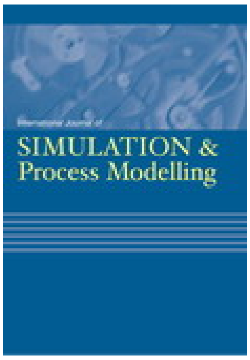


International Journal of Simulation and Process Modelling

ISSN online: 1740-2131 - ISSN print: 1740-2123
https://www.inderscience.com/ijspm

Discrete-event simulation modelling of inventory turnover under supply chain financial collaboration

Zhifeng Qu

DOI: 10.1504/IJSPM.2025.10075344

Article History:

Received:	14 October 2025
Last revised:	16 November 2025
Accepted:	18 November 2025
Published online:	06 March 2026

Discrete-event simulation modelling of inventory turnover under supply chain financial collaboration

Zhifeng Qu

School of Management,
Wuhan Technology and Business University,
Wuhan, 430065, China
Email: WTBU080806@163.com

Abstract: This study addresses the critical gap between operational and financial objectives in supply chain inventory management by proposing a novel framework that integrates discrete-event simulation with deep reinforcement learning. We formulate a dual-objective reward function incorporating both traditional costs (holding, shortage, ordering) and the financial metric of cash conversion cycle. Trained and tested on the real-world M5 forecasting accuracy dataset, our model, cognitive load dynamic assessment model-proximal policy optimisation, demonstrates superior performance. Results show it achieves a total cost of 285.4 ± 8.7 (in thousands), significantly lower than state-of-the-art baselines ($p < 0.01$), while maintaining a 98.2% service level and reducing cash conversion cycle to 35.2 days. This result highlights the framework's effectiveness in achieving operational-financial synergy, offering a data-driven decision-support tool for enhancing both efficiency and financial health in dynamic supply chain environments.

Keywords: supply chain finance; deep reinforcement learning; DRL; inventory optimisation; cash conversion cycle; M5 forecasting accuracy dataset.

Reference to this paper should be made as follows: Qu, Z. (2026) 'Discrete-event simulation modelling of inventory turnover under supply chain financial collaboration', *Int. J. Simulation and Process Modelling*, Vol. 23, No. 1, pp.11–23.

Biographical notes: Zhifeng Qu is an Associate Professor in the School of Management at Wuhan Technology and Business University, China. She obtained her Bachelor's degree in Management from Wuhan University of Technology (2001) and Master's degree in Management from Wuhan University (2006), China. She has published over one EI Compendex-indexed paper. Her research interests include ESG, simulation modelling, and green management accounting.

1 Introduction

As the core of modern enterprise operation, supply chain management directly determines the competitiveness and survivability of enterprises in the globalised market. Inventory management, serving as the critical link that connects procurement, production, and sales, also profoundly influences the operational efficiency and financial health of enterprises. Ideal inventory levels can minimise the costs associated with capital consumption, storage, and wastage while meeting uncertain demand, thus achieving a dynamic balance between operational flexibility and financial health. However, this balance is extremely fragile in reality. Increasingly short life cycles of demand, increased market volatility, and global emergencies (e.g., epidemics, geopolitical conflicts) have made traditional inventory management methods based on static assumptions of historical data frequently ineffective, resulting in companies being caught in the dilemma of inventory backlogs and shortages – the former eroding profits, the latter damaging customer relationships and market share. This classic dilemma of optimising inventory under uncertainty has been a central focus of research in operations management, management science, and industrial engineering for decades.

Throughout the evolution of inventory management, its theoretical cornerstones are deeply rooted in operations research and dynamic planning. Early studies such as Arrow et al. (1951) formalised the classical economic order lot model, providing a mathematical framework for cost trade-offs. Subsequently, Bellman (1966) pioneered the theory of dynamic programming, which laid the methodological foundation for dealing with multicycle stochastic inventory problems. Iglehart (1960)'s proof of the optimality of the (s, S) policy under certain conditions was a milestone in the field, and its influence continues to this day. These seminal works constructed a paradigm for the pursuit of analytically optimal solutions. However, as Mauldin (2017) systematically explained in his authoritative work, when the problem size is increased and the assumptions are relaxed (e.g., introduction of non-smooth demands, multi-hierarchical structures), the complexity of the model exceeds the capacity of analytic methods, and encounters the 'curse of dimensionality'. Although subsequent research, such as Silver et al. (1998), has made excellent contributions to pragmatic heuristics, these methods inherently rely on a high degree of simplification of reality, and their parameters, once set, lack the adaptability to cope with the continuous and rapid environmental changes that characterise the current volatility, uncertainty, complexity, ambiguity (VUCA) era.

In order to break through the limitations of analytical models, simulation and optimisation methods have emerged with advances in computational technology, providing researchers with a 'digital laboratory' for exploring complex systems. In his classic textbook, Law et al. (2007) systematically discusses how discrete-event simulation can model supply chain systems with stochasticity, feedback loops, and complex logic with high fidelity. For example, Fu (1994) provides a comprehensive review of simulation-based optimisation in a number of domains, including queuing networks and inventory systems, highlighting its 'model-driven' advantages – i.e., it eliminates the need for explicit mathematical modelling of real systems. However, the fundamental flaw of this approach is its 'offline' nature. The optimisation process is computationally intensive, time-consuming, and results in a static set of policy parameters. In the event of unforeseen structural changes in the real environment, the performance of the policy degrades rapidly, requiring the system to restart a time-consuming optimisation cycle, and lacks the 'online intelligence' to learn and adjust in real-time during operation.

In recent years, the third wave of artificial intelligence has merged reinforcement learning with deep learning, giving rise to deep reinforcement learning (DRL), which has brought about a paradigm shift in solving sequential decision problems. The textbook by Sutton and Barto (1998) laid the theoretical foundations of this field, describing how intelligences learn optimal strategies through trial and error by interacting with the environment. The deep Q-network (DQN) proposed by Mnih et al. (2015) successfully combined deep learning with Q-learning, demonstrating the ability of DRL to learn superior strategies directly from high-dimensional perceptual inputs, opening a new era of DRL. This breakthrough has quickly attracted the attention of operations management scholars. For example, Oroojlooyjadid et al. (2022) applied DQN to the famous 'beer game' multilevel supply chain, demonstrating its ability to outperform traditional heuristics in a complex network environment. Selukar et al. (2022) employs an actor-critic framework for the specific constrained problem of perishable inventory, demonstrating the flexibility of DRL to handle domain-specific constraints. However, this burgeoning frontier is facing a critical and common bottleneck: the vast majority of current studies are still too narrow and isolated in their vision when designing their core, the reward function. They usually narrowly define the optimisation objective as the sum of operating costs (e.g., holding costs, ordering costs, and out-of-stock costs), while completely or largely ignoring the profound impact of inventory decisions on the other lifeline of the firm: financial liquidity.

This disconnect between operational-financial decision-making is one-sided in theory and dangerous in practice. Inventory is a key balance sheet current asset, and its turnover efficiency directly drives the core financial metric of the 'cash conversion cycle (CCC)'. Excessively high inventories mean that a large amount of working capital is frozen, lengthening the CCC and increasing the need for external financing and finance costs, while excessively low inventories can lead to lost sales due to out-of-stocks, which can also hurt cash inflows. Thus, an inventory strategy that aims only to minimise the visible costs of operations may well unknowingly worsen a firm's cash flow position. The intersection of operations and finance is not an entirely new concept, with earlier pioneering studies such as Buzacott and Zhang (2004)'s asset finance-based inventory management model revealing the profound impact of external financing constraints on optimal inventory strategy. Kouvelis and Zhao (2012) further analysed the value of trade credit-based financing under capital constraints in the newsvendor model. However, most of these studies stayed at the strategic or tactical level of co-modelling or relied on strict financial assumptions, failing to sink into the dynamic learning process that drives daily inventory decisions. In the context of DRL, Hubbs et al. (2020) although the holding cost of work-in-process inventory is considered in chemical production scheduling, its financial perspective is still limited to static cost parameters and does not touch the dynamic cash flow turnover. The empirical study by Frésard and Salva (2010), on the other hand, confirms that holding excess cash can lead to agency problems and reduce operational efficiency, which warns of the potential risks of optimising financial or operational objectives in isolation. In summary, there is a lack of a unified framework that can endogenise financial liquidity metrics into DRL learning mechanisms at the micro, dynamic level of daily operations. This theoretical gap makes it difficult for the existing intelligent inventory decision-making system to support enterprises to truly realise the ultimate goal of 'value creation', and its decision-making is inherently flawed in its scientific and comprehensive nature.

Based on this, this study aims to bridge the growing 'operational-financial' decision-making gap. We recognise that an intelligent inventory management system that can survive in the complex business environments of the 21st century must have two core competencies: the ability to simulate with high fidelity and respond dynamically to complex, non-stationary operating environments, and the ability to optimise the integrated value (including operational efficiency and financial health) of inventory decisions. Therefore, the research work in this paper is based on the core proposition that by constructing a collaborative optimisation framework that integrates discrete-event simulation and DRL, and by designing a novel reward function embedded with both operational cost and financial liquidity metrics, we are able to guide the intelligent agent to learn a completely new class of inventory strategies. These strategies can not only effectively cope with demand fluctuations and guarantee service levels, but also actively manage cash flow, the framework enables the realisation of overall synergistic optimisation of supply chain operations and financial performance. In order to verify this proposition in real-world scenarios, this study relies on the Walmart M5 forecasting accuracy dataset, a publicly available benchmark dataset in the retail industry, which provides several years of sales and related information at the item-store level, providing a solid foundation for constructing a close-to-reality simulation environment and conducting rigorous empirical tests. A solid foundation for building a realistic simulation environment and conducting rigorous empirical tests.

2 Related work

2.1 Classical inventory control theory and its limitations

The theoretical foundations of inventory management are rooted in operations research and dynamic planning, where the core objective is to minimise costs or maximise service levels under uncertain demand. Early seminal work such as the economic order euantity (EOQ) model proposed by Arrow et al. (1951) aimed at balancing ordering costs with holding costs in its classical form, with the optimal order quantity given by $\sqrt{\frac{2KD}{h}}$, where K is the fixed ordering cost, D is the annual demand rate, and h is the annual holding cost per unit of product. The model provides a basic cost trade-off framework for subsequent studies. Iglehart (1960) proved the optimality of the (s, S) strategy, defined as placing an order to raise the inventory level to a target level S when the inventory level drops to the reorder point s, under certain conditions. This result became a central cornerstone of the theory of multi-period stochastic inventories. The theory of dynamic programming established by Bellman (1966), on the other hand, provides a generalised mathematical tool for sequential decision problems, as encapsulated in the Bellman equation:

$$V_t(s) = \min_a \left[C_t(s,a) + \gamma V_{t+1}(s') \right] \tag{1}$$

Which expresses the optimal value function as the sum of the immediate cost and the discounted value of the subsequent state.

However, the elegance of these classical models relies heavily on a set of strict assumptions, such as a stable demand distribution, a fixed lead time, and a single decision objective. In his authoritative work, Mauldin (2017) systematically pointed out that when the problem size scales up to multi-product, multi-tier, or non-stationary demand, the solution space of dynamic programming grows exponentially, leading to so-called 'dimensional catastrophe' and making exact solutions computationally infeasible. Although subsequent studies such as Silver et al. (1998) developed a variety of practical heuristics to cope with complex environments, Graves (1999) emphasised that the parameters of these methods are usually based on static estimates of historical data, which significantly degrades the performance of these methods when faced with non-stationary demand due to promotions or sudden changes in the market, as is common in today's supply chains. Armenzoni et al. (2015) further analysed that in the context of demand forecast updating, the rigidity of the traditional strategies leads to systematic biases and fails to take full advantage of real-time information. Thus, an unbridgeable gap persists between the analytical rigor of classical theories and their applicability in complex real-world environments.

2.2 Application of simulation optimisation methods to inventory management

To overcome the limitations of analytical models, simulation optimisation methods have emerged, which evaluate the performance of different strategies by constructing a digital twin of the system. Law et al. (2007) detailed how discrete-event simulation can reproduce the dynamic behaviour of complex supply chains with high fidelity by modelling the stochastic processes of entities (e.g., orders), resources (e.g., inventory), and events (e.g., demand arrivals, replenishment arrivals). Simulation models do not have optimisation capabilities by themselves and need to be combined with metaheuristic algorithms. Fu (1994) referred to such methods collectively as 'simulation optimisation' and reviews the paradigm of their application to stochastic systems. For example, the taboo search proposed by Goldberg (1989) avoids local optimality by introducing a memory structure, which centres on maintaining a taboo list of recently visited solutions that directs the search towards new regions. Genetic algorithms pioneered by Goldberg (1989), on the other hand, mimic the process of natural selection by evolving better combinations of policy parameters in the solution space through selection, crossover and mutation operations.

In the area of inventory management, Almeder et al. (2009) successfully applied simulation optimisation to configure the (s, S) parameters of a multilevel inventory system, significantly reducing the total system cost:

$$\text{Total Cost} = \sum_{t=1}^{T} \left[K \cdot \mathbb{I}_{I_t < s} + h \cdot I_t^{+} + p \cdot I_t^{-} \right] \tag{2}$$

where T is the total number of cycles, $\mathbb{I}_{I_t < s}$ is an indicator function (1 if inventory is below s, 0 otherwise), I_t^{+} is the positive inventory at the end of the cycle, I_t^{-} is the out-of-stock quantity, and p is the cost of out-of-stock per unit. Although simulation optimisation provides a powerful tool for dealing with complexity and stochasticity, Figueira and Almada-Lobo (2014) poignantly points out its fundamental flaw: the method is inherently ‘offline’. The optimisation process typically requires thousands of simulations runs, which are computationally expensive and time-consuming. The final output is a static set of parameters, such as a fixed set of (s, S) values. Gumte et al. (2021) points out that in the event of unforeseen structural changes in the real environment (e.g., permanent changes in demand patterns or supply disruptions), the original optimised strategy fails rapidly, and the system lacks the ability to learn online and adaptively adjust to the situation, and must restart a time-consuming optimisation cycle. This lag limits its usefulness in fast-changing markets.

2.3 *Progress and shortcomings of reinforcement learning in inventory optimisation*

Reinforcement learning, especially its DRL branch combined with deep learning, provides a completely new paradigm for solving sequential decision problems. Unlike simulation optimisation, RL intelligences learn strategies directly through online interaction with the environment. Sutton and Barto (1998) defines the core problem of RL as learning a strategy $\pi(a \mid s)$ that maximises cumulative expected returns:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{3}$$

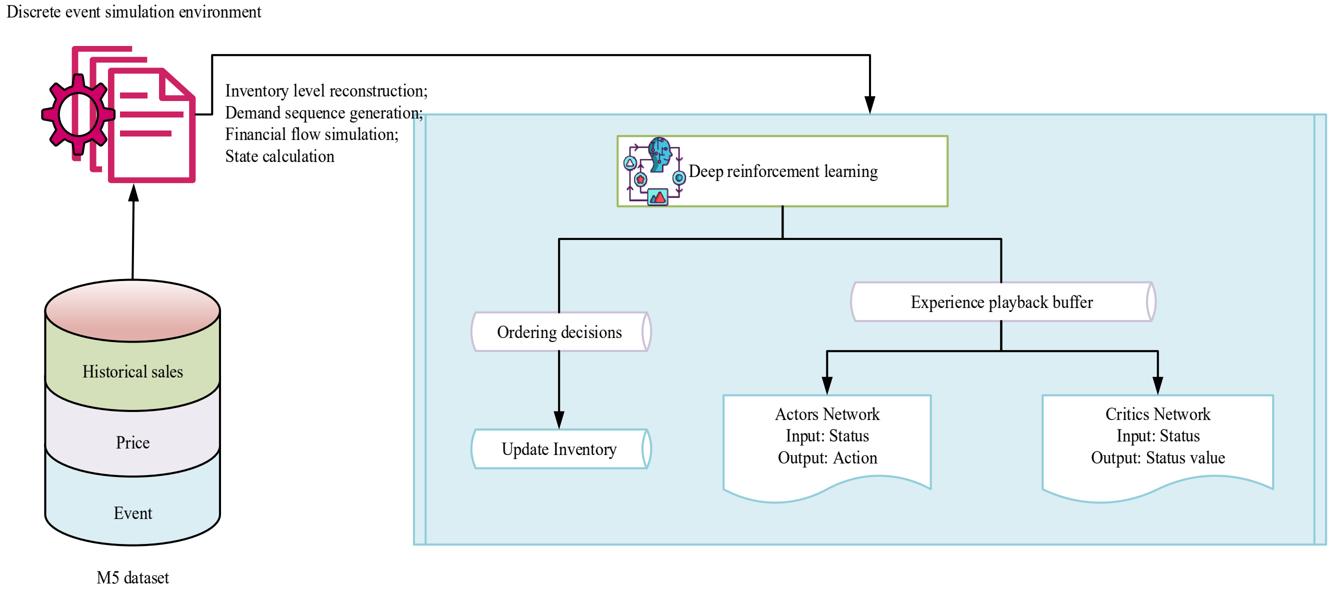
Where R_t is the reward obtained at moment t. The landmark work DQN of Mnih et al. (2015) achieves learning strategies from high-dimensional perceptual inputs to superhuman levels in several domains by approximating the Q-value function $Q(s, a; \theta)$ with a deep neural network.

This breakthrough was quickly introduced to the field of inventory management. Oroojlooyjadid et al. (2022) applied DQN to a multilevel ‘beer game’ supply chain with a reward function designed as:

$$R_t = -\left(h \cdot I_t^+ + p \cdot I_t^- + K \cdot \delta\left(a_t\right)\right) \tag{4}$$

where a_t is the ordering action, demonstrating the ability of DRL to outperform traditional heuristics in complex networks. Selukar et al. (2022), on the other hand, addresses the inventory problem of perishables and proposes to optimise their inventory management using DRL techniques that model retailers’ inventory constraints and real parameters such as expiration dates and shortages, and simulations show that they can reduce the inventory cost and spoilage rate when the delivery time, lifecycle, and demand distribution are known. Stranieri et al. (2024) proposed a novel heuristic combining DRL (setting production lots) and multi-stage stochastic planning (making transportation decisions) for solving the inventory management problem in two echelons of different supply chains, which outperforms pure DRL in reducing the total cost and overcomes the limitations of multi-stage stochastic planning, and released a publicly available software environment that can simulate multiple supply chain settings. We also release a public software environment that can simulate multiple supply chain settings and experimentally validate the effectiveness of the method.

Figure 1 DRL research topic evolution networks in supply chain simulation (see online version for colours)



However, there is a general and critical limitation of the current research. Kober et al. (2013) has emphasised that the design of the reward function directly determines the orientation and value of the strategies learned by the agent. Throughout the existing literature, the vast majority of studies follow a similar paradigm to Oroojlooyjadid et al. (2022) and Buzacott and Zhang (2004), narrowly defining the reward function as a linear combination of operating costs (holding costs, out-of-stock costs, ordering costs). This design completely ignores the profound impact of inventory decisions on a firm's financial liquidity. The evolutionary network of DRL research topics in supply chain simulation shown in Figure 1 clearly indicates that current research is mainly focused on the field of operational cost optimisation, while the attention to the topic of financial synergy is obviously insufficient, which provides a clear direction for the innovation of this study.

Richards and Laughlin (1980) defines CCC as:

$$\mathrm{CCC} = \mathrm{DIO} + \mathrm{DSO} - \mathrm{DPO} \tag{5}$$

where DIO is days, days of inventory outstanding, DSO is days sales outstanding, and DPO is Days Payable Outstanding. A strategy that only optimises operating costs may result in an inappropriately long DIO, freezing large amounts of cash and damaging the value of the business.

While research at the intersection of operations and finance has long existed, Buzacott and Zhang (2004) studied inventory decisions under asset finance constraints and Kouvelis and Zhao (2012) analysed the value of financing under trade credit contracts, much of this work has focused on joint modelling at the strategic or tactical level. At the micro-decision level of DRL, the work of Hubbs et al. (2020) fails to incorporate dynamic cash flow turnover into the learning objective, although inventory costs are considered. Moreover, achieving true integration of operations and finance in a dynamic learning framework remains an open challenge. Therefore, the current DRL-driven inventory optimisation research is significantly one-sided in terms of reward function design and fails to guide the intelligent agent to learn the globally optimal strategy that can synergistically improve operational efficiency and financial health, which leaves a clear scope for innovation in this study.

3 Methodology

3.1 *Finance-operations synergy optimisation problem formalisation*

This study models the supply chain inventory management problem as a sequential decision-making process, with the core lying in constructing an intelligent decision-making system that can simultaneously optimise operational efficiency and financial performance. We use the Markov decision process framework to formally describe this problem. At each decision-making moment t, the system state $s_t \in \mathcal{S}$ observed by the agent is a multivariate vector, which specifically contains the following key components: The current inventory level is I_t, the inventory in transit is O_t (indicating goods that have been ordered but have not yet arrived), the historical demand sequence of the past w periods is $\{D_{t-w},\dots,D_{t-1}\}$, the current cash position C_t, and the accounts receivable aging distribution $vecAR_t$ and accounts payable aging distribution $\overrightarrow{AP_t}$. Based on this state, the agent needs to select an action a_t, specifically the order quantity q_t. The action space can be discretised into multiple predefined levels or treated as a continuous variable.

The core innovation of this study lies in the design of the reward function, which breaks through the traditional limitation of considering only operating costs and introduces financial liquidity indicators. We design the synergistic reward function $R(s_t, a_t)$ in the following specific form:

$$R\left(s_t, a_t\right) = -\left[h \cdot I_t^+ + p \cdot I_t^- + K \cdot 1_{q_t>0}\right] - \lambda \cdot \mathrm{CCC}_t \tag{6}$$

where h denotes the unit cycle holding cost rate, $I_t^+ = \max(I_t, 0)$ is the end-of-cycle positive inventory, p is the unit shortage penalty cost, $I_t^- = \max(-I_t, 0)$ is the quantity of shortages, K is the fixed cost of ordering, and $1_{q_t>0}$ is an indicator function that equals 1 if an order is placed (i.e., $q_t > 0$), and 0 otherwise. λ is a key hyperparameter that weighs the relative importance of operating costs versus financial liquidity. CCC$_t$ is the CCC, which is calculated as:

$$\mathrm{CCC}_t = \mathrm{DIO}_t + \mathrm{DSO} - \mathrm{DPO} \tag{7}$$

where $DIO_t = \dfrac{I_t}{D_{avg}} \times T$ denotes the number of days inventory is held, where D_{avg} is the average daily demand, and T is the cycle length in days. The DSO and DPO are the days of accounts receivable and days of accounts payable, respectively, set based on industry standards or corporate policies. The reward function guides the intelligence to not only minimise traditional operating costs, but also proactively reduce the time cash is tied up, thus enabling dynamic adjustment of the balance between operational and financial objectives.

3.2 *Environmental reconstruction and preprocessing based on M5 data*

In order to validate our approach in a realistic scenario, we build a high-fidelity simulation environment based on the M5 forecasting-accuracy dataset provided by Wal-Mart. The dataset contains daily sales data, selling price information, and holiday events for 3,049 items in 10 Walmart stores in three US states over a period of 1,941 days. Since the dataset does not provide inventory levels directly, we reconstruct the inventory series using a backward imputation technique. Specifically, for each item-store combination, we back-calculate historical inventory levels using the following recursive formula given the final inventory state:

$$I_{t-1} = I_t + S_t - q_{t-L} \tag{8}$$

where S_t is the sales volume on day t and q_{t-L} is the order quantity sent on day $t - L$ that arrives on day t after a lead time L. Initially, we assume that the ordering pattern prior to the simulation start point follows a simple demand-averaging strategy to initiate the computation. This approach has been extensively validated in the inventory analysis literature for its soundness.

In the data preprocessing phase, we performed the following key steps on the raw data to ensure data quality and model applicability: First, we cleaned the data and filled in the small amount of missing data using linear interpolation, removing records with negative or clearly anomalous sales records. Second, to address the problem of large variations in quantiles across items, we normalised the sales data and inventory data by scaling them to the [0, 1] interval. Finally, we performed feature engineering. In addition to the basic demand and inventory features, we constructed time-series features (e.g., the mean and standard deviation of demand within a sliding window), event features (e.g., holiday signs), and interaction features (e.g., price elasticity coefficients). These features, especially holiday markers, allow the simulation environment to capture typical patterns of demand fluctuations triggered by promotions and holidays. Together, these features form the state vector s_t of the DRL intelligence's observation environment, providing it with a comprehensive and informative basis for decision-making.

3.3 DRL algorithm design

To address the properties of this problem-continuous or high-dimensional discrete action spaces, and the need to learn complex strategies from high-dimensional states-we adopt the proximal policy optimisation (PPO) algorithm as the cornerstone of our approach. PPO belongs to the actor-critic architecture of policy gradient methods, which is known for its excellent training stability and sample efficiency (Schulman et al., 2017). The core idea of PPO is to limit the difference between the new strategy and the old one by a cropped agent objective function at each strategy update, thus avoiding destructive huge update step sizes.

The objective function of PPO $L^{\text{CLIP}}(\theta)$ consists of three components: a trimming term for strategy loss, a value function error term, and an entropy reward term that encourages exploration. The core clipped surrogate objective is defined as the average over a batch of sample:

$$L^{\text{CLIP}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \min\left(r_t^{(i)}(\theta)\hat{A}_t^{(i)}, \text{clip}\begin{pmatrix} r_t^{(i)}(\theta), \\ 1-\epsilon, 1+\epsilon \end{pmatrix} \hat{A}_t^{(i)} \right) \tag{9}$$

where N is the batch size, $r_t(\theta) = \frac{\pi\theta\left(a_t \middle| s_t\right)}{\pi\theta_{\text{old}}\left(a_t \middle| s_t\right)}$ is the probability ratio between new and old policies, $\hat{A}_t$ is the estimated advantage function at time t, and ϵ is a hyperparameter (typically 0.1 or 0.2) that defines the clipping range (Schulman et al., 2015).

Our network architecture contains two main components: an actor network that determines actions, and a value estimation network that evaluates states. The actor network $\pi_\theta(a_t \mid s_t)$ takes as input the state s_t and outputs either the probability distribution (for discrete actions) or the mean and standard deviation (for continuous actions) of the action a_t (order quantity). The value estimation network also takes s_t as input and outputs a scalar estimate of the long-term cumulative payoff. Both networks consist of fully connected layers with ReLU activation functions. We use the Adam optimiser to update the network parameters and employ a discount factor γ (usually 0.99) to compute cumulative return.

3.4 Integrated simulation-optimisation framework

Figure 2 shows the overall architecture of the simulation-DRL integrated optimisation framework constructed in this paper. The framework realises the landing of the method through a tightly coupled simulation-optimisation closed-loop system. Specifically, first we construct a supply chain inventory simulation environment based on Python environment using SimPy library. This environment accurately simulates a series of key events such as demand arrivals, inventory checks, sales fulfilment, out-of-stock records, replenishment order issuance (consider a stochastic lead time to model realistic uncertain transportation delays), and cash flow updates (including accounts receivable generation and collection, and accounts payable payments). The overall architecture of our proposed simulation-DRL integrated optimisation framework is illustrated in Figure 3.

Figure 2 Simulation-DRL integrated optimisation framework (see online version for colours)

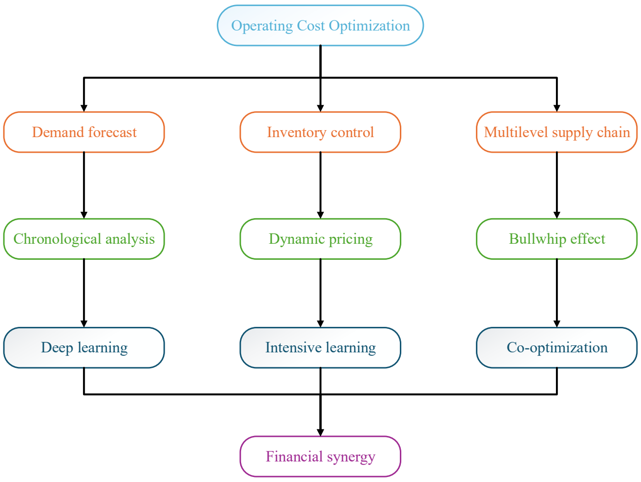
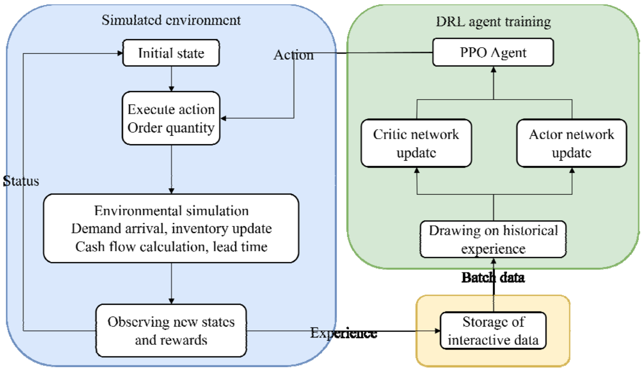


Figure 3 Integrated simulation-reinforcement learning workflow (see online version for colours)



At the end of each simulation step (e.g., one day), the environment passes its current state s_t (containing inventory, demand history, cash flow data, etc.) to the PPO intelligent body. The actor network of the intelligent body decides to order action a_t based on s_t. Subsequently, the action a_t is passed back to the simulation environment and executed, where the environment evolves to the next state s_{t+1} according to its intrinsic logic and computes the immediate reward r_t (i.e., the value of the co-reward function we designed). This interaction tuple (s_t, a_t, r_t, s_{t+1}) is stored in an experience replay buffer.

During training, the PPO algorithm periodically samples a small batch of experience tuples from the experience playback buffer. The critic network uses this data to learn more accurate estimates of state values, updated by minimising the mean-square time-series difference error. The actor network, in turn, uses the PPO's trimmed objective function and the estimated dominance function to update its strategy so that it tends to select actions that yield higher long-term cumulative rewards. Through thousands of such interactions and iterations, the agent eventually learns a complex, adaptive inventory strategy. This strategy not only responds to fluctuations in demand, but also proactively manages inventory to optimise overall operational and financial performance, enabling end-to-end learning from data to intelligent decision making.

3.5 Validation and verification of simulation models

To ensure the reliability of the simulation environment, we adhered to the principles of validation and verification. For validation, we employed modular testing and traced the logic of critical events – such as order arrivals, inventory updates, and cash flow recording – to confirm the code implemented the predefined business rules. For validation, we employed benchmarking methodology. Specifically, we implemented a classical (*s*, *S*) inventory strategy within the simulation environment and compared its outputs under steady-state demand scenarios – such as average inventory holdings and periodic service levels – against theoretical calculations. Results demonstrate that simulation outputs align closely with theoretical expectations, with errors falling within acceptable statistical bounds (e.g., relative error in service level <2%). This process confirms the fidelity of our simulation model in replicating real inventory system dynamics, providing a reliable foundation for subsequent DRL agent learning within this environment.

4 Experimental validation

4.1 Experimental setup

In order to comprehensively assess the performance of the proposed financial-operational co-optimisation framework, we conducted an in-depth empirical study based on the M5 forecasting-accuracy dataset. The experimental data covers the complete sales records of 5 Walmart stores in Wi State totalling 1,125 items for 1,941 days. We strictly follow the chronological principle in data partitioning, and use the data of the first 1,500 days as the training set to drive the simulation environment and train the DRL intelligences, and the data of the last 441 days as the test set to evaluate the generalisation performance of each method, which effectively avoids data traversal and ensures the rigor of the experiments.

We selected four representative baseline methods for performance comparison. The first is the classical (*s*, *S*) strategy (Iglehart, 1960), whose parameters are calibrated by statistical analysis of the training set data, and represents the baseline of traditional inventory control theory. The second is a rolling horizon-based model predictive control (MPC) approach, which integrates a lightweight long short-term memory (LSTM) demand prediction model to plan optimal actions within a finite number of steps based on the prediction at each decision point, reflecting the idea of combining prediction optimisation. The third is a deep recurrent Q-network (Oroojlooyjadid et al., 2022), denoted as DRL-Inv, which minimises only the inventory-related costs (holding costs vs. out-of-stock costs) as a reward function and represents the current frontier of DRL research focusing on operational efficiency only. The fourth is a hybrid approach integrating artificial neural network (ANN) forecasting and optimisation models (Makridakis et al., 2022), denoted ANN-Ensemble, which is one of the top-performing benchmarks in the M5 competition. Our proposed PPO algorithm incorporating a financial-operational synergistic reward function is denoted as cognitive load dynamic assessment MModel-PPO (CLDAM-PPO). All deep learning methods are implemented in the same computational environment using the PyTorch framework with full hyperparameter tuning. The key parameters are set as follows: unit holding cost rate h = 0.02/day, unit out-of-stock penalty p = 0.5, fixed ordering cost K = 1.0, CCC weight λ = 0.01, days of accounts receivable DSO = 15 days, and days of accounts payable DPO = 30 days, which are referenced to the common practice in retailing industry and the related academic literature. In order to simplify the experimental design and focus on the comparison of the core methodology, all commodities adopt this unified parameter set in this validation. It should be noted that our proposed framework is inherently flexible and fully supports the application of differentiated cost and financial parameters for different commodity classes.

The evaluation system consists of four core metrics in both operational and financial dimensions:

$$\text{Total Cost} = \sum \begin{pmatrix} \text{Holding Cost} + \text{Shortage Cost} \\ +\text{Ordering Cost} \end{pmatrix},$$

$$\text{Service Level} = 1 - \begin{pmatrix} \text{Total Shortage Quantity} \\ /\text{Total Demand} \end{pmatrix},$$

$$\text{Inventory Turnover} = \text{Total Sales} / \text{Average Inventory}$$

and

$$\text{CCC} = \text{DIO} + \text{DSO} - \text{DPO}.$$

To eliminate the effect of randomness, all experimental results are the average of five independent runs of each method using five different random seeds on the test set. means after independent runs with different random seeds and their standard deviations are reported. This multi-seeded experimental design was designed to assess the stability and reproducibility of the method performance, and the lower standard deviations observed indicate that the performance differences reported in this paper are stable and reliable. Statistical significance is assessed by two-sided t-tests and effect sizes are measured by Cohen's d-value.

4.2 *Comprehensive performance comparison and synergies*

The comprehensive performance comparison of each method on the test set is shown in Table 1. From the systematic assessment of key indicators, it can be found that the CLDAM-PPO framework proposed in this paper demonstrates significant advantages in the collaborative optimisation of operational efficiency and financial health. As shown in Table 1, in terms of the core indicator total related cost, CLDAM-PPO reached 285.4±8.7 (Unit: Thousand yuan) It was significantly better than all baseline methods, and significantly lower than 412.3 ± 15.2 of the (*s*, *S*) strategy, 365.8 ± 12.1 of the MPC method, 326.1 ± 9.9 of DRL-Inv and 348.5 ± 11.5 of ANN-Ensemble. Paired t-tests with the next-best-performing A paired t-test with the next best performer, DRL-Inv, showed that the performance enhancement was statistically highly significant (p-value < 0.01) and the effect size, Cohen's d = 0.85, belonged to a large effect level, indicating that the enhancement of the CLDAM-PPO was not only statistically significant, but also of great practical importance.

Figure 4 visualises the performance of each method on key metrics, showing the combined advantages of CLDAM-PPO on four core metrics: lowest total associated costs, maintaining a high level of service level comparable to DRL-Inv (98.2% vs. 98.5%, p > 0.05), achieving an inventory turnover of 6.5, significantly higher than the other methods, and reducing the cash-circulation cycle is shortened to 35.2 days. Of particular note, while DRL-Inv performs well in terms of pure operating cost control, its CCC is significantly worse than that of CLDAM-PPO at 41.5 days, a comparison that strongly demonstrates the necessity of including financial metrics in the optimisation objective and the effectiveness of our proposed synergistic reward function. The superiority of this model over classical strategies stems from its inherent adaptive and learning capabilities. Classical (*s*, *S*) strategies rely on static parameters; when confronted with the non-stationery and promotion-driven demand exhibited in the M5 dataset, their rigid decision rules lead to either excessively high inventory levels (resulting in soaring holding costs) or excessively low inventory levels (leading to increased stockout costs). Whilst MPC methods exhibit adaptability through rolling optimisation, their performance is severely constrained by the accuracy of demand forecasts within the planning horizon, with computational load increasing significantly as the planning scope expands. By contrast, our CLDAM-PPO framework employs end-to-end reinforcement learning to directly derive dynamic strategies from historical data and simulation interactions. This approach eliminates the need for predefined policy forms or reliance on precise short-term forecasts, enabling more flexible and robust responses to uncertainties in real-world retail environments.

Table 1 Performance comparison results of different inventory optimisation methods

Performance indicators	*(s, S) strategy*	*MPC Method*	*DRL-Inv*	*ANN-Ensemble*	*CLDAM-PPO*
Total associated costs (1000)	412.3±15.2	365.8±12.1	326.1±9.9	348.5±11.5	285.4±8.7
Service level (%)	95.1±1.8	96.8±1.2	98.5±0.6	97.9±0.9	98.2±0.5
Inventory turnover ratio	4.8±0.3	5.4±0.4	5.9±0.3	5.6±0.4	6.5±0.2
CCC (days)	48.2±2.1	45.6±1.8	41.5±1.5	43.8±1.6	35.2±1.2
Calculation time (minutes)	5.2±0.3	28.5±2.1	156.3±8.7	42.6±3.2	182.5±9.8

Figure 4 Performance comparison of different inventory optimisation methods (see online version for colours)

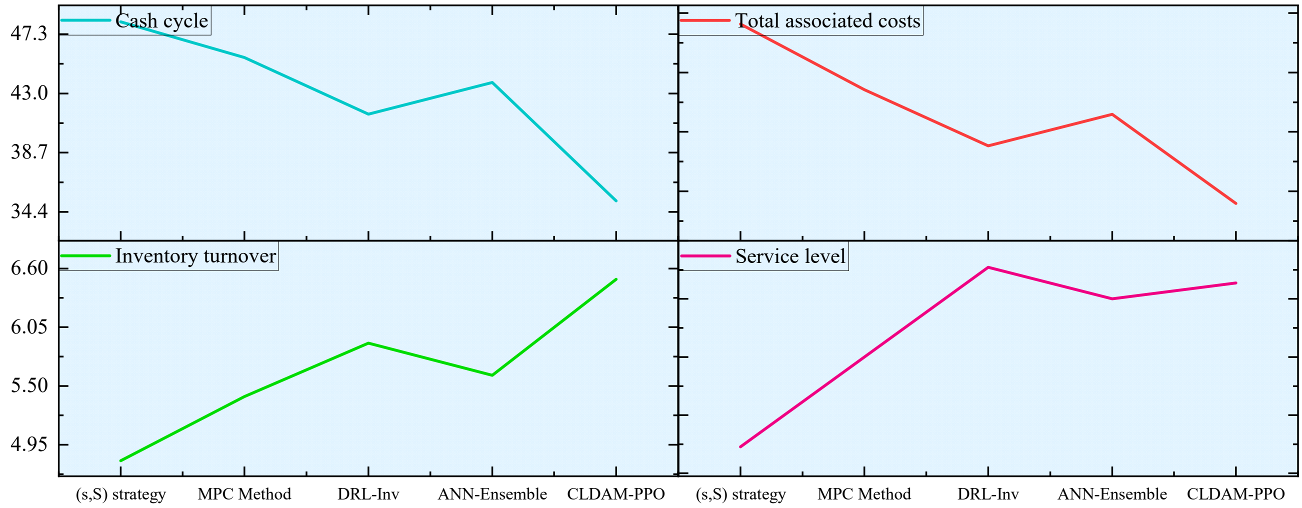
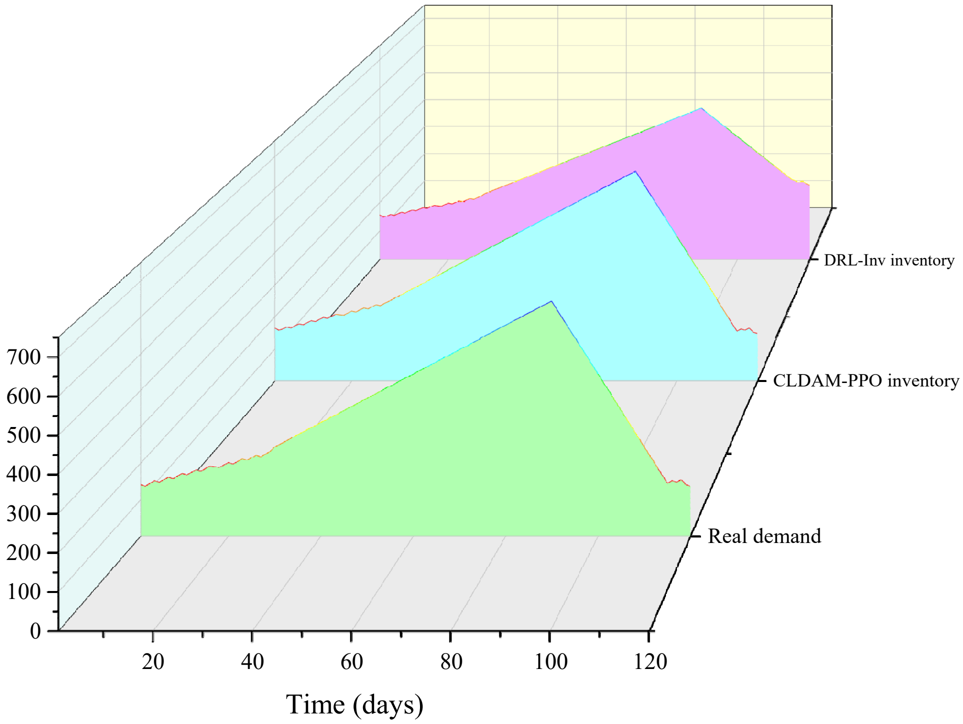


Figure 5 Inventory levels and demand dynamics during typical time periods (see online version for colours)



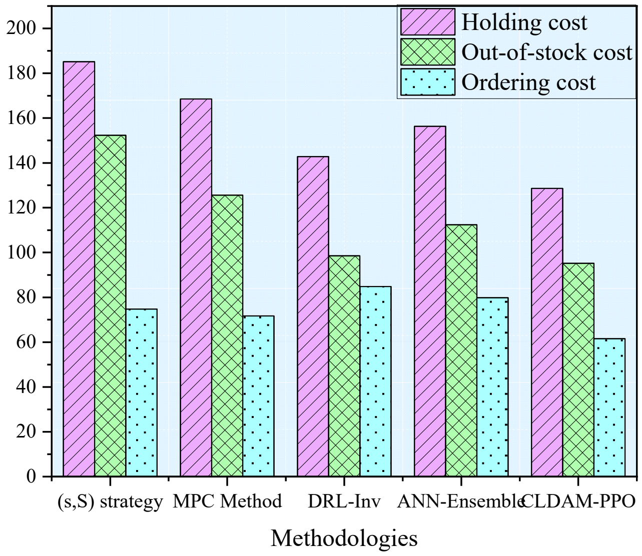
4.3 *Inventory dynamics and demand response behaviour analysis*

To gain insight into the dynamic behaviour of the CLDAM-PPO strategy, we selected a typical 120-day period in the test set containing seasonal peaks and sudden promotional events and plotted its inventory level versus demand time series plot against the best performing baseline method (DRL-Inv) (Figure 5). Analysis of this plot reveals that both DRL methods are able to respond more quickly than the traditional method by increasing order quantities in the face of a steep rise in demand. However, CLDAM-PPO shows more forward-looking inventory management wisdom: it builds up safety stocks more gently and in batches before the demand peak, instead of ordering large quantities urgently at the last minute, which helps to smooth the production and logistics pressure. After the demand peak, CLDAM-PPO also de-stocks faster, proactively reducing inventory levels to a more reasonable range and avoiding long-term capital tie-ups. This 'pre-build and fast depletion' model is inherent in its ability to achieve both high service levels and efficient inventory turnover.

4.4 *Explanation of cost structure decomposition and decision-making mechanism*

Figure 6 provides an insight into the differences in the cost structures of the different approaches, which breaks down the total associated costs into three components: holding costs, out-of-stock costs, and ordering costs. Analysing this figure reveals a key phenomenon: CLDAM-PPO succeeds in significantly reducing the inventory holding cost by moderately increasing the number of small-lot, high-frequency ordering (with a slight increase in the ordering cost), while avoiding a significant increase in the out-of-stock cost through intelligent inventory planning. In contrast, the (*s*, *S*) strategy produces the highest holding and stockout costs due to rigid parameters; the MPC approach improves on stockout control but still has high holding costs; and DRL-Inv, although it has lower total costs, has a cost structure that suggests that it reduces holding costs primarily by aggressively lowering inventory, a strategy that reduces the visible cost of operations but implicitly reduces the cost of stockouts due to the Increased risk of potential stock-outs due to low inventories and – more importantly – negative impact on cash flow efficiency due to failure to optimise order timing and lot sizes, which is explaining its poor performance on the CCC metrics.

Figure 6 Analysis of cost components of different methods (see online version for colours)



4.5 *Parameter sensitivity and strategy robustness validation*

To verify the model's adaptability under different parameter configurations, we systematically assessed the impact of key financial parameters on the performance of CLDAM-PPO. By varying the holding cost ratio h from 0.01 to 0.05, a steady linear growth trend (coefficient of determination $R^2 = 0.94$) is observed for the total associated cost, which rises from 272.5 in the baseline to 312.8, an increase of about 14.8%. It is noteworthy that the service level is consistently maintained within a narrow range of 97.5%-98.5% during this process, indicating that the model can reliably safeguard the service quality under cost pressure. In the sensitivity test of the CCC weight λ (variation range 0.001–0.05), the CCC is significantly shortened from 42.3 days to 29.7 days, a decrease of 29.8%, while the total cost only increases by about 11.7% accordingly. This asymmetric response relationship reveals the phenomenon of increasing marginal benefits from improved financial efficiency. Of particular interest is the fact that in all parameter combination tests, the changes in all performance indicators show smooth and predictable trends without any sudden performance changes or cliff drops, which fully demonstrates the strong robustness of the CLDAM-PPO strategy under different parameter environments. These findings provide decision makers with clear guidance for parameter tuning: by appropriately adjusting the λ value, companies can achieve significant cash flow improvement at a small cost price, thus realising a performance balance that matches their strategic goals. Additionally, we conducted a sensitivity analysis on the core DRL hyperparameters. When the discount factor γ increased from 0.90 to 0.99, the total cost exhibited a downward trend as the agent prioritised long-term returns. The learning rate demonstrated optimal stability and convergence speed at 5e-4; excessively high values (1e-3) caused training oscillations, while excessively low values (1e-4) resulted in slow convergence. Finally, we examined the impact of variance in the lead time. When variance increased by 50% and 100%, the total cost increase for CLDAM-PPO (<8%) remained substantially lower than that of the baseline method (>15%), demonstrating its robust resilience to supply uncertainty.

4.6 *Experimental results and analysis*

This study achieves important theoretical breakthroughs and practical insights in the field of supply chain inventory management by constructing a financial-operational synergistic optimisation framework. Experimental results show that our proposed CLDAM-PPO approach significantly reduces the total associated cost and shortens the CCC time while maintaining high service levels. This success mainly stems from the innovative design of the synergistic reward function, which guides the intelligent agent to simultaneously trade-off operational efficiency and financial health in the learning process. Unlike traditional DRL approaches that focus only on inventory-related costs, our framework internalises the CCC, a core financial metric, and motivates the intelligent agent to develop a more forward-looking inventory strategy-smoothly building up inventories before demand peaks, and quickly de-inventorying them after demand falls back. This behavioural pattern is highly consistent with Kouvelis and Zhao (2012) theoretical claims about the importance of working capital efficiency for supply chain value creation, but this study operationalises this idea for the first time in a dynamic learning framework.

At the theoretical level, the contributions of this study are mainly in three aspects. First, we successfully construct a theoretical-technical bridge between operations management and corporate finance by integrating the traditionally separated inventory turnover optimisation and cash flow management into a unified decision-making framework. This integration responds to Shi and Yu's (2013) call for research at the intersection of operations and finance, and provides a micro-decision-making foundation for the emerging field of 'supply chain finance'. Second, we extend the boundaries of DRL in supply chain management by demonstrating that with well-designed multi-objective reward functions, intelligences are able to learn complex strategies that go beyond single-dimensional optimisation. This finding provides strong support for Kober et al.'s (2013) argument that "reward function design determines the ability of an intelligent body" and provides lessons for solving other operational decision problems with multiple conflicting objectives. Finally, the simulation-learning closed-loop system proposed in this study provides a feasible technological path for organisations to achieve adaptive decision-making in highly uncertain environments, which is in line with the concept of building digital supply chain twins for resilience advocated by Ivanov and Dolgui (2021).

At the practical level, this study provides important insights for business managers. First, the findings confirm the necessity and feasibility of incorporating financial metrics into daily operational decisions. As Cannon (2008) points out, inventory improvement must be translated into financial performance improvement to be meaningful, and the CLDAM-PPO framework is an effective tool to realise this translation. Second, firms can use the method proposed in this study to adjust the weighting parameter (λ) of operating costs and financial liquidity in the reward function according to their own strategic priorities to achieve a customised inventory strategy that meets their needs. For example, capital-strapped firms can increase the value of λ to prioritise cash flow, while firms pursuing market share can appropriately decrease the value of λ to ensure service levels. In addition, the experimental results of this study based on real retail data enhance the credibility of the scheme in real-world scenarios and provide a reference case for enterprises to deploy similar intelligent decision-making systems.

However, there are several limitations to this study. First, there is a gap between the simplifying assumptions of the financial model for DPO and DSO and the dynamic nature of these parameters in reality. Future research could introduce dynamic financial parameter modelling by constructing DPO and DSO as a function of changes in supplier relationships, customer credit status, and market interest rates, or by using a stochastic process to simulate their volatility, in order to enhance the model's realistic adaptability. Second, the complexity and 'black-box' nature of the algorithms may affect their applicability to small and medium-sized enterprises (SMEs) and the transparency of their decisions. Improvements include the development of model compression techniques to reduce computational requirements, as well as the integration of attention mechanisms and counterfactual interpretation methods to provide an understandable basis for inventory decisions. Third, although the validation based on retail industry data is representative, its generalisability to other industries needs to be further tested. In the future, cross-domain validation should be conducted in industries with different inventory characteristics, such as manufacturing and pharmaceuticals, and domain-adaptive algorithms should be established to enhance the migration capability of the framework. These limitations provide a clear direction for future research. Finally, this study focuses on the financial and operational synergy of inventory, and the simulation environment does not explicitly model physical inventory losses such as spoilage and breakage of goods. In reality, such risks are usually indirectly included in the generalised inventory holding cost, and the economic impact is partially covered by the holding cost term in this model.

Recent research (2021–2024) has achieved further progress in integrating simulation with reinforcement learning, highlighting the potential of this paradigm. For instance, Stranieri et al. (2024) proposed an innovative heuristic approach combining DRL with multi-stage stochastic programming specifically for supply chain inventory management problems, demonstrating the hybrid method's potential to outperform pure DRL or pure optimisation approaches in complex environments. Within the broader domain of supply chain simulation optimisation, Gumte et al. (2021) developed data-driven robust optimisation models to handle uncertainty, illustrating the value of integrating simulation with advanced optimisation techniques. Concurrently, digital twins – as embodiments of high-fidelity simulation – have been employed to construct training environments for RL. Collectively, these works establish simulation environments as pivotal for training agents to navigate real-world complexity. However, whilst these studies have achieved significant results in optimising operational costs, as this research indicates, exploration remains limited in systematically integrating financial liquidity metrics into reward functions. The scope of most studies remains confined to traditional operational cost minimisation, failing to internalise the core impact of inventory decisions on corporate cash flow efficiency within learning objectives. Consequently, this research aims to address this critical gap by introducing the CCC into the synergistic reward function.

Future research can be carried out in the following directions, which will significantly enhance the practical application value of the research results: first, extend the framework to a multi-echelon supply chain environment, and study the mechanism of coordinating operational and financial objectives among different tiers. Second, it explores the impact of more complex financial factors, such as interest rate fluctuations and exchange rate risks, on inventory decisions. Once again, consider incorporating sustainability indicators into the optimisation goals to build a decision-making framework that is synergistic with the triple bottom line of operations, finance, and the environment. Finally, we will study how to deploy the framework in real enterprise systems, addressing implementation challenges such as data integration, computational efficiency and organisational change. The advancement of these research directions will jointly promote the transformation of supply chain management from the traditional empirical model to the new development paradigm of intelligence, value and sustainability.

5 Conclusions

This study proposes an innovative solution to the long-standing operational-financial decision-making disconnect in supply chain inventory management. By fusing discrete event simulation with DRL techniques and designing a synergistic reward function that integrates operational cost and financial liquidity, the CLDAM-PPO framework is developed to achieve synergistic decision-making in inventory management and cash flow optimisation. The experimental validation based on the M5 Forecasting Accuracy dataset shows that the framework not only outperforms traditional methods and existing DRL methods in traditional operational metrics (total associated cost, service level), but also demonstrates significant advantages in financial metrics (CCC period), confirming the effectiveness and advancement of the collaborative optimisation framework.

The theoretical contribution of this study is that it promotes the evolution of inventory management theory from pure operational optimisation to value creation orientation, and provides a new paradigm and methodological tool for the intersection of operations and finance. From a practical standpoint, this study provides actionable, intelligent decision-making solutions that enable enterprises to simultaneously enhance operational efficiency and financial health in dynamic market environments.

Declarations

Author declares no conflicts of interest.

References

Almeder, C., Preusser, M. and Hartl, R.F. (2009) 'Simulation and optimization of supply chains: alternative or complementary approaches?', *OR Spectrum*, Vol. 31, No. 1, pp.95–119.

Armenzoni, M., Montanari, R., Vignali, G., Bottani, E., Ferretti, G., Solari, F. and Rinaldi, M. (2015) 'An integrated approach for demand forecasting and inventory management optimisation of spare parts', *International Journal of Simulation and Process Modelling*, Vol. 10, No. 3, pp.233–240.

Arrow, K.J., Harris, T. and Marschak, J. (1951) 'Optimal inventory policy', *Econometrica: Journal of the Econometric Society*, Vol. 1, pp.250–272.

Bellman, R. (1966) 'Dynamic programming', *Science*, Vol. 153, No. 3731, pp.34–37.

Buzacott, J.A. and Zhang, R.Q. (2004) 'Inventory management with asset-based financing', *Management Science*, Vol. 50, No. 9, pp.1274–1292.

Cannon, A.R. (2008) 'Inventory improvement and financial performance', *International Journal of Production Economics*, Vol. 115, No. 2, pp.581–593.

Figueira, G. and Almada-Lobo, B. (2014) 'Hybrid simulation-optimization methods: a taxonomy and discussion', *Simulation Modelling Practice and Theory*, Vol. 46, pp.118–134.

Frésard, L. and Salva, C. (2010) 'The value of excess cash and corporate governance: Evidence from US cross-listings', *Journal of Financial Economics*, Vol. 98, No. 2, pp.359–384.

Fu, M.C. (1994) 'Optimization via simulation: A review', *Annals of Operations Research*, Vol. 53, No. 1, pp.199–247.

Goldberg, D.E. (1989) 'Genetic algorithms in search, optimization, and machine learning', *Addion Wesley*, Vol. 1989, No. 102, p.36.

Graves, S.C. (1999) 'A single-item inventory model for a nonstationary demand process', *Manufacturing & Service Operations Management*, Vol. 1, No. 1, pp.50–61.

Gumte, K.M., Pantula, P.D., Miriyala, S.S. and Mitra, K. (2021) 'Data driven robust optimization for handling uncertainty in supply chain planning models', *Chemical Engineering Science*, Vol. 246, p.116889.

Hubbs, C.D., Li, C., Sahinidis, N.V., Grossmann, I.E. and Wassick, J.M. (2020) 'A deep reinforcement learning approach for chemical production scheduling', *Computers & Chemical Engineering*, Vol. 141, p.106982.

Iglehart, D.L. (1960) 'Optimality of (s, S) policies in the infinite horizon dynamic inventory problem', *Management Science*, Vol. 9, pp.259–267.

Ivanov, D. and Dolgui, A. (2021) 'A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0', *Production Planning & Control*, Vol. 32, No. 9, pp.775–788.

Kober, J., Bagnell, J.A. and Peters, J. (2013) 'Reinforcement learning in robotics: a survey', *The International Journal of Robotics Research*, Vol. 32, No. 11, pp.1238–1274.

Kouvelis, P. and Zhao, W. (2012) 'Financing the newsvendor: Supplier vs. bank, and the structure of optimal trade credit contracts', *Operations Research*, Vol. 60, No. 3, pp.566–580.

Law, A.M., Kelton, W.D. and Kelton, W.D. (2007) *Simulation Modeling and Analysis*, Vol. 3, p. 1, Mcgraw-Hill, New York.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2022) 'The M5 competition: background, organization, and implementation', *International Journal of Forecasting*, Vol. 38, No. 4, pp.1325–1336.

Mauldin, M. (2017) 'Foundations of inventory management', *CABI Databases*, Vol. 1, p.1.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K. and Ostrovski, G. (2015) 'Human-level control through deep reinforcement learning', *Nature*, Vol. 518, No. 7540, pp.529–533.

Oroojlooyjadid, A., Nazari, M., Snyder, L.V. and Takáč, M. (2022) 'A deep q-network for the beer game: Deep reinforcement learning for inventory optimization', *Manufacturing & Service Operations Management*, Vol. 24, No. 1, pp.285–304.

Richards, V.D. and Laughlin, E.J. (1980) 'A cash conversion cycle approach to liquidity analysis', *Financial Management*, Vol. 9, pp.32–38.

Schulman, J., Moritz, P., Levine, S., Jordan, M. and Abbeel, P. (2015) 'High-dimensional continuous control using generalized advantage estimation', *Learning Representations*, Vol. 6, p.2438.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017) 'Proximal policy optimization algorithms', *Advances in Neural Information Processing Systems*, Vol. 30, pp.1190–1200.

Selukar, M., Jain, P. and Kumar, T. (2022) 'Inventory control of multiple perishable goods using deep reinforcement learning for sustainable environment', *Sustainable Energy Technologies and Assessments*, Vol. 52, p.102038.

Shi, M. and Yu, W. (2013) 'Supply chain management and financial performance: literature review and future directions', *International Journal of Operations & Production Management*, Vol. 33, No. 10, pp.1283–1317.

Silver, E.A., Pyke, D.F. and Peterson, R. (1998) *Inventory Management and Production Planning and Scheduling*, Vol. 3, p.30, Wiley, New York,.

Stranieri, F., Fadda, E. and Stella, F. (2024) 'Combining deep reinforcement learning and multi-stage stochastic programming to address the supply chain inventory management problem', *International Journal of Production Economics*, Vol. 268, p.109099.

Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning: An Introduction*, Vol. 1, No. 1, p.1, MIT Press.