Integrating simulation and reinforcement learning for optimized working capital management in supply chains

BADAKHSHAN, Ali, BADAKHSHAN, Ehsan <http://orcid.org/0000-0002-5298-764X>, SAAD, Sameh M <http://orcid.org/0000-0002-9019-9636> and BAHADORI, Ramin <http://orcid.org/0000-0001-6439-7033>

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 277 (2026) 263–270

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Industry of the Future and Smart Manufacturing
(former International Conference on Industry 4.0 and Smart Manufacturing)

Integrating simulation and reinforcement learning for optimized working capital management in supply chains

Ali Badakhshan[a], Ehsan Badakhshan[b]*, Sameh M Saad[b], Ramin Bahadori[b]

[a]*University of Durham, Durham, UK*
[b]*Sheffield Hallam University, Sheffield, UK*

Abstract

Effective working capital management in supply chains requires the joint coordination of physical and financial flows. While existing literature on simulation-integrated reinforcement learning has primarily focused on inventory dynamics, this study extends the scope to include financial dynamics across supply chain tiers. We propose a framework that integrates discrete-event simulation (DES) with deep reinforcement learning (DRL) to optimize both inventory and financial management in a multi-echelon supply chain. A Proximal Policy Optimization (PPO) algorithm is used to train an agent within the simulated environment, enabling it to learn adaptive policies for inventory replenishment, production planning, and cash collection. Comparative results against a Genetic Algorithm (GA)-based benchmark demonstrate that the DRL agent outperforms heuristic policies in terms of convergence stability, cumulative rewards, and responsiveness to stochastic demand. The findings highlight the potential of simulation-integrated DRL frameworks to improve coordination between financial and operational decisions in supply chains. Practically, the framework can be embedded in digital twins to support real-time decision-making in supply chains and offers actionable insights for managers seeking to improve working capital efficiency through adaptive policies.

Keywords: Working capital management; Deep reinforcement learning (DRL); Discrete-event simulation (DES); Proximal policy optimization (PPO); Supply chain

* Corresponding author.
E-mail address: e.badakhshan@shu.ac.uk

1. Introduction

Working capital management is a critical component of supply chain (SC) performance, involving the efficient control of short-term assets and liabilities, including inventory, cash, accounts receivable, and accounts payable. Effective working capital management supports liquidity, reduces the cost of capital tied up in operations, and enhances operational flexibility [1]. While traditional approaches often focus on optimizing these elements within individual firms, such as minimizing local inventory or extending payment terms, such siloed practices can lead to suboptimal outcomes at the SC level. Recent research increasingly advocates for a SC-oriented perspective on working capital management, in which partners collaborate to reduce capital costs and strengthen financial resilience [1, 2].

This integrated view requires the alignment of operational and financial decisions across SC tiers. However, achieving such alignment is challenging due to the inherent uncertainty and dynamic nature of SCs. Volatile demand, fluctuating lead times, and disruptions can easily result in excess inventory or stockouts, both of which erode working capital efficiency and profitability. Additionally, mismatches in financial flows, such as misaligned payment terms or delayed receivables, can exacerbate liquidity risks even when operational performance remains strong [3]. These complexities highlight the need for advanced analytical tools capable of jointly optimizing physical and financial SC decisions under uncertainty.

Simulation has long served as a foundational method for modeling SC behavior under stochastic conditions. Techniques such as discrete-event simulation (DES), agent-based modeling (ABM), and system dynamics (SD) enable modelers to replicate interactions among SC entities and assess the implications of various operational and financial policies [4]. By modeling the flows of products, information, and capital, simulation provides a controlled environment to experiment inventory control strategies, order fulfillment, and financial decision-making under variable conditions. Among these methods, DES is particularly effective for analyzing inventory dynamics, lead times, and discrete SC events [5]. Nevertheless, simulation alone does not prescribe optimal policies; rather, it offers a controlled environment for evaluating candidate strategies. Traditionally, optimization within simulation models has been achieved through the incorporation of metaheuristics, such as genetic algorithms, to search for near-optimal decision rules [6]. While effective, these approaches often require extensive expert input and can be computationally intensive, particularly when addressing high-dimensional decision spaces.

In recent years, the rise of machine learning techniques has opened new possibilities for data-driven, adaptive decision-making in SCs. In particular, reinforcement learning (RL) has emerged as a powerful approach for sequential decision optimization under uncertainty. RL algorithms enable an autonomous agent to learn effective policies through interaction with an environment, receiving feedback in the form of rewards. Unlike supervised learning, RL does not require labeled optimal actions; instead, the agent discovers superior strategies through exploration and outcome observation. The appeal of RL in supply chain management (SCM) lies in its model-free nature, meaning it does not require an explicit analytical model of the system, and its ability to continually improve decisions in complex, dynamic environments [7]. As SCs grow in complexity and data availability increases, researchers have investigated RL-based approaches for challenges ranging from inventory control to transportation and production scheduling. A recent literature review highlights that inventory management remains the most common application domain for RL methods, underscoring the pivotal role of inventory in SC coordination [8].

Several studies have demonstrated that modern deep reinforcement learning (DRL) algorithms can match or exceed the performance of traditional heuristics in complex SC problems. For instance, Gijsbrechts et al. [9] showed that an advanced DRL method, using an asynchronous advantage actor-critic (A3C) algorithm, could achieve near-optimal performance in inventory control problems involving lost-sales and multi-echelon systems. These findings suggest that RL can be a valuable tool for dynamically optimizing operational and financial decisions in SCs under uncertainty.

Integrating simulation and RL offers a synergistic approach to addressing complex SC problems. In an integrated simulation–RL framework, the simulation model serves as the environment for the learning agent: it replicates the SC's behavior in response to the agent's actions, such as inventory replenishment, production scheduling, or credit allocation, and generates rewards based on performance metrics such as cost, service level, or cash flow outcomes. The agent, in turn, updates its policy to maximize cumulative rewards, thereby progressively improving its decision-making capabilities. For example, reward functions can be designed to reflect profitability or working capital efficiency, directly aligning learning objectives with financial performance. This closed-loop interaction constitutes a

flexible form of simulation-based optimization, enabling continuous policy adaptation to complex stochastic environments, even when analytical solutions are intractable [10].

A number of studies have applied integrated simulation and RL approaches to SC problems, providing insights into their effectiveness. For example, Chaharsooghi et al. [11] developed an RL-based model for the well-known Beer Game, a multi-echelon SC simulation that illustrates inventory oscillations under uncertainty. They used a Q-learning agent to adjust ordering policies at each echelon and found that the RL-driven coordination policy outperformed traditional decision rules, effectively dampening the bullwhip effect and improving overall cost performance. Subsequently, Mortazavi et al. [12] proposed a simulation-based optimization framework employing Q-learning for a multi-tier SC ordering system, incorporating more sophisticated performance measures such as value-at-risk for costs. Building on these earlier contributions, Preil and Krapp [13] introduced a novel integration of simulation and RL using multi-armed bandit algorithms to optimize base-stock levels in a stochastic multi-echelon SC. These studies collectively established a foundation for treating SC decision-making problems as Markov decision processes and solving them through learning agents interacting with simulation models.

With advances in computational power and algorithmic design, recent studies have increasingly adopted deep reinforcement learning (DRL) within SC simulations to address larger and more complex state and action spaces. For example, Fujii et al. [14] applied a deep multi-agent RL approach to an extended Beer Game scenario, using neural networks to approximate policies and introducing an evolutionary mechanism to co-evolve the agents' learning processes. Their findings demonstrated superior cost and service level outcomes compared to classical methods. Similarly, Oroojlooyjadid et al. [15] developed a deep Q-network (DQN) agent for the Beer Game, showing that the learned policies could match or exceed human performance and generalize effectively across varying cost structures.

Beyond inventory management, simulation–RL integration has also been applied to dynamic delivery strategies [16], production scheduling under uncertain manufacturing conditions [17], and supplier selection in competitive SC environments [18]. In each case, simulation captures the complexity and stochasticity of the environment, while RL enables the autonomous development of responsive and adaptive policies.

While integrated simulation and RL approaches have demonstrated success in operational domains such as inventory management and logistics, their application to working capital management has focused narrowly on inventory dynamics, without explicitly modeling the financial aspects that are central to working capital efficiency. Considering that working capital extends beyond the management of physical goods to include cash flow timing and financing strategies, there is a clear need for integrated frameworks that combine the simulation of both operational and financial SC elements with RL-based decision-making. To address this gap, we develop a DES model to train an RL agent for optimized working capital management in SCs.

2. Material and methods

2.1. SC configuration and Simulation model

In this study, we examine a single-product, three-tier serial SC comprising a manufacturer, a wholesaler, and a retailer, representative of a typical fast-moving consumer goods (FMCG) SC [19]. The distribution lead time between each SC stage is set to one week. There is no lead time between the retailer and the end customer, as customers collect their orders directly from the retailer. The manufacturer operates with a weekly production capacity of 40,000 units. Customer demand is stochastic and follows a uniform distribution between 10,000 and 17,000 units per week. If the retailer's inventory is insufficient to meet this demand, unmet orders are backlogged, which negatively affects the SC's service level defined as the ratio of the retailer's fulfilled sales to total customer demand.

All SC members follow a periodic review inventory policy with a review interval of one week. Accordingly, each week, members assess their inventory and work-in-progress (WIP) levels and place replenishment orders upstream.

The operational sequence for each SC node includes the following steps: (1) receipt of products ordered one week prior, added to inventory; (2) use of inventory to fulfill downstream orders and any existing backlog; (3) dispatch of products downstream, update of inventory and WIP levels, and creation of backlogs if inventory is insufficient; and (4) issuance of a non-negative replenishment order to the upstream node.

DES is employed to model the dynamic behavior of the SC. Working capital performance is assessed using the cash conversion cycle (CCC). To verify the simulation model, output data analysis and simulation run monitoring

were conducted. For validation, the model was run 100 times using a fixed set of simulation parameters, and results across replications were compared to assess robustness. Furthermore, the model demonstrates the bullwhip effect when configured according to the Beer Game assumptions, as described in [20].

2.2. Reinforcement learning

To optimize working capital decisions dynamically, this study integrates a Deep Reinforcement Learning (DRL) approach with a discrete-event simulation model. The problem is formulated as a Markov Decision Process (MDP), in which an agent learns optimal policies for inventory replenishment and cash collection by interacting with the simulated environment. A Proximal Policy Optimization (PPO) algorithm with an actor–critic architecture is employed to train the agent. The state space captures current inventory levels and backlog positions across the manufacturer, wholesaler, and retailer. The action space includes (i) continuous cash collection policy parameters for the manufacturer and wholesaler, (ii) discrete order quantities for each supply chain tier, and (iii) the manufacturer's production rate, subject to capacity constraint. The reward function is defined as a weighted linear combination of cash conversion cycles for all supply chain entities and the backlog at the retail node.

3. Results and discussion

The performance of the proposed DRL framework, implemented using PPO, was evaluated against a benchmark Genetic Algorithm (GA)-based policy from the literature [21]. This comparative analysis involved three key aspects: convergence behavior, policy performance, and reward distribution across evaluation episodes.

During PPO training, convergence of the critic's value function was assessed using the mean squared error (MSE). As illustrated in Fig. 1, the value-function loss declined steadily from approximately 1.0 to below 10^{-4}, indicating rapid and stable convergence. The low final loss suggests the value network was able to accurately estimate expected returns, thereby enabling efficient policy updates via PPO's clipped surrogate objective. This behavior confirms PPO's stability and suitability for high-dimensional control in stochastic SC environments.

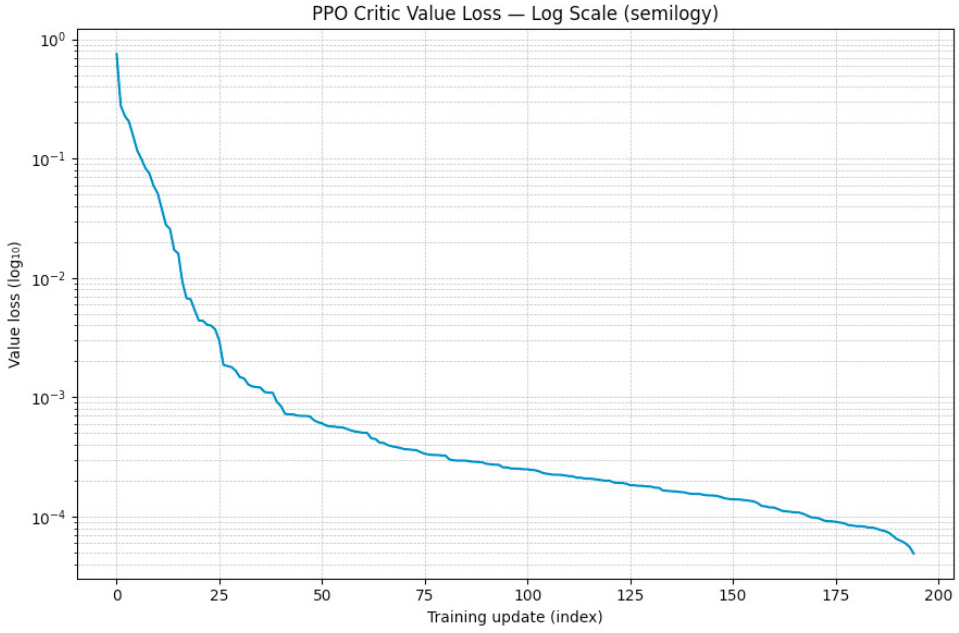

Fig.1. log-scaled critic value loss during PPO training.

The Genetic Algorithm was subjected to an identical decision space and reward structure for fairness of comparison. As shown in Fig. 2, GA demonstrated fast initial gains in both mean and maximum fitness within the first 20 generations. However, convergence quickly plateaued, indicating stagnation near suboptimal local optima. This phenomenon reflects the GA's limited capacity to explore and exploit complex temporal dependencies in supply chain dynamics, especially under uncertainty.

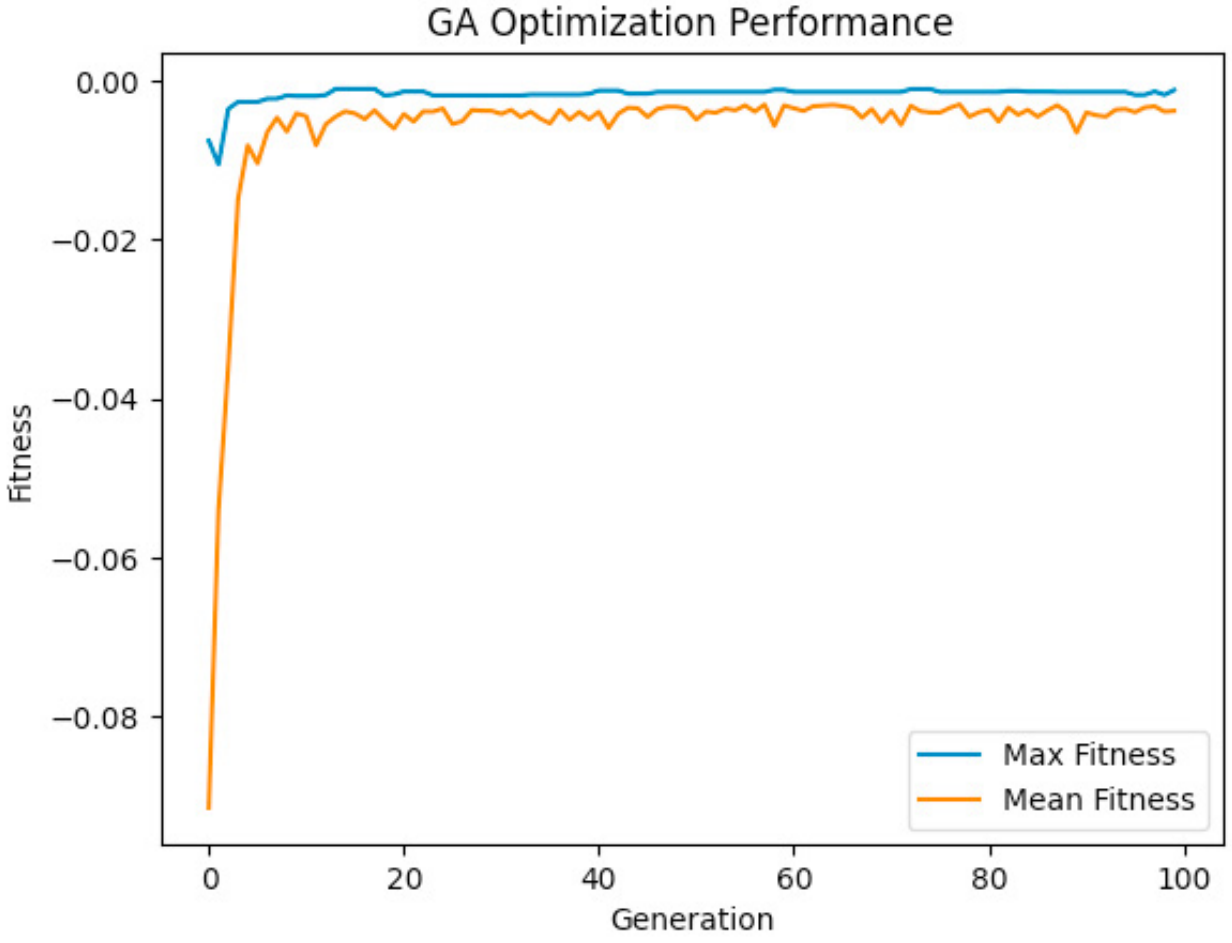

Fig. 2. convergence of GA during policy optimization.

To quantify real-world performance, both PPO and GA policies were evaluated across 1,000 simulation episodes, each spanning a 50-week operational horizon. As shown in Fig. 3, the PPO-trained policy mostly achieved positive cumulative rewards, whereas the GA-derived policies resulted in negative returns. These outcomes underscore the critical advantage of DRL methods in learning adaptive and resilient policies that account for long-term state transitions and feedback effects.

The performance disparity highlights two important findings. First, DRL's ability to model sequential decision-making allows for more robust inventory and cash flow strategies compared to static or heuristic-based optimization methods. Second, while GA may provide quick heuristics in simpler settings, it lacks the dynamic adaptability required for real-time supply chain decision-making under uncertainty. The results support the broader applicability of simulation-integrated DRL frameworks for optimizing financial and operational objectives in supply networks.

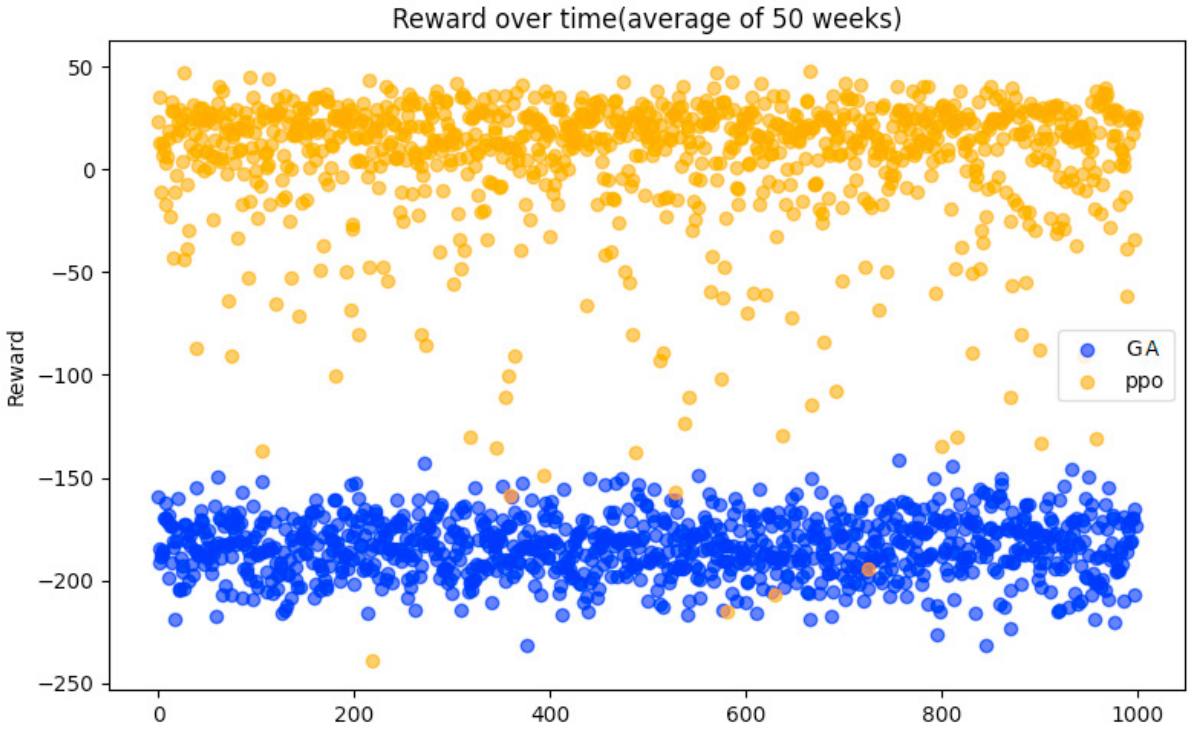

Fig. 3. reward distribution across 1,000 evaluation episodes for PPO and GA policies.

The behaviors of the PPO-trained and GA-trained policies were further analyzed by inspecting state-action mappings at selected decision steps. Table 1 summarizes two representative examples. The comparative results reveal notable differences between the GA and PPO policies in terms of responsiveness, adaptability, and strategic alignment with SC conditions. In Step 1, the system state indicates that the manufacturer has zero inventory and is experiencing a backlog of 8,000 units, while the retailer and wholesaler hold moderate inventory levels and no backlog. The GA policy responds with relatively aggressive ordering behavior, placing substantial orders at all levels, including a 14,789 unit order to the manufacturer despite its inability to fulfill it due to stockouts and backlog. This action suggests that the GA does not effectively incorporate state feedback into its decision-making and applies a static policy that may result in infeasible or suboptimal decisions. In contrast, the PPO agent demonstrates a more cautious and context-aware approach. It issues moderately sized orders downstream (15,352 at the retail level and 30,000 at the wholesale level) but places no order to the manufacturer. This indicates that the PPO agent recognizes the manufacturer's constrained capacity and avoids contributing to the existing backlog.

In Step 2, the state changes slightly: retailer inventory rises to 37,896, while wholesaler inventory drops to a critically low 2,000 units. Manufacturer inventory remains at zero, and its backlog increases to 25,000. The GA policy, however, repeats the same set of actions from Step 1, i.e., issuing identical cash collection rates and order quantities, demonstrating a lack of dynamic adjustment to state transitions. This further reinforces the conclusion that the GA-derived policy lacks state sensitivity and generalization ability. On the other hand, the PPO policy adapts its ordering behavior by increasing the retail order to 20,000, reflecting an awareness of upstream constraints and the need to support downstream availability. Still, it refrains from placing an order with the manufacturer, consistent with its recognition that the manufacturer is unable to fulfill new orders due to lack of inventory and existing backlog.

Overall, the PPO agent exhibits significantly more robust and adaptive behavior, maintaining stable and conservative cash collection policies while dynamically adjusting inventory decisions based on real-time state conditions. In contrast, the GA policy is static and potentially counterproductive under supply constraints, highlighting the value of reinforcement learning in capturing temporal dependencies and SC interdependencies. These observations

underscore the advantage of DRL-based methods for joint financial and operational decision-making in SC environments.

Table 1. example state-action mapping from PPO agent in the simulation environment.

Step	Adjusted State Vector	Transformed GA Action	Transformed PPO Action
1	r_inventory: 26,785 w_inventory: 22,000 m_inventory: 0 r_backlog: 0 w_backlog: 0 m_backlog: 8,000	cash_col_retail: 0.58 cash_col_wholesale: 0.53 order_retail: 17,351 order_wholesale: 40,556 order_manufacturer: 14,789	cash_col_retail: 0.5 cash_col_wholesale: 0.5 order_retail: 15,352 order_wholesale: 30,000 order_manufacturer: 0
2	r_inventory: 37,896 w_inventory: 2,000 m_inventory: 0 r_backlog: 0 w_backlog: 0 m_backlog: 25,000	cash_col_retail: 0.58 cash_col_wholesale: 0.53 order_retail: 17,351 order_wholesale: 40,556 order_manufacturer: 14,789	cash_col_retail: 0.5 cash_col_wholesale: 0.5 order_retail: 20,000 order_wholesale: 30,000 order_manufacturer: 0

4. Conclusions

This study proposes a simulation-based deep reinforcement learning (DRL) framework for optimizing working capital decisions across supply chain tiers. By modeling the decision problem as a Markov Decision Process and integrating a Proximal Policy Optimization (PPO) agent into a discrete-event simulation (DES) environment, we addressed both operational and financial control dimensions in a unified setting. The results demonstrate that the DRL-based policy outperforms a genetic algorithm (GA) benchmark in terms of cumulative reward, policy stability, and responsiveness to stochastic demand. The agent learned to adaptively coordinate inventory replenishment, production planning, and cash collection, achieving improved working capital efficiency.

This research contributes to the literature by extending reinforcement learning applications in supply chains beyond traditional inventory control, incorporating financial flow into the optimization process. It also introduces a scalable, data-driven framework that can be embedded in digital twin architectures [22] to enable adaptive learning and feedback loops for continuously optimized decision-making in dynamic SC environments.

From a practical perspective, the framework offers a decision-support tool for managers aiming to optimize liquidity and service performance in integrated supply chain systems. This study has a few limitations: it assumes a single-product supply chain, uniform lead times, and centralized decision-making. Future research could expand the framework to multi-product and multi-agent settings, integrate variable lead times or dynamic payment terms, and explore risk-aware reward structures for more realistic financial modeling. Moreover, integrating the model with emerging technologies such as blockchain [23] could enhance transparency, traceability, and trust in financial transactions.

References

[1] Badakhshan, Ehsan, and Ramin Bahadori. (2024) “A simulation-based optimization model for balancing economic profitability and working capital efficiency using system dynamics and genetic algorithms.” *Decision Analytics Journal* **12**: 100498.

[2] Hofmann, Erik, Juuso Töyli, and Tomi Solakivi. (2022) “Working capital behavior of firms during an economic downturn: an analysis of the financial crisis era.” *International Journal of Financial Studies* **10 (3)**: 55.

[3] Haralambides, Hercules, and Girish Gujar. (2023) “The ‘new normal’, global uncertainty and key challenges in building reliable and resilient supply chains.” *Maritime Economics & Logistics* **25 (4)**: 623-638.

[4] Korder, Benjamin, Julien Maheut, and Matthias Konle. (2024) “Simulation methods and digital strategies for supply chains facing disruptions: Insights from a systematic literature review.” *Sustainability* **16 (14)**: 5957.

[5] Badakhshan, Ehsan, and Peter Ball. (2024) "Deploying hybrid modelling to support the development of a digital twin for supply chain master planning under disruptions." *International Journal of Production Research* **62(10)**: 3606-3637.

[6] Badakhshan, Ehsan, Paul Humphreys, Liam Maguire, and Ronan McIvor. (2020) "Using simulation-based system dynamics and genetic algorithms to reduce the cash flow bullwhip in the supply chain." *International Journal of Production Research* **58 (17)**: 5253-5279.

[7] Yan, Yimo, Andy HF Chow, Chin Pang Ho, Yong-Hong Kuo, Qihao Wu, and Chengshuo Ying. (2022) "Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities." *Transportation Research Part E: Logistics and Transportation Review* **162**: 102712.

[8] Rolf, Benjamin, Ilya Jackson, Marcel Müller, Sebastian Lang, Tobias Reggelin, and Dmitry Ivanov. (2023) "A review on reinforcement learning algorithms and applications in supply chain management." *International Journal of Production Research* **61 (20)**: 7151-7179.

[9] Gijsbrechts, Joren, Robert N. Boute, Jan A. Van Mieghem, and Dennis Zhang. (2019) "Can deep reinforcement learning improve inventory management." *Performance on dual sourcing, lost sales and multi-echelon problems*.

[10] Badakhshan, Ehsan, Navonil Mustafee, and Ramin Bahadori. (2024) "Application of simulation and machine learning in supply chain management: A synthesis of the literature using the Sim-ML literature classification framework." *Computers & Industrial Engineering*: 110649.

[11] Chaharsooghi, S. Kamal, Jafar Heydari, and S. Hessameddin Zegordi. (2008) "A reinforcement learning model for supply chain ordering management: An application to the beer game." *Decision Support Systems* **45 (4)**: 949-959.

[12] Mortazavi, Ahmad, Alireza Arshadi Khamseh, and Parham Azimi. (2015) "Designing of an intelligent self-adaptive model for supply chain ordering management system." *Engineering Applications of Artificial Intelligence* **37**: 207-220.

[13] Preil, Deniz, and Michael Krapp. (2022) "Bandit-based inventory optimisation: reinforcement learning in multi-echelon supply chains." *International Journal of Production Economics* **252**: 108578.

[14] Fuji, Taiki, Kiyoto Ito, Kohsei Matsumoto, and Kazuo Yano. (2018) "Deep multi-agent reinforcement learning using dnn-weight evolution to optimize supply chain performance." *Proceedings of the 51st Hawaii International Conference on System Sciences*.

[15] Oroojlooyjadid, Afshin, MohammadReza Nazari, Lawrence V. Snyder, and Martin Takáč. (2022) "A deep q-network for the beer game: Deep reinforcement learning for inventory optimization." *Manufacturing & Service Operations Management* **24 (1)**: 285-304.

[16] Zou, Guangyu, Jiafu Tang, Levent Yilmaz, and Xiangyu Kong. (2022) "Online food ordering delivery strategies based on deep reinforcement learning." *Applied Intelligence*: 1-13.

[17] Serrano-Ruiz, Julio C., Josefa Mula, and Raul Poler. (2022) "Development of a multidimensional conceptual model for job shop smart manufacturing scheduling from the Industry 4.0 perspective." *Journal of Manufacturing Systems* **63**: 185-202.

[18] Lee, Yoon Sang, and Riyaz Sikora. (2019) "Application of adaptive strategy for supply chain agent." *Information Systems and e-Business Management* **17**: 117-157.

[19] Badakhshan, Ehsan, Peter Ball, and Ali Badakhshan. (2022) "Using digital twins for inventory and cash management in supply chains." *IFAC-PapersOnLine* **55(10)**: 1980-1985.

[20] Sterman, John. "System Dynamics: systems thinking and modeling for a complex world." (2002).

[21] Badakhshan, Ehsan, Paul Humphreys, Liam Maguire, and Ronan Mclvor. (2018) "Simulation-based system dynamics optimization modelling of supply chain working capital management under lead time uncertainty." *International Conference on Intelligent Systems (IS),* pp. 934-938.

[22] Badakhshan, Ehsan, and Peter Ball. (2023) "Applying digital twins for inventory and cash management in supply chains under physical and financial disruptions." *International Journal of Production Research* **61(15)**: 5094-5116.

[23] Badakhshan, Ehsan, and Dmitry Ivanov. (2025) "Integrating digital twin and blockchain for responsive working capital management in supply chains facing financial disruptions." *International Journal of Production Research*: 1-35.