Towards Efficient and Sustainable Operations in Supply Chain Management: A Reinforcement Learning Approach

Djonathan Quadras[1], Romeo Bandinelli[1], Virginia Fani[1,2]

[1]Department of Industrial Engineering, University of Florence, e-mail: djonathan.quadras@unifi.it, romeo.bandinelli@unifi.it
[2] Department of Engineering and Sciences, Universitas Mercatorum, e-mail: virginia.fani@unimercatorum.it

KEYWORDS

Supply Chain Management, Reinforcement Learning, Inventory Control, Sustainable Operations, Artificial Intelligence.

ABSTRACT

Modern supply chain management requires balancing economic performance, operational efficiency, and environmental sustainability among dynamic demand uncertainties. Traditional analytical models often fail to capture the stochastic nature of these complex, multi-echelon networks, driving the need for advanced data-driven methodologies. This study investigates the application of Reinforcement Learning to address the gap in jointly optimizing procurement, production, and replenishment decisions while explicitly minimizing carbon emissions and maintaining high service levels. A simulation-based approach was developed to model a three-echelon juice supply chain, comprising two suppliers, one manufacturing factory, and five regional warehouses facing stochastic seasonal demand. Two algorithms, Proximal Policy Optimization and Advantage Actor-Critic, were trained and evaluated within this continuous, multi-objective environment. The computational results reveal a significant divergence in policy stability and performance. The Proximal Policy Optimization algorithm demonstrated superiority, converging to a low-variance strategy that generated approximately 35% more profit and reduced total accumulated operational costs by 33%. Furthermore, it successfully optimized the distribution of goods, achieving a superior customer fulfilment rate while simultaneously operating with lower inventory levels. On the other hand, the Advantage Actor-Critic agent achieved lower carbon emissions.

INTRODUCTION

Supply Chain Management (SCM) has evolved into a complex and highly interconnected discipline in which operational efficiency, economic performance, and environmental sustainability must be addressed simultaneously. Modern Supply Chains (SCs) are typically structured as multi-echelon networks composed of suppliers, manufacturers, warehouses, and retailers that coordinate procurement, production, transportation, and distribution activities to meet customer demand while maintaining service levels and minimizing costs (Hu et al., 2025; Zhou et al., 2024). Within this context, inventory management plays a central role, as firms must carefully balance holding costs against the risk of stockouts and disruptions (Kou et al., 2025; Rizqi and Chou, 2024). However, the environmental footprint associated with these operations has become a shared concern among researchers in literature. Transportation procedures contribute significantly to greenhouse gas emissions. Furthermore, production processes and end-of-life handling of products strongly influence the total environmental impact across the SC lifecycle (Fani et al., 2025). As organizations face regulatory pressures and societal expectations to reduce emissions, managing SCs in a sustainable manner has become a strategic priority (Wang et al., 2025).

At the same time, the operational environment of SCs has grown more uncertain and dynamic. Demand volatility, variable lead times, disruptions, and complex interdependencies across echelons make decision-making increasingly challenging (Deshpande et al., 2025). Traditional analytical models and heuristics, such as deterministic inventory policies, often rely on simplified assumptions that fail to capture the stochastic and nonlinear nature of real systems, leading to sub-optimal performance in practice (Deshpande et al., 2025; Shar et al., 2022). The increasing dimensionality of decision variables in large networks also makes classical optimization approaches computationally expensive or intractable (Abdellahi et al., 2025). To address the limitations of traditional analytical models, data-driven simulation and optimization methodologies have been widely recognized as effective tools for resolving complex production scheduling and finite capacity constraints within multi-level SCs (Bandinelli et al., 2022). It facilitated the integration of Artificial Intelligence (AI) into SCM, using data-driven methods to learn adaptive policies capable of responding to complex and changing environments.

Among AI paradigms, Reinforcement Learning (RL) has emerged as a particularly promising approach for sequential decision-making problems (Quadras et al., 2024). RL is characterized by an agent that interacts with an environment, observes system states, takes actions, and improves its strategy through feedback with the objective of maximizing long-term rewards (Chen et al.,

Communications of the ECMS, Volume 40, Issue 1,
Proceedings, ©ECMS Filippo Sanfilippo, Florenc Demrozi, Fabio Sgarbossa, Mohammad Poursina (Editors) 2026 ISSN 2522-2414
ISBN: 978-3-937436-90-6 / 978-3-937436-89-0(CD)

2025; Wang et al., 2025). Unlike supervised learning, it does not rely on labelled data but instead acquires knowledge through experience, making it well suited for dynamic and uncertain domains (Rizqi and Chou, 2024). When combined with deep learning, RL can handle high-dimensional state spaces and capture complex patterns, enabling effective learning in environments where system dynamics are partially unknown or constantly evolving (Abdellahi et al., 2025; Hu et al., 2025).

The suitability of RL for SC optimization lies in its capacity to manage sequential and interdependent decisions across multiple echelons. Decisions taken at one stage influence downstream performance, often generating phenomena such as the bullwhip effect and inefficient resource allocation (Mutyala and Kaur, 2024). RL can learn coordinated policies that stabilize inventory levels, reduce total system costs, and maintain service performance while adapting to stochastic demand and lead time variability (Mao et al., 2023; Shar et al., 2022). Moreover, recent studies indicate that AI-driven strategies can dynamically balance economic objectives with environmental considerations, supporting emission reduction while preserving operational efficiency (Wang et al., 2025).

Despite the growing body of literature on AI applications in SCM, the joint optimization of environmental and operational objectives remains a developing research area. In particular, the integration of RL to simultaneously reduce emissions and holding costs while preserving service levels and integrating the decision for procurement, production and distribution has not yet been fully explored in structured multi-echelon environments. To address this gap, this paper investigates the application of two RL algorithms to manage inventory and replenishment decisions in a SC context. The objective is to develop adaptive policies that minimize carbon emissions and inventory holding costs while maintaining a satisfactory service level.

This paper is structured as follows: Section 2 presents the theoretical background to present correlated works that implemented RL into SCM. Section 3 provides the methodological framework, simulation environment, and the algorithm configurations. Section 4 provides the training and testing results. Section 5 discusses the results and compare the algorithms. In the end, section 6 concludes the paper and provides suggestions for future works.

THEORETICAL BACKGROUND

RL is a computational approach to learning from interaction, focused on goal-directed learning from sequences of states, actions, and rewards (Sutton and Barto, 1998). RL has increasingly been adopted as a decision-support paradigm for SCM due to its capacity to model sequential decision-making under uncertainty. In a typical RL formulation, presented in Figure 1, an agent (algorithm) interacts with an environment (the SC) described by a set of states, selects actions, and receives feedback through a reward signal that reflects system performance (Quadras et al., 2024).

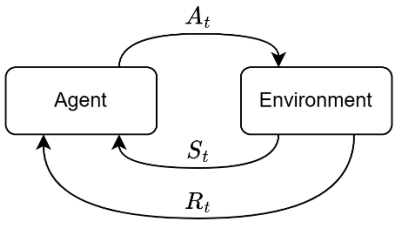

Figure 1 – RL Framework

The application of RL to SCM has progressively evolved from single-node inventory control to complex, multi-echelon, multi-objective, and sustainability-aware systems. Early adaptive inventory models such as those proposed by Yang and Jihui (2015) positioned the replenishment problem as a sequential decision process under stochastic demand. In their centralized and decentralized configurations, the state was primarily characterized by the current demand pattern at each decision epoch, while the action space consisted of selecting a safety factor that determined safety stock and lead time. The reward structure implicitly targeted cost minimization through the adjustment of inventory buffers. In the following works the state representation expanded beyond demand signals to encompass richer operational descriptors. For example, the approaches applied by Shar et al. (2022) and Deshpande et al. (2025) modelled the environment through detailed inventory positions, in-transit orders, supplier lead times, and demand trends. This shift toward multi-objective optimization is further exemplified in the fashion industry, where frameworks now combine mathematical programming with discrete-event simulation to balance conflicting targets like cost, delays, and processing times (Fani et al., 2017). The action space shifted toward discrete replenishment quantities, and the reward was explicitly defined as total cost, typically combining holding, ordering, and stockout penalties. The decision process became more granular, moving from adjusting safety factors to determining explicit order quantities at each period, thereby embedding tactical inventory control directly within the agent's policy.

The complexity further increased with the introduction of multi-echelon and multi-agent formulations. Zhou et al. (2024) proposed EM-VDTD3, integrating value decomposition and modified experience replay into a twin-delayed actor–critic framework. Each agent managed a subsystem within a large-scale network, sharing parameters while processing heterogeneous subsystem characteristics. The state space captured on-hand inventory, in-transit stock, and demand rates for each echelon. Actions corresponded to replenishment quantities, and the reward aggregated multiple cost components including holding, transshipment, emergency ordering, and lost production costs. Similarly, Hu et al. (2025) adopted a decentralized multi-agent Proximal Policy Optimization (PPO) approach. In this configuration, each participant operated with local observations such as on-hand inventory, backlogs, arriving shipments, and on-order quantities, while

collectively minimizing total system cost. The transition from centralized state visibility to decentralized local observations marks a conceptual shift in RL applications: the state becomes partial and agent-specific, actions remain replenishment decisions, and rewards reflect joint performance. Learning thus navigates the tension between autonomy and coordination.

More recent contributions embed prediction and procurement decisions into the loop. Chen et al. (2025) integrated demand forecasting and decision-making within a centralized training and decentralized execution paradigm. The state space incorporated historical sequences, neighbouring information, carbon price signals, congestion rates, weather risks, and strike probabilities. Actions extended beyond replenishment to include transportation route adjustments, while rewards reflected total cost under environmental and operational uncertainty. Compared to inventory-focused models, the state evolved into a multidimensional representation of SC risk and environmental signals, and the action space incorporated both material and logistical levers. Rizqi and Chou (2024) introduced uncertain supplier discounts and delivery options, representing the state through inventory positions and discount events, and allowing actions that jointly determined order quantities and transport modes. In this approach, RL explore trade-offs between cost and lead time, transforming the agent from a replenishment planner into a tactical procurement strategist.

Sustainability-focused studies further internalized environmental externalities into the reward mechanism. Wang et al. (2025) modelled a closed-loop SC under carbon tax policy using PPO. The state included current and in-transit inventories, recyclable quantities, demand histories, and temporal information. Actions encompassed collection planning, manufacturing, recycling decisions, and external procurement. The reward incorporated carbon tax costs alongside traditional economic terms, effectively reshaping the objective from pure cost minimization to eco-efficient optimization. Additional dimensions such as perishability and waste were incorporated by Kou et al. (2025), where the state explicitly tracked residual lifetimes of goods, and rewards penalized disposal costs for expired products.

Rl approaches may also be integrated with a diverse set of tools and algorithms to increase security, transparency, and compliance. Ma et al. (2024) combined Soft Actor–Critic learning with blockchain-based verification mechanisms. Kumar and Khedlekar (2025) further integrated emission constraints, perishability, and traceability into a hybrid framework, demonstrating how RL can collaborate with predictive filtering, genetic optimization, and decentralized validation mechanisms.

Across these contributions, a trend emerges. The state set has expanded from demand and inventory observations to high-dimensional, multimodal, and environmentally focused signals. Actions have evolved from selecting safety factors to determining replenishment quantities, transportation routes, collection plans, procurement modes, and production levels. Reward functions have shifted from simple cost formulations toward multi-objective formulations integrating service level, carbon taxation, perishability losses, and profit maximization. However, current literature still focus on isolated decision domains, such as inventory control, procurement choices, or production planning, rather than addressing them as a fully integrated system. No studies were identified that propose a unified framework capable of jointly optimizing procurement, production, and replenishment decisions with an explicit focus on sustainability objectives while simultaneously maintaining service levels. Although some works incorporate environmental costs or carbon-related signals into the reward structure, and others expand the action space to include procurement or logistics decisions, these dimensions are not considered together.

MODEL DESCRIPTION AND CONFIGURATION

This study adopts a simulation-based approach to evaluate two RL algorithms in an integrated procurement, production, and replenishment problem. The simulation environment was developed entirely in Python as a custom event-driven model utilising the SimPy library. To enable the integration of RL algorithms, the SC environment was wrapped using the Gymnasium framework. This approach provided a standardised interface for the agents to interact with the simulated multi-echelon network. The simulated system represents a three-echelon juice SC composed of two suppliers, one dedicated factory, and five warehouses located in real Italian regions. The network manages four finished products and five raw materials, four for juice production and one for packaging. Customer demand is generated at the warehouse level on a weekly basis according to a different Poisson distribution at each location and affected by summer-winter seasonality.

At each decision period, warehouses place replenishment orders to the factory. The factory simultaneously determines production quantities for each product and procurement quantities of raw materials from suppliers. Suppliers are assumed to have unlimited capacity, purchasing costs are ignored, and transportation is modelled with unlimited vehicle capacity and an average speed of 80 km/h. Physical feasibility constraints are enforced: negative inventories are not allowed, production requires available raw materials, and replenishment orders can only be fulfilled if finished goods are available at the factory.

The environment evolves in discrete weekly steps. The state and action spaces were explicitly defined to capture the complex dynamics of the network. The observation space consists of 61 continuous variables. These include 10 variables tracking raw material inventory and transit, 4 for factory product inventory, 40 for warehouse inventory and transit, 5 monitoring recent lost sales per warehouse, and 2 variables for time embedding (the sine and cosine of the day of the year to model seasonality). The action space is a continuous vector comprising 58

dimensions, normalised between -1.0 and 1.0. This vector dictates 10 variables for raw material decisions (reorder points and order quantities), 8 variables for production decisions (reorder points and batch sizes), and 40 variables for warehouse decisions (reorder points and order quantities across all locations).

The reward function is mathematically formulated to penalise the total cost at each step. Specifically, it is calculated as the negative total cost divided by a scaling factor of 1,000,000. This scaling factor is a necessary technique applied to prevent numerical instability and exploding gradients during neural network training, as raw financial costs within the simulation can grow excessively large. The total cost comprises the sum of holding costs, shortage penalties, transport freight costs, and carbon costs. Precise financial parameters were established to weight the agent's behaviour: holding costs are 0.01 per unit per day for raw materials and 0.50 for finished products. To prioritise service levels, the shortage penalty is weighted heavily at ten times the product's sale price. Furthermore, carbon emissions are penalised at a rate of 0.50 per kilogram of CO_2, while freight transport costs 1.5/km.

Two policy-gradient algorithms are implemented: Advantage Actor–Critic (A2C) and PPO. Both algorithms utilise a standard Multi-Layer Perceptron policy architecture featuring two hidden layers of 128 neurons. The network employs the hyperbolic tangent (Tanh) non-linearity activation function to ensure bounded outputs suitable for the continuous action space. A2C is configured with a learning rate of 0.0005, 20 rollout steps, discount factor of 0.99, generalized advantage estimation with $\lambda = 0.95$, and symmetric actor–critic networks with two hidden layers of 128 neurons. PPO is configured with a learning rate of 0.0003, 2048 rollout steps, batch size of 64, and discount factor of 0.99.

Training is conducted over multiple episodes representing complete operational horizons under stochastic demand. After convergence, the learned policies are evaluated through independent simulation runs. Performance is assessed using total cost, emission levels, and achieved service level, enabling a direct comparison of how each algorithm balances economic and environmental objectives within an integrated decision framework.

RESULTS

This section presents the training and testing results applying the PPO and A2C algorithms. Figure 2 shows the total cost evolution. It illustrates a pronounced learning curve for both algorithms. During the initial training phase, both agents exhibit a fast decline, which suggests the successful early adaptation of their respective policies to minimize operational inefficiencies and mitigate penalties. As the training progresses into the stabilization phase, a distinct bifurcation in performance becomes evident. The PPO agent establishes a smooth, stable trajectory, converging to a visibly lower baseline cost while exhibiting minimal variance. On the other hand, even though the A2C agent also reaches a steady-state region, it stabilizes at a higher cost level and is characterized by volatility.

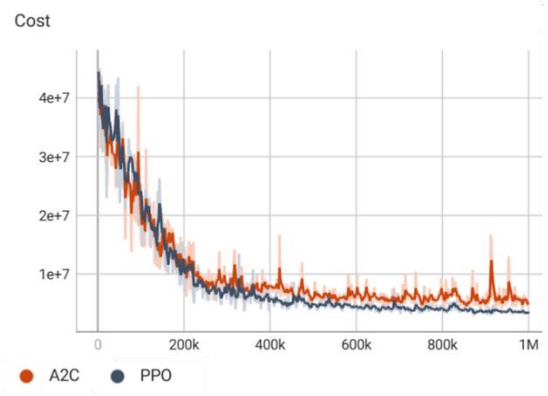

Figure 2 – Evolution of Cost During Training

Figure 3 presents the inventory level evolution. Initially, both the A2C and PPO algorithms exhibit a synchronized upward trajectory, rapidly accumulating stock as they learn to establish safety buffers against unfulfilled demand. As the training advances, their strategic behaviours visibly diverge. The PPO agent transitions into a stable level, maintaining a steady inventory with consistent fluctuations. In contrast, the A2C agent struggles to achieve a similar steady state, displaying pronounced volatility and chaotic variance. Its trajectory is marked by aggressive spikes and sharp depletions.

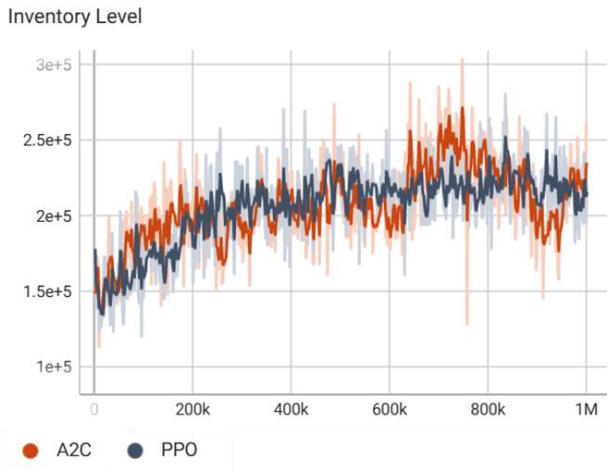

Figure 3 – Evolution of Inventory Level During Training

The comparative analysis of the carbon emission profiles is presented in Figure 4. Both algorithms show an initial fast increase in emissions. However, the PPO agent establishes a trajectory characterized by a significantly higher environmental footprint, ascending steeply to a prolonged level. In contrast, the A2C agent maintains a consistently lower emission profile throughout the entire training duration.

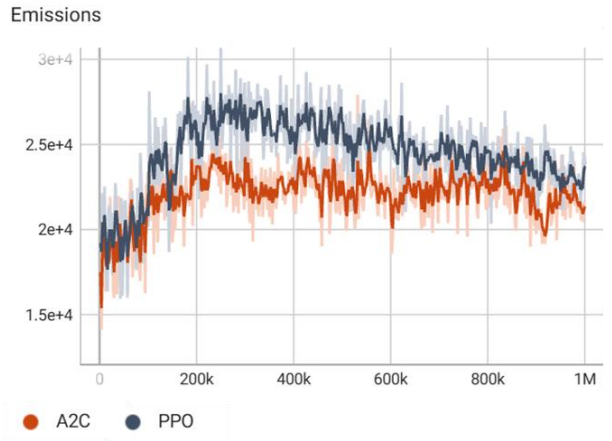

Figure 4 – Evolution of Emissions During Training

The analysis of the lost sales trajectory is presented in Figure 5. It illustrates the direct consequence of the inventory strategies adopted by each agent. Both algorithms demonstrate a rapid initial improvement, characterized by an exponential decay in unfulfilled orders. The PPO agent achieves a highly efficient steady state, driving lost sales down and maintaining this level with stability. On the other hand, the A2C agent converges to underperforming baseline. Its trajectory also presents volatility, indicating that its reactive inventory policy lacks the buffer necessary to absorb stochastic demand.

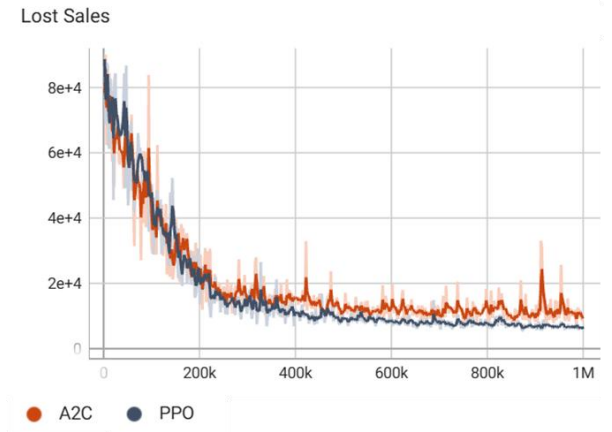

Figure 5 – Evolution of Lost Sales During Training

After training the agents, a one-year simulation was conducted to test the model's efficiency. The results suggest a superiority of the PPO model. Financially, the PPO agent demonstrated better efficiency, generating a total profit about 35% superior compared to A2C. This profitability is driven by PPO's ability to reduce inefficiencies, achieving a performance about 33% better in the total accumulated costs. In the same direction, PPO proved to be more efficient in using resources, with reduction of about 7% in total holding inventory achieving a better service level of 97.13%, compared to A2C's 95.58%. The only metric where A2C achieved a better result is related to the CO_2 emissions, with a value about 13% lower when compared to PPO.

Table 1: Testing results for one year

	PPO	A2C
Total Cost ($\times 10^6$)	2,89	4,32
Profit ($\times 10^6$)	6,29	4,67
Service Level	97,13%	95,58%
Emissions ($\times 10^3 kg\ CO_2$)	21,28	18,42
Average Stock (k)	196,23	210,30

DISCUSSION

The analysis of the PPO and A2C algorithms suggest a divergence in their ability to formulate stable and efficient SC control policies. While both agents successfully demonstrate an initial learning phase, characterized by a fast adaptation to minimize operational inefficiencies, their steady-state behaviours show an interesting contrast. The PPO algorithm proves to converge to a stable trajectory that establishes a lower baseline cost with minimal variance. On the other hand, the A2C agent struggles to transcend its reactive learning phase, stabilizing at a higher cost level marked by continuous volatility.

This divergence in policy stability is most clearly reflected in the agents' respective inventory and service level management. Both models initially accumulate stock to build safety buffers to attend the demand, but the PPO agent transitions into a constrained level, maintaining a steady inventory level with regulated fluctuations. This stable replenishment strategy directly translates into superior customer fulfilment, allowing PPO to drive lost sales down to an efficient result. Conversely, the A2C agent presents increased variance in its inventory levels, characterized by aggressive increase and decreases. This reactive inventory policy lacks the resilient buffer necessary to absorb demand shocks, causing the A2C model to converge to an underperforming baseline.

The single dimension where the A2C agent appears to visually outperform the PPO model is in its carbon emission profile. It maintains a consistently lower environmental footprint throughout training and achieving approximately 13% fewer emissions during the testing phase. However, when contextualized within the broader operational metrics, this reduction is not indicative of advanced environmental optimization but rather points to logistical stagnancy. The PPO agent's higher emission trajectory is a direct consequence of the active distribution required to reliably satisfy market demand. The A2C model's lower transport activity is intrinsically linked to its inability to adequately replenish stock and fulfil orders, highlighting a trade-off where the A2C's apparent environmental efficiency comes at the expense of operational reliability.

The one-year testing simulation proves the superiority of the PPO model over the A2C. In financial terms, the PPO agent generated approximately 35% more profit, driven by a 33% reduction in total costs. Furthermore, PPO demonstrates a sophisticated capability for resource optimization, as it managed to secure a superior service level of 97.13% while simultaneously operating with 7% less total average holding inventory. This ability to achieve higher fulfilment rates with leaner average stock levels confirms that the PPO algorithm successfully optimized the spatiotemporal distribution of goods. It is important to highlight that this is a considerably difficult challenge, as the factory must order the raw materials at time t, to produce at time $t + 1$ in answer to the expected demand at $t + 2$. Thus, PPO is suggested to be able to provide a viable and resilient solution for complex SC deployment.

CONCLUSIONS

This study investigated the application of RL algorithms to jointly optimize procurement, production, and replenishment decisions within a multi-echelon SC. A simulation-based approach was developed for a three-echelon juice SC facing stochastic seasonal demand. The comprehensive evaluation of these algorithms demonstrated that the PPO model significantly outperformed the A2C baseline across different metrics. While both agents successfully learned to mitigate early operational inefficiencies, PPO converged to a highly stable policy, reducing total costs and achieving a superior customer service level while maintaining leaner average stock levels, even though achieving higher environmental impacts to meet the demand.

From an academic perspective, this research bridges a gap in the existing literature by providing a unified RL framework that jointly optimizes procurement, production, and replenishment decisions while explicitly incorporating sustainability objectives. It advances the theoretical understanding of how different algorithms handle high-dimensional and stochastic environments. For industry, the study offers a data-driven tool to face complex trade-offs between holding costs, service levels, and carbon emissions. The findings provide managers with actionable insights into deploying autonomous decision-support systems that enhance both economic resilience and operational agility to face demand volatility.

Despite the results, this study is subject to limitations that must be acknowledged. The simulation environment relies on specific structural simplifications, such as assuming unlimited supplier capacity and unconstrained vehicle transportation capacity. Additionally, the economic model excludes raw material purchasing costs and possible discounts based on the amount purchased, focusing strictly on holding costs, shortage penalties, and transportation-related emissions. The scope of the SC is also localized to a specific three-echelon juice network with five regional warehouses, meaning the learned policies and algorithmic performance dynamics might require substantial recalibration if applied to vastly different industries, highly perishable goods with strict expiration constraints, or more extensive global networks facing international trade disruptions.

Future research should incorporate more realistic logistical limitations, such as finite vehicle capacities, constrained supplier production rates, and dynamic purchasing costs. Expanding the state and action spaces to include multi-modal transportation options or dynamic routing could provide a deeper exploration of the trade-off between operational efficiency and carbon emissions. Furthermore, investigating decentralized multi-agent RL architectures, where different echelons or facilities operate as independent learning agents with partial observability. Also, future research should focus on developing tools to support the human interaction with the system to build a human-based decision system. Finally, future models should integrate SC disruptions and stochastic lead times to test the resilience and adaptability of PPO-driven policies.

REFERENCES

Abdellahi, L.S.A., Zoubeir, Z., Mohamed, Y., Haouba, A., Hmetty, S., 2025. Deep Reinforcement Learning for Optimal Replenishment in Stochastic Assembly Systems. Mathematics 13. https://doi.org/10.3390/math13142229

Bandinelli, R., Nunziatini, A., Fani, V., Bindi, B., 2022. A data-driven approach for process simulation optimization: a case study, in: Proceedings-European Council for Modelling and Simulation. pp. 36–42.

Chen, X., Wang, Y., Lei, S., Li, Z., Huang, C., 2025. Optimization Design and Simulation of Supply Chain Demand Forecast Model Based on Multi-Agent Reinforcement Learning, in: Proceedings of 3rd International Conference on Mathematics and Machine Learning. pp. 374–382. https://doi.org/10.1145/3783779.3783842

Deshpande, M., Sahasrabudhe, V., Agarwal, P., Zodpe, A., 2025. Deep Reinforcement Learning for Supply Chain Optimization: A DQN and LSTM-Based Approach, in: International Conference on Emerging Trends in Engineering and Technology, ICETET. IEEE Computer Society. https://doi.org/10.1109/ICETETSIP64213.2025.11156380

Fani, V., Bandinelli, R., Rinaldi, R., 2017. Optimizing products allocation with simulation in the fashion industry: a multi-company case study, in: Proceedings of the 2017 Winter Simulation Conference. IEEE.

Fani, V., Bucci, I., Bandinelli, R., da Silva, E.R., 2025. Sustainable reverse logistics network design using simulation: Insights from the fashion industry. Cleaner Logistics and Supply Chain 14. https://doi.org/10.1016/j.clscn.2024.100201

Hu, J., Xia, L., Huang, T., Wu, H., 2025. A multi-agent deep reinforcement learning approach for multi-echelon inventory optimization and its application to the beer game. Transp. Res. E Logist. Transp.

Rev. 203, 104367. https://doi.org/10.1016/j.tre.2025.104367
Kou, A., Cheng, Y., Huang, X., Jin, J., 2025. Dynamic replenishment policy for perishable goods using change point detection-based soft actor-critic reinforcement learning. Expert Syst. Appl. 270. https://doi.org/10.1016/j.eswa.2025.126556
Kumar, L., Khedlekar, U.K., 2025. A hybrid AI and blockchain framework for intelligent supply chain optimization with reinforcement learning-based inventory management, stochastic demand forecasting, and sustainable supplier selection. Soft comput. https://doi.org/10.1007/s00500-025-10947-8
Ma, Z., Chen, X., Sun, T., Wang, X., Wu, Y.C., Zhou, M., 2024. Blockchain-Based Zero-Trust Supply Chain Security Integrated with Deep Reinforcement Learning for Inventory Optimization. Future Internet 16. https://doi.org/10.3390/fi16050163
Mao, S., Wang, Y., Guan, Q., Xu, Y., 2023. Structural Charging and Replenishment Policies for Battery Swapping Charging System Operation Under Uncertainty. IEEE Transactions on Intelligent Transportation Systems 24, 14598–14609. https://doi.org/10.1109/TITS.2023.3298320
Mutyala, P.K., Kaur, A., 2024. Inventory Optimization of Multi-Echelon Supply Chain Under Market Uncertainties, in: 1st International Conference on Pioneering Developments in Computer Science and Digital Technologies, IC2SDT 2024 - Proceedings. Institute of Electrical and Electronics Engineers Inc., pp. 257–262. https://doi.org/10.1109/IC2SDT62152.2024.10696052
Quadras, D.L.O., Pereira Mafia, M.M., Mendes, L.G., Braghirolli, L.F., Frazzon, E.M., 2024. Perspectives for the Application of Reinforcement Learning for the Integrated Order-Dispatching and Maintenance Scheduling, in: IFAC-PapersOnLine. Elsevier B.V., pp. 79–84. https://doi.org/10.1016/j.ifacol.2024.08.054
Rizqi, Z.U., Chou, S.Y., 2024. Neuroevolution reinforcement learning for multi-echelon inventory optimization with delivery options and uncertain discount. Eng. Appl. Artif. Intell. 134. https://doi.org/10.1016/j.engappai.2024.108670
Shar, I. El, Sun, W., Wang, H., Gupta, C., 2022. Deep Reinforcement Learning toward Robust Multi-echelon Supply Chain Inventory Optimization, in: IEEE International Conference on Automation Science and Engineering. IEEE Computer Society, pp. 1385–1391. https://doi.org/10.1109/CASE49997.2022.9926659
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction, 1st ed. Cambridge: MIT press.
Wang, J., Zhou, S., Li, M., Ren, G., Ren, X., Xiong, X., Zhang, Y., 2025. Multi-echelon inventory optimization of waste electrical and electronic equipment closed-loop supply chain based on reinforcement learning under carbon tax policy. Eng. Appl. Artif. Intell. 154. https://doi.org/10.1016/j.engappai.2025.110987
Yang, S., Jihui, Z., 2015. Adaptive inventory control and bullwhip effect analysis for supply chains with non-stationary demand, in: The 27th Chinese Control and Decision Conference. Qingdao, China, pp. 3903–3908. https://doi.org/10.1109/CCDC.2015.7162605
Zhou, Y., Guo, K., Yu, C., Zhang, Z., 2024. Optimization of multi-echelon spare parts inventory systems using multi-agent deep reinforcement learning. Appl. Math. Model. 125, 827–844. https://doi.org/10.1016/j.apm.2023.10.039

AUTHOR BIOGRAPHIES



DJONATHAN QUADRAS holds a bachelor and master's degree in industrial engineering provided by Universidade Federal de Santa Catarina, in Brazil. He is currently a PhD Candidate in Industrial Engineering at the University of Florence, in Italy. His areas of interest include the digitalization of supply chains, simulation, optimization algorithms and AI algorithms applied to industrial processes. His e-mail address is: djonathan.quadras@unifi.it.



ROMEO BANDINELLI is an Associate Professor at the University of Florence, Department of Industrial Engineering. He is member of the IFIP 5.1 "Global Product development for the whole lifecycle" and Scientific Director of the Laboratories Fashion for Future (www.fashionforfuture.it) and AGRItech COmmunity for Learning and Application (https://agricola.unifi.it/). He has authored more than 80 papers on proceedings and international peer-reviewed journals. His e-mail address is: romeo.bandinelli@unifi.it.



VIRGINIA FANI is a Postdoctoral Research Fellow at the Department of Industrial Engineering, University of Florence. She graduated with full marks in Management Engineering from Politecnico di Milano in 2014 and obtained her PhD at the Department of Industrial Engineering, University of Florence. Her research interests include sustainability in supply chains, with a particular focus on the fashion industry. Dr. Fani has contributed to various international journals and conferences, sharing her expertise in these areas. Her e-mail address is: virginia.fani@unifi.it.