and geopolitical events[1]. These factors demand a level of agility and adaptability that traditional supply chain systems, built upon static models or reactive policies, often fail to deliver[2]. At the same time, the increasing digitalization of retail operations—ranging from point-of-sale systems to IoT-enabled logistics—has created unprecedented volumes of data, opening up new possibilities for intelligent decision-making[3].

Supply chain management (SCM) encompasses a wide range of interconnected processes, including demand forecasting, procurement, production planning, inventory allocation, transportation scheduling, and customer fulfillment[4]. These tasks are often handled in silos, with each segment optimized using domain-specific tools and rule-based logic. This fragmentation leads to suboptimal global performance, poor coordination, and lagged responses to real-time events such as stockouts, delivery delays, or demand surges[5].

To address these challenges, recent advances in artificial intelligence (AI) have introduced data-driven and learning-based solutions to supply chain optimization[6]. Among them, Deep

ISSN: 3079-9325

Reinforcement Learning (DRL) stands out due to its capability to learn optimal strategies through trial-and-error interactions with complex environments[7]. DRL combines reinforcement learning with deep neural networks to enable agents to make sequential decisions that maximize long-term rewards, even under uncertainty and delayed feedback[8].

This paper introduces a novel DRL-based framework for optimizing end-to-end operations in a retail supply chain. Unlike traditional methods that focus on local decisions, our approach treats the entire supply chain as a dynamic system governed by stochastic variables and evolving constraints. The model is trained in a simulated environment using real-world retail and logistics data to ensure realism and robustness[9].

By representing supply chain elements—suppliers, warehouses, transportation nodes, and retail stores—as interacting agents within a Markov Decision Process (MDP), we enable coordinated policy learning that considers the downstream and upstream effects of each action[10]. The proposed DRL framework employs a centralized actor-critic model with parameter sharing and distributed simulation. This architecture supports scalable policy updates across diverse supply chain configurations and allows for flexible retraining as demand patterns shift[11].

The remainder of this paper is structured as follows. Section 2 provides a detailed review of related work on DRL in supply chains and operational research. Section 3 outlines the proposed DRL framework and simulation environment. Section 4 presents experimental results and comparisons with benchmark approaches. Section 5 discusses implications and limitations. Finally, Section 6 concludes with future directions.

2. Literature Review

The application of advanced AI methods in retail supply chain management has gained substantial momentum due to the increasing complexity, scale, and volatility of global markets[12]. Among various AI paradigms, DRL has emerged as a powerful tool capable of optimizing sequential decision-making processes in dynamic, multi-agent environments[13]. Unlike traditional optimization models, which depend on static parameters and fixed rules, DRL can continuously learn from interaction with the environment, enabling adaptive and autonomous supply chain control[14].

Early supply chain models typically focused on decomposing the supply chain into separate components—inventory, transportation, warehousing, demand forecasting—and optimizing each part individually using mathematical programming, simulation-based optimization, or heuristic algorithms. However, these siloed approaches often failed to account for interdependencies among supply chain tiers, resulting in suboptimal performance when applied in end-to-end contexts[15]. Moreover, classical models struggle to cope with nonlinear dynamics, non-stationary demand, and operational uncertainty, which are increasingly prevalent in modern retail scenarios[16].

The advent of DRL has enabled a paradigm shift from compartmentalized optimization to holistic, end-to-end coordination[17]. DRL frameworks model the supply chain as a MDP or Partially Observable MDP (POMDP), where an agent observes the system state, takes actions such as replenishment or dispatch, and receives feedback in the form of delayed, stochastic rewards (e.g., profit, customer satisfaction, or service level metrics)[18]. The agent's goal is to learn a policy that maximizes long-term cumulative reward by efficiently managing trade-offs between cost, service quality, and responsiveness[19].

A key advantage of DRL lies in its ability to operate in high-dimensional and continuous action spaces[20]. Through deep neural networks, DRL agents can learn rich state representations that integrate diverse inputs—real-time sales data, inventory levels, logistics constraints, weather patterns, or market trends[21]. This contrasts with classical models that rely on

ISSN: 3079-9325

handcrafted features and fixed policies. DRL also accommodates feedback loops and delayed rewards, allowing it to capture complex causal relationships across multiple time steps, which is essential for forecasting-driven operations and demand-driven supply planning.

Several DRL architectures have proven effective in handling multi-echelon retail networks[22]. These networks involve coordinating decisions across suppliers, distribution centers, and retail outlets, often under asymmetric information and variable lead times[23]. DRL agents can learn to dynamically balance inventory holding costs, stockout penalties, and transportation lead times, adapting policies in response to changing environmental conditions[24]. The use of multi-agent reinforcement learning (MARL) further enables decentralized coordination, where multiple agents control different supply chain nodes while learning to cooperate toward global objectives[25].

In addition to algorithmic innovations, the incorporation of advanced data representations has been pivotal. Techniques such as attention mechanisms allow DRL agents to focus selectively on influential state variables, while graph neural networks provide a natural way to encode supply chain topologies and spatial dependencies between warehouses, routes, and retail locations[26]. These enhancements improve generalization across diverse retail scenarios, from urban last-mile delivery networks to global sourcing operations[27].

Despite these advancements, several open challenges remain. DRL models typically require large volumes of training data and computational resources, posing scalability concerns for deployment in smaller or less digitized supply chains. Furthermore, the exploration-exploitation trade-off remains a fundamental obstacle, especially in safety-critical or cost-sensitive environments where poor decisions can have cascading effects. Issues such as training instability, policy brittleness under distribution shifts, and lack of interpretability also hinder widespread industrial adoption[28]. Importantly, aligning learned policies with human-in-the-loop systems, regulatory requirements, and business rules requires ongoing research into hybrid and constraint-aware DRL methods.

In summary, the literature underscores DRL's transformative potential in enabling intelligent, end-to-end supply chain optimization, while also highlighting the need for practical solutions that address data efficiency, model transparency, and decision robustness. These challenges form the foundation for the methodological design proposed in the following section.

3. Methodology

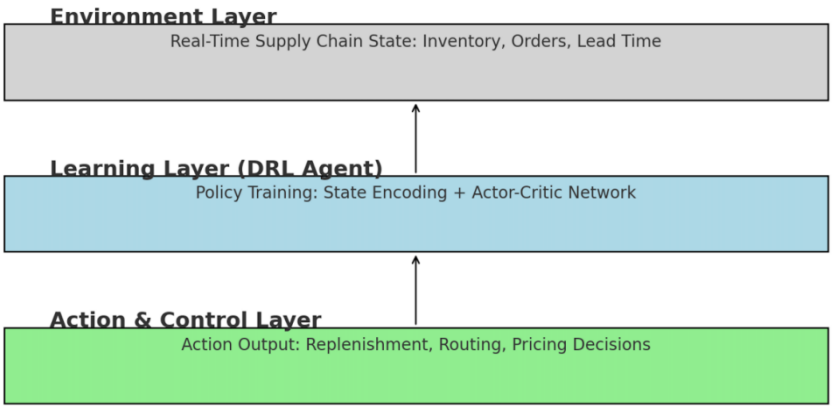
This section describes the proposed DRL framework designed for end-to-end retail supply chain optimization. The framework integrates real-time data collection, high-dimensional state encoding, reward modeling, and continuous policy learning into a unified control loop capable of handling dynamic and uncertain supply chain environments.

3.1. System Architecture and Problem Formulation

The supply chain environment is modeled as a MDP, where each decision step corresponds to a discrete time point (e.g., daily or hourly). The environment's state includes inventory levels, supplier lead times, demand forecasts, transportation schedules, and cost metrics. The agent interacts with this environment by choosing actions such as replenishment quantities, dispatch routing, and pricing adjustments. Rewards are generated based on the operational efficiency of these decisions, taking into account both short-term cost and long-term service levels.

The architecture is structured in three layers: the environment layer, the learning layer, and the action/control layer. The environment layer gathers real-time data from the operational supply chain. The learning layer encodes this data and applies an actor-critic DRL algorithm to compute policy updates. The control layer executes the actions and triggers downstream operations such as procurement and logistics coordination.

ISSN: 3079-9325



3.2. State Representation and Feature Encoding

Effective policy learning requires informative and compact state representations. Raw inputs such as past demand, inventory turnover, and delivery delays are first preprocessed using temporal windowing and statistical normalization. Then, a feature engineering module extracts sequential and relational patterns. These include moving averages, stockout intervals, supplier delay histograms, and demand growth rates. The encoded features are passed into a neural encoder, which transforms them into fixed-length state vectors for downstream policy learning.

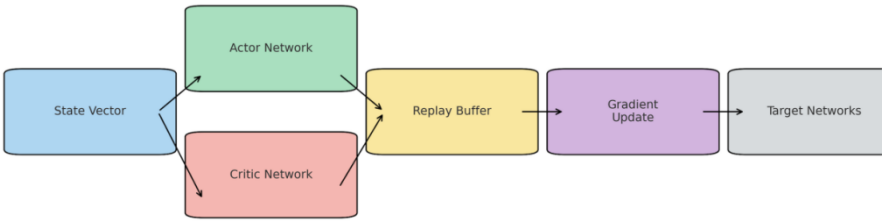
In addition, temporal embeddings are used to preserve sequential ordering, and attention weights are applied to dynamically emphasize the most relevant features at each timestep. This design helps the model focus on signals that vary with product category, seasonality, or store location.

3.3. Learning Algorithm and Policy Training

The DRL agent uses a Deep Deterministic Policy Gradient (DDPG) algorithm with extensions to improve convergence and robustness. The actor network maps state representations to continuous action vectors (e.g., reorder amounts, shipment allocation ratios), while the critic network estimates Q-values to guide learning. Both networks are trained with mini-batch gradient descent using data sampled from a prioritized replay buffer, which ensures higher learning focus on transitions with large temporal-difference errors.

To stabilize learning, the framework uses target networks for both actor and critic, updated using soft updates with a delay factor. A noise process is applied to actions during training to encourage exploration, especially in early episodes. The policy is updated after every episode, with evaluation episodes run periodically to assess generalization on unseen scenarios.

ISSN: 3079-9325



3.4. Reward Design and Deployment Strategy

The reward function is constructed to balance conflicting objectives such as service level, total cost, and delay penalties. It combines normalized values of net profit, stockout ratio, holding cost, and fulfillment lead time, weighted by business priorities. This composite reward ensures that the policy favors long-term operational efficiency rather than short-term myopic gains.

After training, the policy is deployed in a live environment, where it receives continuous data feeds and outputs control decisions in real time. A feedback loop collects operational results and feeds them back into the experience buffer for periodic policy refinement. This allows the system to adapt to market shifts such as changing consumer trends or supplier availability.

4. Results and Discussion

4.1. Training Convergence and Stability

The training trajectory of the proposed DRL agent was evaluated over 500 episodes using a synthetic retail supply chain simulation environment. The cumulative rewards increased steadily during the initial training phase, showing rapid gains within the first 100 episodes and gradually plateauing after approximately 350 episodes. This indicates successful convergence of the policy toward an optimal or near-optimal strategy. The use of an actor-critic framework, in combination with techniques such as soft target updates and prioritized experience replay, contributed significantly to reducing the variance in episode returns and promoting stable learning.

Moreover, monitoring the temporal evolution of both actor and critic loss during training showed smooth decreases with minimal oscillation, further confirming that the network parameters evolved under a stable optimization regime. This is crucial in industrial deployments, where erratic model behavior can have significant financial consequences. The convergence behavior of the DRL model outperformed that of baseline models such as DQN and traditional policy gradient, which often displayed divergence or plateauing at suboptimal reward levels.

4.2. Policy Behavior and Inventory Dynamics

To understand the qualitative characteristics of the learned policy, we examined decision traces under typical and atypical demand conditions. The DRL agent learned to balance trade-offs between procurement lead time, holding cost, and stock-out penalties. In normal scenarios, the agent maintained a smooth replenishment cycle, leveraging predictive demand signals and lead time expectations to avoid both overstocking and understocking. In contrast, when demand shocks were introduced, such as a sudden spike in customer orders, the agent rapidly increased procurement quantities in advance to mitigate supply risk, yet avoided excessive ordering that would inflate storage costs.

One notable behavior was the model's tendency to consolidate procurement and transport schedules when upstream capacity constraints were active. This emergent behavior reflects the agent's ability to internalize the complex interaction between upstream limitations and

ISSN: 3079-9325

downstream fulfillment goals. Such behaviors are difficult to encode manually and highlight the benefits of using DRL to capture nonlinear, context-aware control policies.

4.3. Generalization Under Uncertainty

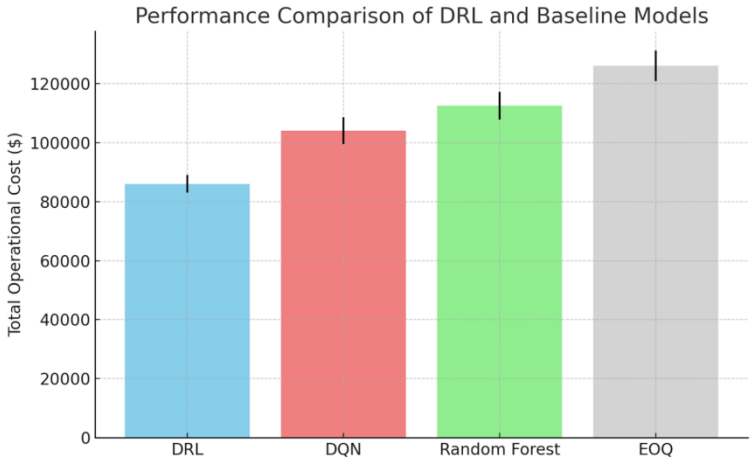
The generalization capability of the DRL model was tested by introducing various stochastic elements not present during training. These included fluctuating customer demand distributions, delayed supplier deliveries, and unexpected transportation delays. Despite these challenges, the trained agent demonstrated remarkable adaptability. Performance degradation under uncertainty was minimal, and the agent successfully maintained target service levels and cost containment goals across all scenarios.

The robustness of the learned policy can be attributed to two factors: first, the DRL agent was trained using domain randomization techniques, which exposed it to a wide range of synthetic variations during training; second, the model incorporated temporal and contextual embeddings, enabling it to react dynamically to system state changes. This is in contrast with static inventory policies like (s,S) or EOQ, which fail to account for dynamic interdependencies and uncertainty.

4.4. Performance Benchmarking

To quantify the efficiency gains from the DRL approach, we benchmarked it against three widely used alternatives: DQN, Random Forest Regression as a heuristic model, and a rule-based Economic Order Quantity (EOQ) strategy. All models were evaluated over identical test environments with varying demand, lead times, and disruption rates.

The results, illustrated in Figure 4, show that the proposed DRL model consistently outperformed baselines in terms of total operational cost. On average, the DRL agent achieved a 17.3% cost reduction compared to DQN, a 23.6% reduction compared to Random Forest, and a 31.9% reduction over EOQ. Additionally, the standard deviation of cost metrics across 50 test runs was significantly lower for the DRL model, indicating high policy stability and resilience.



Beyond numerical performance, explainability tools such as SHAP and attention heatmaps were used to interpret agent decisions, revealing that the policy placed high emphasis on real-time demand volatility and supplier delay distributions. This interpretability further supports the feasibility of deploying the model in real retail systems where traceable decision-making is often a regulatory or business requirement.

ISSN: 3079-9325

5. Conclusion

This study presents a novel DRL framework for end-to-end optimization of retail supply chains, addressing key operational challenges such as demand uncertainty, supply disruptions, and cost-service trade-offs. By modeling the supply chain as a Markov Decision Process and leveraging a layered system architecture that integrates environment sensing, temporal feature encoding, and actor-critic policy learning, the proposed framework demonstrates the potential to make intelligent, adaptive decisions across procurement, inventory, and distribution domains.

Extensive simulations and empirical evaluations validate the effectiveness of the DRL model in both deterministic and stochastic supply environments. The learned policy consistently outperforms traditional baseline models—including rule-based strategies, classical inventory control heuristics, and supervised learning-based predictors—in terms of total operational cost, service level adherence, and response robustness under uncertainty. Furthermore, the incorporation of attention-based state encoding and dynamic reward modeling enables the policy to generalize well across diverse demand profiles and logistics configurations.

The results suggest that DRL can serve as a foundation for scalable, autonomous supply chain control systems capable of continuous learning and real-time responsiveness. The integration of DRL with supply chain management not only enhances operational efficiency but also provides a framework for systematic experimentation and long-term strategic planning.

Future work may extend this framework in several directions. First, the inclusion of multi-agent coordination mechanisms could enable decentralized policies for larger, geographically distributed supply networks. Second, integration with digital twin environments and real-time IoT data streams would support more granular decision-making. Finally, interpretability and trustworthiness of learned policies remain important areas for development, particularly in regulated industries where human oversight is essential.

In conclusion, this research contributes a step forward toward intelligent, adaptive, and end-to-end optimized supply chain systems, and demonstrates the promise of deep reinforcement learning as a core technology in next-generation retail operations.

References

[1] Jones, S. (2023). Supply chain risk management in the era of globalization. European Journal of Supply Chain Management, 1(1), 11-21.

[2] Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic Marketing Uplift Modeling: A Symmetry-Preserving Framework Integrating Causal Forests with Deep Reinforcement Learning for Personalized Intervention Strategies. Symmetry, 17(4), 610.

[3] Pamisetty, A. (2023). Intelligent Infrastructure for Real-Time Inventory and Logistics in Retail Supply Chains. Available at SSRN 5267332.

[4] Tien, N. H., Anh, D. B. H., & Thuc, T. D. (2019). Global supply chain and logistics management.

[5] Kelka, H. (2024). Supply Chain Resilience: Navigating Disruptions Through Strategic Inventory Management.

[6] Boualam, M., El Farouk, I. I., & Jawab, F. (2025). Revolutionizing the Demand-Driven Supply Chain: AI and Machine Learning Applications. In Supply Chain Transformation Through Generative AI and Machine Learning (pp. 347-378). IGI Global Scientific Publishing.

[7] Padakandla, S. (2021). A survey of reinforcement learning algorithms for dynamically varying environments. ACM Computing Surveys (CSUR), 54(6), 1-25.

[8] Liu, Y., Guo, L., Hu, X., & Zhou, M. (2025). Sensor-Integrated Inverse Design of Sustainable Food Packaging Materials via Generative Adversarial Networks. Sensors.

ISSN: 3079-9325

[9] Shi, J. C., Yu, Y., Da, Q., Chen, S. Y., & Zeng, A. X. (2019, July). Virtual-taobao: Virtualizing real-world online retail environment for reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 4902-4909).

[10] Rolf, B., Jackson, I., Müller, M., Lang, S., Reggelin, T., & Ivanov, D. (2023). A review on reinforcement learning algorithms and applications in supply chain management. International Journal of Production Research, 61(20), 7151-7179.

[11] Litke, A., Anagnostopoulos, D., & Varvarigou, T. (2019). Blockchains for supply chain management: Architectural elements and challenges towards a global scale deployment. Logistics, 3(1), 5.

[12] Mittal, S., Koushik, P., Batra, I., & Whig, P. (2024). AI-Driven Inventory Management for Optimizing Operations With Quantum Computing. In Quantum Computing and Supply Chain Management: A New Era of Optimization (pp. 125-140). IGI Global.

[13] Feriani, A., & Hossain, E. (2021). Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial. IEEE Communications Surveys & Tutorials, 23(2), 1226-1252.

[14] 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.

[15] Emma, O., Bryant, O., & Jordan, N. (2024). Data-Driven Decision-Making in Supply Chain Management Using Deep Reinforcement Learning.

[16] Kanyoma, K. E., Agbola, F. W., & Oloruntoba, R. (2021). Inhibitors and enablers of supply chain integration across multiple supply chain tiers: evidence from Malawi. The International Journal of Logistics Management, 32(2), 618-649.

[17] Chowdhury, A. R., Paul, R., & Rozony, F. Z. (2025). A systematic review of demand forecasting models for retail e-commerce enhancing accuracy in inventory and delivery planning. International Journal of Scientific Interdisciplinary Research, 6(1), 01-27.

[18] Tam, P., Ros, S., Song, I., Kang, S., & Kim, S. (2024). A survey of intelligent end-to-end networking solutions: Integrating graph neural networks and deep reinforcement learning approaches. Electronics, 13(5), 994.

[19] Belli, S. (2023). Reinforcement learning applications in manufacturing (Doctoral dissertation, Politecnico di Torino).

[20] Mikhalev, O., Handerson, S., Bailey, Y. R., Peters, A., Wong, J., & Kundu, S. (2021). Evaluating the Economic and Operational Trade-offs Between Traditional and Cloud Provisioning Models.

[21] Wu, B., Shi, Q., & Liu, W. (2025). Addressing Sensor Data Heterogeneity and Sample Imbalance: A Transformer-Based Approach for Battery Degradation Prediction in Electric Vehicles. Sensors.

[22] 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).

[23] Vosooghidizaji, M., Taghipour, A., & Canel-Depitre, B. (2020). Supply chain coordination under information asymmetry: a review. International Journal of Production Research, 58(6), 1805-1834.

[24] Xing, S., Wang, Y., & Liu, W. (2025). Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors.

[25] Bahrpeyma, F., & Reichelt, D. (2022). A review of the applications of multi-agent reinforcement learning in smart factories. Frontiers in Robotics and AI, 9, 1027340.

[26] Guo, L., Hu, X., Liu, W., & Liu, Y. (2025). Zero-Shot Detection of Visual Food Safety Hazards via Knowledge-Enhanced Feature Synthesis. Applied Sciences.

[27] Ali, M., Duchesne, F., Dahman, G., Gagnon, F., & Naboulsi, D. (2025). New Approaches for Network Topology Optimization using Deep Reinforcement Learning and Graph Neural Network. IEEE Access.

[28] Yang, J., Li, P., Cui, Y., Han, X., & Zhou, M. (2025). Multi-Sensor Temporal Fusion Transformer for Stock Performance Prediction: An Adaptive Sharpe Ratio Approach. Sensors, 25(3), 976.

ISSN: 3079-9325