Computers & Industrial Engineering 218 (2026) 112044

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

A cooperative digital twin–multi-agent reinforcement learning for circular supply chains: balanced control across production, logistics, and sustainability☆

Eduardo Guzmán [a,c,*], Beatriz Andrés [b], Marta Torres-Polo [a]

[a] Business Organization Department, Faculty of Economics and Business Administration, Autonomous University of Madrid 28049 Madrid, Spain
[b] Research Centre on Production Management and Engineering (CIGIP), Escuela Politécnica Superior de Alcoy, Universitat Politècnica de València, Alcoy, Spain
[c] Faculty of Business and Technology Sciences, Distance University of Madrid (UDIMA), Collado Villalba, Madrid 28400, Spain

ARTICLE INFO

Keywords:
Digital twin
Multi-agent reinforcement learning
Circular supply chain
Data-driven decision making
Simulation-based optimization

ABSTRACT

Coordination of production, inventory, logistics, and recovery decisions in circular supply chains (CSCs) remains challenging due to demand uncertainty, transport variability, and competing objectives across service, cost, and environmental dimensions. Controlled quantification of trade-offs and information value under matched experimental conditions has rarely been reported in prior digital twin (DT)-enabled control studies. To address this gap, an integrated methodological framework is proposed by coupling a simulation-based DT with cooperative multi-agent reinforcement learning (MARL). Specifically, a five-agent controller was trained to coordinate planning, inventory, logistics, expediting, and recycling decisions under a shared multi-objective reward within a DT–MARL formulation. The primary contribution is methodological: a controlled evaluation protocol with matched seeds, fixed horizons, and 95% confidence intervals is introduced to enable reproducible comparison across baselines, disruption scenarios, and sector archetypes. Moreover, a complementary Value-of-Data (VoD) protocol was used to isolate the marginal impact of cross-functional information integration on controller performance. In benchmark experiments, balanced improvements in lead time and on-time-in-full (OTIF) delivery were observed relative to a No-Op (no-action) baseline, while policy stability was maintained under transport, demand, and energy shocks. Furthermore, transferability was demonstrated across four archetypal operating regimes without retuning. Finally, VoD analysis indicated that integrated observation regimes shifted operating points toward improved resource efficiency.

1. Introduction

Contemporary circular supply chains (CSCs) face a dual imperative: striving for operational excellence, characterized by the optimization of costs, delivery times, and service levels, while simultaneously meeting strict sustainability goals, such as the reduction of waste and carbon dioxide (CO_2) emissions (Ouahabi et al., 2025). This inherent tension is significantly exacerbated within complex industrial sectors by widespread demand volatility, driven by mass personalization and shorter product life cycles (Kuo et al., 2025; Pan et al., 2021). Moreover, unforeseen logistical disruptions stemming from global events and geopolitical tensions (Ivanov, 2023), along with critical data fragmentation across functional silos, contribute to growing complexity in the industry (Liu et al., 2021). Such dynamics often result in operational information that is hysteretic and isolated, hindering agile and coordinated decision-making (Pan et al., 2021). In these dynamic environments, where information is fragmented, a notable gap emerges between the strategic goals of economic efficiency and environmental stewardship, making it necessary to adopt more intelligent and adaptive management paradigms (B. Peng, 2024).

CSCs differ from conventional linear supply chains in ways that fundamentally alter the operational control problem. First, product takeback and value recovery are introduced, making material flows bidirectional. This occurs because forward flows of sourcing, production, and distribution are coupled with reverse flows of collection, inspection, recovery, and reintegration (Guide & Van Wassenhove, 2009). Second,

☆ This article is part of a special issue entitled: 'The value of data in Industrial Engineering' published in Computers & Industrial Engineering.

* Corresponding author at: Avda. Francisco Tomás y Valiente, 5, 28049 Madrid, Spain.
E-mail addresses: brunnel.guzman@uam.es (E. Guzmán), bandres@cigip.upv.es (B. Andrés), marta.torres@uam.es (M. Torres-Polo).

https://doi.org/10.1016/j.cie.2026.112044

Available online 24 April 2026

additional uncertainty is introduced by the reverse stream, as the quantity, timing, and quality of returns are typically stochastic and only partially synchronized with primary demand and production planning (H. Peng et al., 2020). Third, the objective structure is expanded, since cost and service performance are required to be balanced against waste reduction, resource efficiency, and emission control. This, in turn, induces multi-objective trade-offs across production and logistics decisions (Govindan, Soleimani, et al., 2015). Fourth, recovery intensity and the allocation between virgin and recovered inputs are introduced as operational levers that interact with traditional planning, inventory, and transport decisions. As a result, an integrated control architecture is required so that these coupled decisions can be coordinated under heightened uncertainty.

In the face of such dynamism, traditional planning and optimization approaches, including local heuristics, deterministic optimization models, and fixed dispatching rules, are considered insufficient to effectively address the associated complexity (Kuo et al., 2025; Yan et al., 2022). These approaches were largely conceived for static and predictable environments and, as a result, fail to capture the critical interdependencies and stochastic nature inherent in CSCs (Yan et al., 2022). Their ability to adapt to sudden changes (such as new job insertions, equipment failures, or fluctuations in resource availability) is notably slow and often requires complete recalculation, rendering them unsuitable for real-time operational control (Krenczyk, 2024; Ouahabi et al., 2025; Tang et al., 2025).

To enable this level of coordination and address the computational complexity of such integrated systems, two powerful and converging technological streams have emerged: Digital Twins (DTs) (Talla & McIlwaine, 2024) and Reinforcement Learning (RL) (Ngwu et al., 2025). DTs are conceptualized as high-fidelity virtual representations of physical systems, which can be updated with real-time data from sources such as the Internet of Things (IoT) (Badakhshan et al., 2024). By creating a dynamic and synchronized mirror of the physical world, DTs function as simulated environments, tools for real-time monitoring, scenario analysis, and optimization throughout the product lifecycle (Pires et al., 2023). Their main role is to bridge the gap between physical and virtual spaces, enabling a holistic understanding of system behavior and supporting predictive and proactive decision-making (Bakhshi et al., 2024).

At the same time, RL, particularly in its deep (DRL) and multi-agent (MARL) variants, is widely recognized as a highly effective machine learning (ML) paradigm for solving complex sequential decision-making problems under uncertainty (Kreuzer et al., 2024). RL agents learn optimal control policies through direct trial-and-error interaction with an environment, making them exceptionally well-suited for dynamic optimization tasks such as production scheduling (del Real Torres et al., 2022), resource allocation (Schroer et al., 2025), and maintenance planning (Ouahabi et al., 2025; Yan et al., 2022).

Overall, significant opportunities to improve both operational efficiency and sustainability remain untapped, highlighting the need for methodological frameworks capable of holistic management of systemic uncertainty and dynamics without compromising sustainability. In this context, holistic is defined operationally along three dimensions: (i) integration across decision domains is represented by the simultaneous coordination of production planning and scheduling, inventory, logistics, expediting, and recycling actions rather than sequential optimization; (ii) integration across objectives is represented by the joint evaluation of operational, economic, and environmental key performance indicators (KPIs) within a single scalarized reward formulation instead of treating sustainability as a secondary constraint; and (iii) integration across information sources is represented by a consolidated KPI-based state representation in which production, inventory, logistics, and recovery signals are captured through a shared KPI vector, including lead time, service indicators, inventory and backlog exposure, transport time, cost measures, and emission measures.

From an Industrial Engineering perspective, these decision domains correspond to standard operational functions. Performance has been evaluated through standard operational metrics, including lead time, on-time-in-full (OTIF) and service level, inventory and backlog exposure, and cost efficiency, while environmental performance has been assessed through CO_2 emissions, energy-related measures, and material waste. Coordination across these processes has been targeted so that fragmentation between production, logistics, and sustainability objectives can be reduced.

The primary problem addressed in this study has been defined as the absence of an integrated, deployment-oriented control framework for circular supply chains in which cross-functional decisions are coordinated across production, inventory, logistics, expediting, and recycling under a multi-objective formulation that treats sustainability KPIs as first-class criteria. A review of the literature reveals several critical gaps that this research aims to address. First, a holistic and multi-objective integration of DT and MARL remains elusive (Ouahabi et al., 2025). Many existing frameworks are limited to specific tasks such as recommendation systems (Pires et al., 2023) or single-objective scheduling (Yan et al., 2022), rather than system-wide, multi-criteria optimization of production, logistics, and sustainability simultaneously. Furthermore, a significant number of studies refer to systems as DTs when they lack the essential bidirectional feedback loop, functioning instead as simpler digital shadows or models (Kreuzer et al., 2024; Liu et al., 2021). Second, controlled quantification of the Value of Data (VoD), understood as how information integration affects controller performance, has remained limited. Although simulation-based environments have been widely used when live operational data streams are unavailable or proprietary (Kreuzer et al., 2024), the marginal value of cross-functional observability has rarely been isolated through controlled comparisons under otherwise identical physics and objectives. Third, reproducible robustness evidence under realistic shock scenarios has remained scarce, but it has been strengthened here through structured transport, demand, and energy shock tests with matched seeds and confidence-interval reporting.

This paper addresses these gaps by proposing a data-driven framework in which a DT is synchronized with a cooperative MARL controller to optimize CSCs holistically. This framework is hereafter referred to as DT–MARL. A multi-agent formulation was adopted to provide a factorized representation of control across domain-aligned decision functions (planning, inventory, logistics, expediting, and recycling). While a centralized single-agent controller with a rich state representation can remain competitive in stylized settings such as the present single-commodity, two-echelon benchmark, factorized control with parameter sharing has been described as the intended pathway for scalability when the joint action space grows combinatorially in multi-commodity, multi-site, or deeper multi-echelon extensions. In addition, modularity and interpretability have been supported through agent-level ablations, which enable causal attribution of KPI shifts to specific action families under controlled removals. The DT consolidates material flows, production stages, transport logistics, economic costs, and energy and CO_2 emissions into an operational state expressed through industrial and sustainability KPIs. Within this environment, the MARL controller trains domain-aligned agents (planning, inventory, logistics, expediting, and recycling) to coordinate decisions and jointly manage competing objectives. The controller's behavior is shaped by a multi-objective reward function with non-negative weights for final evaluation, ensuring consistent behavior across scenarios.

The DT–MARL framework is evaluated against baselines, including no-action (No-Op), random, heuristic, and single-agent RL policies, under exogenous shocks to transport, demand, and energy (both individually and in combination), and across multiple sectors (construction, discrete manufacturing, process industries, and waste management). A VoD study compares an integrated-information configuration (Full-Data, hereafter Full) with an otherwise identical siloed-information configuration (Silo-Data, hereafter Silo), keeping physics, reward weights, normalizers, seeds, and evaluation horizons constant.

In Full, the controller observes a consolidated state vector that merges cross-functional signals: KPI snapshots together with stage utilizations, work-in-process (WIP), on-hand inventory by echelon, backlog flags, shipment queues and lane-level estimated-time-of-arrival (ETA) statistics, recycling-capacity utilization, and exogenous-shock indicators. In Silo, selected components of this vector are masked to emulate functional silos, while the underlying DT dynamics remain unchanged. VoD is defined as the marginal performance difference between Full and Silo, thereby attributing any improvement to information integration rather than to policy or environment changes.

In addition, ablation studies are conducted to identify causal mechanisms: (i) action-family ablations, in which one decision lever (planning, inventory, logistics, expediting, or recycling) is disabled while all other levers, DT physics, and evaluation settings remain constant; and (ii) reward-term ablations, in which a single term in the multi-objective reward (e.g., cost, CO_2, lead time) is set to zero while the remaining weights and normalizers are held fixed. Changes in KPIs under these controlled removals are interpreted as the marginal effect of the ablated component. Learning-curve diagnostics confirm within-episode stability. For clarity, the delivery reliability KPI "on-time-in-full" (OTIF) denotes the fraction of orders delivered within the due window and in the requested quantity.

The evaluation protocol fixes normalizers, random seeds, horizon length, and reward weights across all scenarios to preserve transitivity of effects. Results are reported as means with 95% confidence intervals; where variance exports are available, effect sizes are also computed. The primary outcome metrics include average lead time, OTIF, service level, operational cost, CO_2 emissions, total material waste, and energy consumption per unit.

The study is guided by four research questions (RQ):

RQ1: Can the proposed DT–MARL reduce lead time and improve OTIF while achieving favorable trade-offs across operational cost and environmental KPIs, including CO_2 emissions, material waste, and energy per unit, under a weighted multi-objective formulation, when compared to realistic baselines?

RQ2: What incremental value does integrated information provide when the objective remains unchanged, and does integration shift the operating point in a measurable and desirable way (VoD)?

RQ3: Are the learned policies robust to material shocks and transferable across sectors without retuning?

RQ4: Which action families and reward terms are determinative of the observed improvements, and how do they correspond to concrete operational levers?

The primary contribution of this study is methodological: an integrated DT-synchronized framework and a controlled evaluation protocol have been designed and validated for cooperative control in circular supply chains. This framework comprises (i) a DT–MARL architecture, in which a simulation-based digital twin is coupled with a five-agent cooperative controller to coordinate production, inventory, logistics, expediting, and recycling decisions under a fixed multi-objective reward; (ii) a controlled evaluation protocol, which enables reproducible comparison across baselines, shock scenarios, and sector archetypes through matched seeds, fixed horizons, consistent normalizers, and the reporting of 95% confidence intervals and effect sizes (Glass's Δ) where baseline variance permits; and (iii) a VoD protocol, which isolates the marginal impact of cross-functional information integration by contrasting Full-Data and Silo-Data observation regimes under otherwise identical conditions. The algorithmic component has been intentionally kept simple through established centralized training with decentralized execution (CTDE) principles with linear action–value functions and ε-greedy exploration, and no algorithmic novelty has been claimed. Empirical evidence has been provided for bounded trade-offs across service, cost, emissions, and waste, robustness under transport, demand, and energy shocks, transferability across sector archetypes without retuning, and causal attribution through agent-level and reward-term ablations.

The remainder of this paper is organized as follows: Section 2 reviews the related literature on DTs, RL, and CSCs, positioning the present contribution and clarifying unresolved gaps. Section 3 formalizes the problem and describes the DT–MARL framework in detail, including the state representation (industrial and sustainability KPIs), action families, multi-objective reward function, feasibility guards, and the training–execution scheme. Section 4 outlines the experimental design: the sector-parametric data generator, shock scenarios, cross-sector settings, baselines, normalization choices, seeds, horizons, and evaluation metrics. Section 5 presents the empirical findings: benchmark results, robustness under transport/demand/energy shocks, cross-sector transferability, the VoD comparison (Full vs. Silo), ablation studies, and within-episode stability diagnostics. Finally, Section 6 concludes with a synthesis of managerial implications, limitations, and directions for future research.

2. Literature review

Modern supply chain management (SCM) has faced complex, multifaceted challenges, which have been further intensified by the transition toward circular models (de Lima et al., 2021). A central dilemma has been balancing operational efficiency objectives with sustainability imperatives (Timperi et al., 2024). Manufacturing and logistics processes are under increasing pressure to reconcile productivity and cost goals with the sustainability of their actions, driven by stricter regulations, climate policies, and performance standards (Ciano et al., 2025). This duality is particularly challenging in uneven global markets, in which the pursuit of profitability often conflicts with sustainability considerations (Badakhshan et al., 2024).

While the primary objective of an SC has traditionally been to maximize total value generated, sustainability has emerged as a critical factor that was frequently overlooked by conventional frameworks (Dey et al., 2022).

2.1. Circular supply chains in uncertain and dynamic environments

The implementation of circular economy (CE) principles, such as the 10R strategies (Ciano et al., 2025), has been hindered by challenges related to economic feasibility, technological limitations, and regulatory barriers, thereby limiting large-scale adoption (Sajadieh & Noh, 2025). The inherent complexity of modern CSCs is reflected in high volatility, intricate interdependencies, and data fragmentation, creating an environment that is difficult to manage (de Lima et al., 2021).

Multiple interacting drivers, including growing demand for product customization, shorter product life cycles, rising product mix complexity, and dynamic production settings, have introduced substantial operational dynamics and uncertainty into production and logistics systems (Kuo et al., 2025; Y. H. Pan et al., 2021; Yuan et al., 2025). As a result, these systems are characterized as large-scale and complex networks comprising multiple operational stages (production, transportation, and storage) and diverse managerial layers (Y. H. Pan et al., 2021). Consequently, significant challenges have arisen for effective production planning and logistics management under supply chain uncertainty, including unpredictable job arrivals, equipment failures, fluctuations in customer demand (Tang et al., 2025), transportation unpredictability, and inventory volatility (Simard et al., 2023). Moreover, this uncertainty has been further exacerbated by process fragmentation, deficiencies in monitoring, insufficient optimization capabilities, and the absence of real-time bidirectional information exchange between assets and stakeholders (Li et al., 2024).

A lack of integrated data sources and real-time information delivery has also been documented in CSC environments (Yang et al., 2025). Moreover, existing production and logistics optimization methods have been proven insufficient to address the complexity and dynamism of CSCs, revealing a gap in their practical applicability (Chen et al., 2023). These methods are often based on deterministic or simplistic stochastic

models that poorly adapt to the inherent uncertainty and operational variability of production and logistics systems (Kuo et al., 2025). Traditional mathematical scheduling typically assumes stable conditions, fixed resources, and predetermined task attributes, assumptions that rarely hold on the shop floor, resulting in suboptimal schedules and frequent rescheduling (Ngwu et al., 2025; Zhang et al., 2026). Logistics management practices such as just-in-time (JIT) and lean production are optimized within relatively static subsystems, concluding that production-logistics systems have not been holistically or dynamically optimized (Y. H. Pan et al., 2021). Both centralized optimization within a single model and sequential optimization of independent modules tend to fall short of global optima due to reduced degrees of freedom and limited optimization scope (Y. H. Pan et al., 2021). Additionally, the manual design of dispatching rules generally requires trial-and-error, substantial time and coding effort, and deep domain expertise (Ouahabi et al., 2025).

In line with the scalability limitations of exact optimization formulations discussed above, metaheuristic approaches have been widely adopted to address NP-hard closed-loop supply chain (CLSC) network design and reverse logistics problems at realistic scales. In sustainable CLSC design, an adapted imperialist competitive algorithm hybridized with variable neighborhood search (VNS) has been proposed and benchmarked against baseline methods, including hybrid variants based on genetic algorithms (GA) and simulated annealing (SA) (Devika et al., 2014). Under stochastic demand, a robust hybrid multi-objective electromagnetism-like metaheuristic combined with VNS has been developed for sustainable order allocation and network design, and it has been compared against the Non-dominated Sorting Genetic Algorithm II (NSGA-II) and multi-objective particle swarm optimization (MOPSO) (Govindan, Jafarian, et al., 2015). In addition, a multi-objective memetic algorithm coupling a GA with dynamic local search has been applied to integrated forward and reverse logistics network design (Pishvaee et al., 2010). Restart-augmented variants of modified differential evolution (MDE) and related evolutionary schemes have also been reported for CLSC network design, together with additional multi-objective evolutionary and swarm optimizers in sustainable CLSC settings (Chaharmahali et al., 2022; Madani et al., 2026).

These studies demonstrate that metaheuristics can effectively solve large-scale CLSC design problems, generate Pareto-optimal fronts for multi-objective trade-offs, and handle complex constraints including capacity limits and environmental regulations. However, metaheuristics face limitations when applied to operational control in dynamic environments (Rajwar et al., 2023): (i) they solve static problem instances and require complete re-optimization when conditions change; (ii) they produce solutions rather than decision policies that generalize across scenarios; (iii) optimization times of minutes to hours are incompatible with real-time operational adjustments; and (iv) they do not incorporate feedback from outcomes to improve future decisions. These limitations motivate RL as a complementary paradigm for operational control, where learned policies map states to actions in real time, generalize through function approximation, and improve through environment feedback.

2.2. *Digital twins and multi-agent reinforcement learning in production, logistics, and sustainability*

Key contributions on the convergence of DTs and RL in production, logistics, and sustainability are reviewed in this section. Across the surveyed literature, a clear trend has emerged: while success has often been demonstrated in localized operational tasks such as scheduling, a significant gap remains in the provision of holistic, end-to-end control frameworks in which sustainability is treated as a first-order objective.

Table 1
Overview of related literature on digital twin–enabled optimization and reinforcement learning for production and supply chain control.

Reference	Domain & Decision Scope	DT Role	Optimization or Control Method	Sustainability & Circularity	Validation & Experimental Scope
Schroer et al. (2025)	EV charging hub (single facility)	Simulation DT used as RL environment	SAC	Cost objective with energy system modeled; explicit circularity not modeled	Scalability and optimality-gap benchmarking reported
Mingorance et al. (2025)	Industrial plant operations	Monitoring & re-optimization	Metaheuristics (GA) + ML surrogates	CO_2/water included as KPIs; circularity not modeled	Self-adaptation under disturbances demonstrated
Zhu et al. (2025)	Port dry bulk terminal	Monitoring & assessment	Heuristic search (OptQuest)	Emissions & energy assessed; circularity not modeled	DT fidelity calibration & resilience assessment emphasized
Krenczyk (2024)	CPPS resource assignment	DES-based DT environment	DRL (A2C & PPO)	No explicit circularity; sustainability not primary	Algorithm comparison & parameter sensitivity reported
Pan et al. (2025)	Hanging workshop (FJSP)	Scheduling environment	DRL scheduling (PPO-based)	No explicit sustainability or circularity	Makespan improvements vs. dispatching rules reported
Yuan et al. (2025)	Welding line (Dynamic FJSP)	Proactive scheduling	Multi-agent PPO	No explicit circularity; energy-related discussion limited	Disturbance responsiveness evaluated in production-line
Siatras et al. (2024)	Bicycle industry production	Multi-agent system (MAS) coupled with DT modules	Hybrid (Math. prog + DRL + Heuristics)	No explicit sustainability; circularity not modeled	Throughput and utilization emphasized for short-term
Anwar et al. (2025)	Federated manufacturing control	Federated DT in DES	PPO-based RL + priority protocols	No explicit sustainability or circularity	Completion time & urgent-job delay improvements
Yan et al. (2022)	Job Shop (FJSP + Maintenance)	Deviation detection & trigger	Double-layer Q-learning (DLQL)	No explicit sustainability or circularity	Performance vs. metaheuristics benchmarks reported
Geng et al. (2025)	Workshop + transport	Rescheduling with trigger	MAPPO-MC (global critic)	No explicit sustainability or circularity	Stable responses to disturbances reported
Xu et al. (2023)	Edge-end computing (MEC)	Offline training and online execution	Actor-critic MARL (CTDE)	Not supply chain focused; sustainability not modeled	Latency and deadline satisfaction benchmarked
Gu et al. (2024)	Intelligent workshop + AGV	Virtual-physical agent structure	Hybrid IGP-PPO	Energy efficiency targeted; circularity not modeled	Generalizability under dynamic events reported
Xiao et al. (2025)	EV battery disassembly	Multi-scenario DT platform	Multi-agent dueling Double-DQN	Circular operation via disassembly; sustainability via operational proxies	Convergence & comparative performance reported
This Study	End-to-end circular supply chain	Simulation-based DT as a KPI-synchronized environment for controlled evaluation	Cooperative MARL (Linear-DQN, CTDE)	CO_2 & waste as first-class KPIs with explicit trade-offs	VoD, ablations, and cross-sector comparisons

To summarize, a cross-study synthesis is provided in Table 1, where the reviewed contributions are compared along Domain & Decision Scope, DT Role, Optimization or Control Method, Sustainability & Circularity, and Validation & Experimental Scope.

Schroer et al. (2025) presented a data-driven planning approach for managing electric-vehicle (EV) charging hubs using soft actor–critic (SAC) reinforcement learning within a DT. Convergence toward near-optimal solutions, scalability, and performance were benchmarked against Deep Q-Learning (DQN), Deep Deterministic Policy Gradient (DDPG), and a perfect-information mathematical program, revealing optimality gaps of 4–15% and scalability to facilities with up to 1,000 charging points. The problem was framed as a high-dimensional, dynamic, stochastic decision process. Although the study addressed asset-and-operations planning for a single hub (chargers, photovoltaic, battery, substation), the scope remained facility-level and data-driven through a detailed DT, rather than a network-logistics model or a full supply chain twin. The objectives focused on total cost minimization, with no explicit treatment of circularity and only limited sustainability assessment beyond energy-efficiency analysis.

Mingorance et al. (2025) proposed a DT-based methodology to dynamically optimize operational strategies and real-time maintenance outages in industrial plants, aiming to reduce costs, CO_2 emissions, and surplus. Plant KPIs were maximized under a rolling-horizon planning scheme enabling self-adaptation to internal and external disturbances. For optimization, the approach integrated metaheuristics (e.g., genetic algorithms – GA) and machine-learning models. While reinforcement learning has been explored for dynamic production planning, the proposed methodology relied on GAs and ML rather than RL due to convergence-time considerations.

Zhu et al. (2025) developed an integrated DT model for dry bulk cargo terminals, focused on biomass handling, to provide a holistic sustainability perspective on port operations, particularly regarding building energy use and related emissions. The scarcity of high-resolution data was addressed via synthetic data and heuristic search (OptQuest) to calibrate the DT, with the objective of minimizing deviations between simulation outputs and historical records to improve fidelity. Although the study demonstrated the utility of DTs for monitoring and assessing sustainability and resilience in port operations, simulation and heuristic calibration/optimization were employed instead of multi-agent reinforcement learning (MARL).

Krenczyk (2024) introduced a DT-based production system for cyber-physical production systems (CPPS) to derive strategies for assigning processes to production resources. The methodology integrated discrete-event simulation (DES) with DRL, employing Asynchronous Advantage Actor–Critic (A2C) and Proximal Policy Optimization (PPO); decision quality surpassed classical dispatching rules and heuristics.

Pan et al. (2025) implemented a DT- and DRL-based framework for scheduling in a hanging workshop, achieving reduced makespan in flexible job-shop scheduling (FJSP) relative to priority dispatching-rule (PDR) baselines.

Yuan et al. (2025) designed a dynamic DT-based scheduling mechanism for the dynamic flexible job-shop scheduling problem (DFJSP), using multi-agent PPO to handle scheduling in a welding line. The primary objectives were to minimize makespan and balance machine utilization. While potential benefits for energy-management contexts were noted, such as improved resource allocation, increased energy efficiency, and reduced carbon emissions, the validation remained confined to welding-line scheduling, with emphasis on traditional operational metrics (makespan, total tardiness, machine deviation). The DT was constructed for a specific production line, enabling proactive scheduling and rapid response to disturbances.

Siatras et al. (2024) presented a framework that integrates a multi-agent system with DTs to support production managers in scheduling within the bicycle industry. A production DT module based on DES was built for each department, and optimization combined mathematical programming, DRL, and heuristics. The objectives focused on daily throughput (minimizing makespan) and maximizing line utilization, with the DT serving as a quantification tool for short-term production system performance under-agent decisions.

Anwar et al. (2025) proposed a hybrid concurrency-control framework for federated digital twins in software-defined manufacturing systems, coupling a PPO-based deep RL scheduler with a dynamically updated Priority Ceiling Protocol (PCP) to mitigate priority inversion in SimPy-based discrete-event simulations. Compared to PCP, Priority Inheritance (PI), DRL-only, and other baselines, the method reduced total completion time by up to 24.27% and urgent-job delay by up to 6.65%, while virtually eliminating priority inversions. The study targets coordination across heterogeneous DTs (autonomous mobile robot fleets, stochastic arrivals) rather than end-to-end network logistics.

Yan et al. (2022) developed a DT-enabled dynamic scheduling approach that jointly addresses flexible job-shop scheduling and flexible preventive maintenance using a double-layer Q-learning (DLQL) algorithm. The DT continuously compares physical and virtual states to trigger rescheduling after disturbances (e.g., job insertions, breakdowns), while the DLQL learns machine and operation assignments in real time. Experiments conducted against two metaheuristics and single-layer Q-learning showed superior solution quality across standard benchmarks and disturbance scenarios.

Geng et al. (2025) introduced a DT-driven dynamic scheduling framework for discrete manufacturing workshops with transportation resource constraints. Their approach incorporates a rescheduling-trigger discrimination mechanism to avoid unnecessary plan churn and proposes MAPPO-MC (multi-agent PPO with a global critic plus per-agent critics) to balance global coordination and local specialization. Case studies show prompt and stable responses to disturbances, along with improved performance compared to PDR and DRL baselines.

Xu et al. (2023) addressed edge–end collaborative scheduling for heterogeneous tasks over multi-access edge computing, using a DT for offline centralized training and online decentralized execution. The problem was formulated as a multi-agent Markov Decision Process (MDP) with a compound latency/deadline reward. Their actor–critic multi-agent deep reinforcement learning (MADRL) algorithm minimized job completion time while meeting diverse task deadlines, outperforming typical benchmarks. Although network- and edge-centric, the DT/MARL pattern is transferable to cyber-physical production settings.

Gu et al. (2024) developed a multi-agent manufacturing architecture for intelligent workshops and a dynamic scheduling mechanism combining improved genetic programming (IGP) with PPO, forming the IGP-PPO approach. Within a DT-like virtual–physical agent structure, the method addresses flexible FJSP under limited automated guided vehicle (AGV) transport, targeting both makespan reduction and energy efficiency. Prototype experiments demonstrated the method's superiority and generalizability under dynamic event conditions.

Xiao et al. (2025) presented a multi-scenario DT platform for human–robot collaborative disassembly of electric vehicle batteries. Process logic was modeled using Dynamic Time Petri Nets, and task sequencing and allocation were optimized via a heterogeneous multi-agent dueling Double-DQN (MADDQN) algorithm. A case study involving a commercial EV battery pack demonstrated efficient, safe, and flexible human–robot collaboration (HRC) in disassembly planning under structural uncertainty.

2.2.1. Multi-agent systems in DT-enabled control: MARL, CTDE, and functional decomposition

Multi-agent reinforcement learning extends single-agent reinforcement learning to settings in which multiple decision-makers interact within a shared environment (Busoniu et al., 2008; Gronauer & Diepold, 2022). In particular, in production and supply chain settings, separate functional responsibilities, such as planning, inventory, logistics, expediting, and recovery, can be represented explicitly as coordinated learners whose actions jointly determine system outcomes. However, a

core difficulty is non-stationarity, since the environment perceived by each agent is altered as other agents update their policies during learning.

Coordination has commonly been structured through CTDE. Under this principle, training information can be centralized through access to the global state and a shared team reward, while execution remains decentralized because each agent selects actions from its own policy using only its permitted observations (Foerster et al., 2017; Lowe et al., 2017). In cooperative settings, a common reward is used to align incentives, ensuring that improvements made by any agent contribute to the joint objective (Oroojlooy & Hajinezhad, 2023). The DT–MARL formulation adopted in this study is specified as fully cooperative, as the same global multi-objective reward is provided to all agents, as defined in Eq. (7).

A key design choice in cooperative multi-agent control is decomposing the joint action space. In this framework, decisions have been factorized into domain-aligned action families that correspond to standard industrial engineering functions, namely planning, inventory, logistics, expediting, and recycling. Furthermore, interpretability has been supported because agent-level ablations attribute KPI shifts to specific decision domains under controlled removals (see Section 5.5). In addition, modularity has been supported because policy updates can be targeted at a single decision function without requiring a full redesign of the controller. Moreover, scalability has been supported because the architecture is structured to extend to multi-site or multi-commodity variants through action factorization and parameter sharing, rather than requiring explicit enumeration of a combinatorially large joint action space.

2.2.2. *Digital twins for disruption management and supply chain resilience*

A growing stream of research has investigated the application of digital twins to disruption management and resilience enhancement in supply chains. Ivanov, (2023) introduced the concept of an intelligent digital twin (iDT) for supply chain stress-testing and viability analysis, emphasizing the role of DTs in providing visibility across supply chain tiers and supporting proactive decision-making under uncertainty. Ivanov & Dolgui (2020) proposed a DT-based anomaly detection model in which deviations between physical operations and their digital representations are used to identify emerging disruptions before they propagate, thereby reducing the ripple effect across supply chain stages. Dolgui et al. (2017) provided a comprehensive review of the ripple effect, established foundational concepts for disruption management, and highlighted the need for integrated modeling approaches capable of capturing cross-tier dependencies. Additional contributions have examined DT-based impact analysis in food retail supply chains under demand uncertainty (Burgos & Ivanov, 2021) and resilience-oriented decision support combining predictive analytics with simulation-based methods, which align with DT capabilities in end-to-end disruption management (Li et al., 2024).

In the present study, disruption management is integrated within a cooperative control framework. While prior work has emphasized digital twins as tools for scenario analysis, anomaly detection, or post-hoc resilience assessment, disruption response has been embedded in the learned control policy in the proposed DT–MARL framework. The robustness experiments reported in Section 5.2 evaluate policy performance under transport, demand, and energy shocks, and directionally consistent trade-off behavior has been observed under the tested adverse conditions. The VoD analysis reported in Section 5.4 complements this evidence by quantifying the marginal contribution of cross-functional information integration to disruption response under otherwise matched conditions, a dimension that has been less frequently quantified in prior DT-enabled disruption management studies.

2.3. *Literature gaps and proposal novelty*

As summarized in Table 1, recurring patterns have been observed across DT-enabled optimization and learning-based control studies in production and supply chain settings. First, decision scope has predominantly been confined to isolated functions or local subsystems, while end-to-end coordination across production, inventory, logistics, expediting, and recycling has rarely been addressed within a single closed-loop controller. Second, the DT role has most often been limited to monitoring or to a simulation testbed, whereas bidirectional coupling for operational decision support has been less frequently operationalized. Third, optimization approaches have been heterogeneous, ranging from mathematical programming and metaheuristics to single-agent reinforcement learning and multi-agent formulations, but cooperative multi-agent reinforcement learning under a shared multi-objective formulation has rarely been reported for cross-functional supply chain control. Fourth, sustainability and circularity have typically been treated as secondary outcome metrics rather than as first-class optimization objectives with explicit KPIs. Fifth, validation designs have often relied on limited statistical control and narrow scenario coverage, while controlled quantification of information value and reproducible robustness evidence under demand, transport, and energy shocks has remained scarce. A fragmented landscape has therefore emerged, in which DT and RL capabilities are often demonstrated in isolation rather than being unified into a comprehensive control framework for circular supply chains. The primary problem can be stated as follows: no existing approach has been shown to provide simultaneous coordination of production, inventory, logistics, expediting, and recycling decisions, joint optimization of operational, economic, and environmental objectives with sustainability KPIs treated as first-class criteria, controlled quantification of the value of cross-functional information integration, and reproducible robustness evidence under realistic demand, transport, and energy shocks.

To address these gaps, a cooperative DT–MARL framework for circular supply chains is proposed. Joint coordination occurs across production, inventory, logistics (consolidation and transport mode), expediting, and recycling, while the operational and sustainability KPIs defined in Section 3.2.1 are treated as first-class criteria.

Multi-objective coordination is achieved through a shared reward with fixed normalizers, ensuring comparability across scenarios. Operational cost, CO_2 emissions, and material waste are optimized jointly through weighted scalarization, while the prioritization of service improvements and the limitation of economic and environmental externalities are encoded through the weight selection procedure, which has been designed to preserve at least 70% of the best observed lead time improvement. The VoD is quantified by comparing integrated (Full) versus siloed (Silo) observations under identical physical conditions and objectives. Ablation analyses, both by action family and reward term, are conducted to identify causal mechanisms.

Results are reported with 95% confidence intervals and, where variance exports permit, effect sizes (Glass's Δ) are provided. Learned policies are stress-tested under transport, demand, and energy shocks (individually and in combination) and across four sector archetypes: construction, discrete manufacturing, process industries, and waste management and recycling.

3. The DT–MARL framework

3.1. *Problem description*

This case addresses CSCs in which forward flows (sourcing, production, storage, distribution) and reverse flows (returns, scrap recovery, recycling) interact under daily operational uncertainty. The focal system represents a single-commodity, two-echelon supply chain consisting of a production site, a consolidation node for outbound shipments, and customer demand aggregated at regional sinks. A reverse loop returns recovered materials to upstream stages, subject to capacity and quality constraints. This configuration captures common patterns in construction, discrete manufacturing, process industries, and waste

management, while avoiding site-specific idiosyncrasies.

In this study, a simulation-based DT environment was used to instantiate the operational logic, stochastic processes, and KPI accounting described in Section 3.2.1. Live data streams from physical operations were not used; instead, synthetic trajectories were generated according to the distributional assumptions and parameter ranges reported in Tables 3 and 4. This design choice enabled controlled experimentation with matched seeds, fixed horizons, and consistent normalizers across all methods and scenarios. Validation on operational data streams, including parameter calibration from historical event logs and shadow-mode deployment, is identified as future work and is discussed in Section 6.

This configuration was used to preserve the core coupled interactions that drive circular supply chain control while keeping experiments reproducible. Forward and reverse flows were coupled through a recovery loop that feeds material back upstream under capacity constraints, and multi-objective trade-offs were induced because lead time, OTIF, cost, CO_2, and waste were jointly determined by production pacing, inventory positioning, shipment consolidation, expediting, and recycling intensity. Congestion effects and backlog propagation were represented through finite buffers and backordering, and exogenous shocks to demand, transport reliability, and energy cost were propagated through multiple KPIs to evaluate robustness. However, multi-site coordination, multi-commodity flows including material substitution, and shared-capacity competition were not represented in the present benchmark and are identified as extensions that can be addressed through state augmentation and agent replication with parameter sharing, while preserving the evaluation protocol.

Operational performance is shaped by two interdependent objectives. On the one hand, service quality must be sustained through short lead times, high on-time-in-full, and stable service levels to meet project schedules or delivery service-level agreements (SLAs). On the other hand, sustainability commitments require reductions in material waste and CO_2 emissions without disproportionate increases in operating cost. These objectives compete through the physical and informational couplings of the system: higher throughput can induce congestion and backlog; aggressive consolidation reduces transport emissions but may extend delivery times; expediting recovers service but increases cost and emissions; and recycling reduces net waste but is limited by recovery capacity and material suitability.

Decisions are taken on a daily cadence over fixed-length episodes used for evaluation. Operational levers include modest adjustments to production throughput or cycle time (e.g., shift patterns, minor line balancing, and resource assignment). Logistics levers include inventory targets, stock adjustments across echelons (including reorder toggles), shipment-consolidation thresholds, and dispatch cadence, and transport mode selection. Sustainability-related levers include the controlled use of expediting for time-critical loads and the intensity/frequency of recycling operations. These levers are routinely available to line managers and planners and allow for measurable trade-offs in time, cost, emissions, and waste.

Exogenous variability is driven by three dominant factors observed across industrial sectors. Demand fluctuates around a sector-specific baseline and may experience persistent uplifts (e.g., project ramps, seasonal campaigns). Transport reliability varies, with delays affecting both the mean and variability of lead times along key lanes. Energy

Table 2
KPI definitions and targets.

KPI	Symbol	Goal	Type	Description
Average Lead Time	L_t	↓	Operational	Mean time from order release to delivery
Service Level	S_t	↑	Operational	Fraction of demand served at t
OTIF	Ot	↑	Operational	Orders delivered on time and in full
Operational Cost	C_t	↓	Economic	Total cost per period
CO_2 Emissions	E_t	↓	Sustainability	Greenhouse Gas emissions (process + transport)
Material Waste	W_t	↓	Sustainability	Net scrap after recycling
Average Inventory	I_t	↓	Operational	Mean on-hand/WIP
Backlog	B_t	↓	Operational	Unserved demand outstanding
Cost per Unit	CPU_t	↓	Economic	C_t/Units served$_t$
Energy per Unit	EPU_t	↓	Sustainability	Energy$_t$/Units served$_t$

Table 3
Data-generator modules and parameterization used in the sector-parametric DT.

Generator Module	Description and Parameters
Demand	Daily demand is modeled as an over-dispersed count process: $D_t \sim \text{Poisson}(\lambda_{dem}(1+\delta_{\text{dem}}))$, $\lambda_{dem} = \lambda_{\text{sector}} \cdot u_\lambda$, $u_\lambda \sim \text{LogNormal}(0, \sigma_\lambda^2)$ where sector-specific means λ_{sector} combine with multiplicative dispersion u_λ. Parameter ranges (Table 4) are selected to yield baseline service 85–95% and utilization rates of 70–85% under No-Op, consistent with regimes reported in make-to-stock/make-to-order operations. Sensitivity to ±20% rescaling does not alter qualitative conclusions.
Production and scrap	Unit processing times follow $p \sim \text{LogNormal}\left(\mu_p, \sigma_p^2\right)$, with capacity $C_t = available_{time}/\mathbb{E}[p]$. Yield $y \in [0.92, 0.98]$ induces gross scrap $W_t^{\text{gross}} = (1-y)\cdot\text{input}_t$. An expedite action multiplicatively reduces the effective cycle time by $\rho_{exp} \in [1.1, 1.5]$, with surcharge c_{exp}.
Inventory and backlog	Stocks evolve by flow conservation. Period costs include holding $C_t^{\text{hold}} = h\,I_t$ and backlog $C_t^{\text{back}} = b\,B_t$, with sector-agnostic ranges $h \in [0.5, 1.5]\$/unit{\cdot}day$ and $b \in [1.5, 3.0]\$/unit{\cdot}day$.
Logistics and transport	Consolidation yields shipment size Q_t and a stochastic lead time: $L_t \sim \text{LogNormal}\left(\phi_{\text{tr}}\mu_L, (\phi_{\text{tr}}\sigma_L)^2\right)$, where ϕ_{tr} scales mean and spread (scenario shock). Transport cost is: $C_t^{\text{tr}} = c_{\text{km}}\cdot d\cdot Q_t + c_{\text{fix}}\cdot 1\{Q_t > 0\}$, with sectoral distances $d \in [50, 200]$ km and $c_{\text{km}} \in [0.04, 0.08]\$/unit{\cdot}km$. The expedite lever switches to a faster mode: $L_t^{ship} \leftarrow L_t^{ship}/\rho_{exp}, \rho_{exp} \in [1.1, 1.5]$ with an additional transport surcharge Δc_{tr}(shorter mean/variance at higher cost and CO_2).
Energy and CO_2.	Process energy is modeled as: $E_t^{\text{proc}} = \alpha_e\,\text{throughput}_t + \beta_e(kWh)$ with $\alpha_e \in [0.8, 1.5], \beta_e \in [100, 300]$. Under shock, the energy price is multiplied by κ_{en}. Total emissions are: $E_t = \underbrace{e_{\text{kWh}}\,E_t^{\text{proc}}}_{\text{process}} + \underbrace{e_{\text{tr}}\,d\,Q_t}_{\text{transport}}$, using mid-range life-cycle factors: $e_{\text{kWh}} \in [0.35, 0.6] kgCO_2/kWh$ and $e_{\text{tr}} \in [40, 90] gCO_2/unit{\cdot}km$.
Waste and recycling	A recovery fraction $r \in [0.2, 0.6]$ converts gross scrap to reusable material; net waste is $W_t = (1-r)\,W_t^{\text{gross}}$. Recovered material offsets future inputs (closed loop) subject to capacity/quality constraints. Recycling incurs a cost $c_{\text{rec}} \in [0.1, 0.3]\$/unit$ and saves $e_{\text{rec}} \in [0.2, 0.5] kgCO_2/unit$.

Table 4
Sector-archetype parametric ranges (daily values).

Sector	λ_{sector}	$d(km)$	$c_{\text{km}}(\$/unit{\cdot}km)$	$\mu L, \sigma L(baseline)$
Construction Sector	350–650	120–200	0.05–0.08	moderate, high variance
Discrete Manufacturing	700–1100	80–150	0.04–0.07	low–moderate
Process Industries	900–1200	50–120	0.04–0.06	low variance
Waste Management and Recycling	450–800	100–200	0.05–0.08	moderate

Exact numeric baselines are logged in the artifacts; sector ranges aim to place No-Op in 85–95% service and 70–85% utilization.

markets introduce price shocks that propagate to operating costs and, indirectly, to emissions profiles depending on the electricity mix. These stressors are represented by scenario multipliers (demand uplift, transport-delay factor, and energy-cost factor) that perturb baseline parameters for controlled stress testing and robustness assessment.

Data are available across functional silos but are not always integrated in practice. Accordingly, the case contrasts a Full-Data view (in which production, inventory, logistics, cost, energy, and recycling signals are consolidated) with a Silo-Data view, where selected features are withheld from the decision layer while the underlying physics remain identical. This contrast enables a quantitative estimate of the Value of Data: the marginal performance gain that stems from integrating cross-functional information under an unchanged objective.

The DT that instantiates the case translates raw operational inputs (including orders released and delivered, WIP levels, shipment queues, energy consumption, and recovery volumes) into a KPI-based state observed at the end of each day. The primary indicators monitored for managerial decision-making include: average lead time and OTIF (service); operational cost and cost per unit (economics); CO_2 emissions and energy per unit (sustainability); total material waste (circularity); and average inventory, WIP, and backlog (flow health). Baseline operating points fall within realistic regimes reported across make-to-stock and make-to-order contexts; conclusions are drawn from relative changes against a no-intervention reference to mitigate dependence on absolute scales.

Boundary conditions are imposed to maintain operational realism. Production and transport capacities are finite; inventories and flows are non-negative; service windows and dispatch calendars are respected; expediting is rate-limited and subject to budgetary or energy constraints; unmet demand is backordered; and recycling capacity bounds material recovery. Within these constraints, the study focuses on how small, frequent tactical decisions, coherently coordinated, can shift the system toward service and sustainability improvements without incurring excessive economic penalties.

3.2. Methodology

3.2.1. DT dynamics and state

The operational environment is instantiated by a DT that maintains a KPI-based state vector $s_t \in \mathbb{R}^d$ and exposes a simulation interface to a set of specialized agents indexed by $i \in M$ (planning and scheduling, inventory, logistics, expediting, and recycling). At decision epoch $t = 0, \cdots, H$, each agent selects an action $a_{i,t} \in \mathscr{A}_i$; actions are aggregated into the joint control $a_t = (a_{i,t})_{i \in M} \in \mathscr{A} \equiv \prod_i \mathscr{A}_i$.

The digital twin has been implemented as a discrete-time, step-based stochastic simulation environment with daily decision epochs. At each step, exogenous variables are sampled from the distributional assumptions reported in Table 3, and the system state is updated through five tightly coupled mechanisms:

(i) Production flow is modeled across parallel and series stages with capacity and cycle-time scaling. Yields and scrap are tracked, and rework/scrap feeds the recycling pool. Queueing effects are reflected through finite buffers limiting WIP.

(ii) Inventory is conserved via mass balance at each echelon. On-hand and WIP stocks are updated; backorders B_t accrue when demand exceeds availability; and holding/backorder costs are charged at the end of each period.

(iii) Logistics and transport respond to consolidation thresholds, dispatch cadence, and mode selection, which jointly determine shipment sizes and travel-time distributions. The scenario multiplier ϕ_{tr} scales transit-time parameters, whereas an optional expedite switch shortens mean and variance at the expense of surcharges and higher emissions.

(iv) Economics accumulate production and setup, holding, transport (by lane and mode), and expediting costs. The energy price is scaled by κ_{en}.

(v) Energy, CO_2, and waste are computed as activity-times emission factors for process and transport components. Net waste is obtained after recovery, with the recycling rate bounded by capacity.

The objective is to achieve multi-objective improvement across industrial and sustainability KPIs, with explicit trade-offs governed through scalarization. The system is modeled as a multi-agent MDP, with a stochastic state transition induced by the DT:

$$s_{t+1} \sim P(s_{t+1} | s_t, a_t; \sigma) \tag{1}$$

Here, σ denotes exogenous operating conditions (scenarios) defined as $\sigma = (\phi_{\mathrm{tr}}, \delta_{\mathrm{dem}}, \kappa_{\mathrm{en}})$, with transport-delay multiplier $\phi_{\mathrm{tr}} \geq 1$, demand uplift $\delta_{\mathrm{dem}} \geq 0$, and energy-cost multiplier $\kappa_{\mathrm{en}} \geq 1$. Each agent follows a policy $\pi_i(a|s)$. Operational constraints (non-negativity, production and transport capacity, service windows, expediting limits) are enforced through a safety and feasibility layer and soft penalization in the reward function.

KPIs and evaluation. Performance is assessed through a comprehensive set of KPIs computed at each step t of every episode. Lower-is-better metrics include: average lead time L_t, operational cost C_t, CO_2 emissions E_t, total material waste W_t, average inventory I_t, backlog B_t, cost per unit $\mathrm{CPU_t}$, and energy per unit $\mathrm{EPU_t}$. Higher-is-better metrics include: service level S_t and OTIF O_t.

The KPIs reported in this study are organized into operational, economic, and environmental performance dimensions (Brandenburg et al., 2014; Hassini et al., 2012). Operational KPIs, including lead time, OTIF and service level, inventory, and backlog, indicate standard production and logistics performance. Economic KPIs, including operational cost and cost intensity measures, represent resource efficiency. Environmental KPIs, including CO_2 emissions, energy related measures, and material waste, reflect sustainability requirements.

The action families are aligned with standard Industrial Engineering functions, as planning and scheduling actions are depicted through production pacing and cycle time compression; inventory actions are represented through stock positioning and replenishment regulation; logistics actions have been represented through consolidation and transport mode selection, expediting actions are shown through time–cost trade-offs in fulfillment; and recycling actions are characterized by recovery intensity under reverse flow constraints.

Let Q_t be the set of orders delivered at step t and $N_t = |Q_t|$. Then:

$$L_t = \frac{1}{N_t} \sum_{o \in Q_t} (\mathrm{delivery}(o) - \mathrm{release}(o)), (Nt > 0) \tag{2}$$

$$S_t = \frac{\text{units served at } t}{\text{units demanded at } t}. \tag{3}$$

$$O_t = \frac{1}{N_t} \sum_{o \in Q_t} 1\{\text{on-time}(o) \wedge \text{in-full}(o)\}, (Nt > 0) \tag{4}$$

$$\mathrm{CPU_t} = \frac{C_t}{\mathrm{Unitsserved_t}} \tag{5}$$

$$\mathrm{EPU_t} = \frac{\mathrm{Energy_t}}{\mathrm{Unitsserved_t}}. \tag{6}$$

If $N_t = 0$ in a given period, L_t and O_t are either evaluated at the episode level or defined by convention to avoid division by zero; the same convention is applied consistently across all experiments. Episode-level KPIs are averaged over $t = 0, \cdots, H$, and then across random seeds. Results are reported with 95% confidence intervals; where variance exports are available, standardized contrasts (Glass's Δ relative to the baseline standard deviation, SD) are provided.

Reward and scalarization. To render the multi-objective problem tractable for RL algorithms, a weighted scalarization method is employed. This transforms the vector of objectives into a single scalar value. The instantaneous reward r_t is defined as the negative of a

weighted sum of normalized KPIs:

$$r_t = -\sum_{k\in K}(\omega_k \widehat{K_{k,t}}) + b_t, \qquad \widehat{K_{k,t}} = \frac{K_{k,t}}{K_k^{norm}}, K_k^{norm} > 0 \tag{7}$$

The sustainability-related KPIs were selected to capture complementary environmental dimensions that are commonly monitored in closed-loop and sustainable supply chain settings. CO_2 emissions serve as a measure of climate impact, material waste indicates resource circularity by quantifying unrecovered material leaving the system, and energy per unit reflects energy intensity and its operational coupling with both cost and emissions (Brandenburg et al., 2014; Hassini et al., 2012). Material waste is treated as a distinct objective because circularity performance is not solely inferred from cost and CO_2 alone, and recovery allocation policies can change waste outcomes even when cost and emission levels remain comparable (Govindan, Soleimani, et al., 2015). Accordingly, the reward in Eq. (7) is constructed through weighted scalarization so that operational, economic, and environmental KPIs are optimized jointly, while relative priorities have been encoded through the weight selection procedure described in Section 4.2.

Where K denotes the set of KPIs included in the reward function (see Table 2). An additive bonus term b_t is applied when specific targets are met. Concretely, the following triggers are evaluated each period: $S_t \geq 0.95$, $O_t \geq 0.95$, and schedule adherence ≥ 0.90, as well as low-waste and budget-favorable regimes; otherwise, $b_t = 0$.

Normalizers K_k^{norm} are estimated once from No-Op rollouts (averaged across seeds and periods) and then held fixed across all experiments (benchmark, shocks, cross-sector, VoD, ablations) to ensure comparability. The cooperative MARL objective maximizes the discounted episodic return:

$$\max_{\{\pi_i\}_{i\in M}} \mathbb{E}\left[\sum_{t=0}^{H-1}\gamma^t r_t\right], \qquad \gamma \in (0,1], \tag{8}$$

With $\gamma \approx 1$, the episodic return approximates average performance over time. Given that managerial service-level agreements and sustainability (ESG) priorities are context-dependent, the weight vector ω is calibrated through a systematic sweep. The selected set retains at least 70% of the best achievable improvement in average lead time within scenario, while minimizing the average $\Delta\text{Cost}\% + \Delta CO_2$. The tuned weights, used consistently across benchmark runs, stress tests, cross-sector evaluations, VoD analysis, and ablations studies, are: $\omega^\star = \{\omega_{co_2} = 1.1, \omega_{\text{cost}} = 1.2, \omega_{\text{inv}} = 0.5, \omega_{\text{lead}} = 0.7, \omega_{\text{waste}} = 1.2, \omega_{\text{backlog}} = 0.6, \omega_{\text{cpu}} = 1.0, \omega_{\text{epu}} = 0.6\}$.

Observations supplied to the agents combine KPI snapshots with operational context. A normalized KPI vector is concatenated with process-level features (e.g., stage utilizations, per-stage WIP, on-hand by echelon, backlog flags, shipment queues, lane-level ETA statistics, recycling-capacity utilization) and exogenous indicators ($\phi_{\text{tr}}, \delta_{\text{dem}}, \kappa_{\text{en}}$).

VoD experiment. The VoD is defined as the performance gain attributable solely to the integration of cross-functional information. A controlled A/B comparison is conducted between two observation regimes: Full-Data (Full) and Silo-Data (Silo). In Full, the controller receives a consolidated observation vector s_t that merges KPI snapshots with operational context (e.g., stage utilizations, per-stage WIP, on-hand by echelon, backlog flags, shipment queues and lane-level ETA statistics, recycling-capacity utilization, and exogenous-shock indicators). In the Silo, a subset of these components is withheld to emulate functional silos by applying a binary mask $m \in \{0,1\}^d$, resulting in $\tilde{s}_t = m \odot s_t$.

Crucially, the DT physics, reward function r_t, weight vector ω, normalizers, scenarios σ, random seeds, and time horizon are held constant across Full and Silo conditions. Therefore, any performance difference is attributable to information integration rather than changes in policy or environment. For each KPI K, the VoD metric is $VoD_K = \mathbb{E}[K]_{Full} - \mathbb{E}[K]_{Silo}$, computed over matched episodes. Under this definition, negative values are associated with improved performance under Full for lower-is-better KPIs (e.g., lead time, cost, CO_2, waste, CPU, EPU), whereas positive values are associated with improved performance under Full for higher-is-better KPIs (e.g., service level, OTIF).

Actions. Control is exerted through five disjoint action families whose Cartesian product forms the joint action a_t.

(i) Planning and Scheduling adjusts throughput and cycle time via a discrete multiplier $\{0.9, 1.0, 1.1\}$ lowering the multiplier to lengthen cycle time or raising it to increase capacity at potential energy/cost penalties.

(ii) Inventory toggles reorder policy and adjusts targets by $\pm\Delta$ around nominal levels, trading off stockouts/backlog and holding cost.

(iii) Logistics sets consolidation thresholds, dispatch cadence (e.g., daily vs. every two days), and transport mode (ground/air), thereby affecting shipment lead times, cost, and CO_2 emissions.

(iv) Expedite selects an urgency level $e \in \{0,1,2\}$ that reduces transit time and variance while incurring surcharges and additional emissions; and.

(v) Recycling tunes recovery intensity and frequency under capacity limits, increasing the recycling rate and reducing net waste.

From a control perspective, these action families influence the system's objectives through specific channels: planning affects lead time, service level, and OTIF (L, S, O) via capacity and WIP regulation; inventory influences (S, O, I, C) through stock positioning; logistics impacts (L, O, C, E) through consolidation and mode selection; expediting improves (L, O) at the expense of (C, E); and recycling reduces waste and emissions (W, E).

A safety and feasibility layer is applied before each DT step. Actions that would violate non-negativity or capacity constraints are projected onto the nearest feasible point. Let $\mathscr{A}^{\text{adm}} \subseteq \prod_i A_i$ denote the admissible set (capacity, non-negativity, service windows, expedite caps, optional budget/energy limits), defined via Euclidean projection:

$$\tilde{a}_t = \Pi_{\mathscr{A}^{\text{adm}}}(a_t) := \arg\min_{u\in\mathscr{A}^{\text{adm}}} \|u - a_t\|_2. \tag{9}$$

The period roll-out, including feasibility guards and reward computation, is summarized in Algorithm 1.

Algorithm 1: DT step with safety/feasibility guards

function DT-Step $(s_t, a_t, \sigma; DT_{params})$
1: # Feasibility layer
2: $\tilde{a}_t \leftarrow$ Project-To-Feasible $(a_t; A_{adm}, caps, service_{windows})$
3: if Expedite-Guard $(\Delta\widehat{O}_t, \theta_O) ==$ false then
4: $\quad \tilde{a}_t.exp \leftarrow$ Reduce-To-Last-Admissible $(\tilde{a}_t.exp)$
5: end if
6: # Advance DT modules under scenario σ
7: (prod, inv, log) ← Flow-Update $\left(s_t, \tilde{a}_t; \sigma\right)$
8: (cost, energy, CO_2, waste, backlog) ← Period-Accounting (prod, inv, log)
9: $KPIs_t \leftarrow$ Compute-KPIs(cost, energy, CO_2, waste, backlog, prod, inv, log)
10: $s_{t+1} \leftarrow$ Assemble-State $\left(KPIs_t, operational_features, exogenous(\sigma)\right)$
11: $r_t \leftarrow -\Sigma_k \omega_k \cdot Normalize\left(KPI_{k,t}; K_k^{norm}\right) + bonus_t$
12: return $(s_{t+1}, r_t, KPIs_t)$
end function

Policy gating for expedite. Expedite actions were permitted only when a predicted OTIF shortfall justified intervention. The shortfall was defined as: $\widehat{\Delta O_t} = \max\{0, O_{\text{target}} - \widehat{O}_{t+\Delta}\}$ where O_{target} denote the service target and $\widehat{O}_{t+\Delta}$ is the forecasted on-time-in-full rate at the next dispatch, estimated from lane-level ETA predictive distributions (e.g., quantiles of current transit-time posteriors) and due-window constraints for the orders in queue. An expedite level e_t was admissible only if $\widehat{\Delta O_t} \geq \theta_O$; otherwise e_t was reduced to the largest admissible value. Whenever the guard was binding, a small penalty term was added to the reward to discourage repeated reliance on guarded proposals.

3.2.2. MARL formulation and learning

Learning follows the CTDE paradigm. The principle and the rationale for functional decomposition of the decision space were summarized in Section 2.2.1, and the specific parameterization and update rules adopted in this study are detailed below.

Each agent i is parameterized by a linear action–value function:

$$Q_i(s,a;\theta_i) = \theta_i^{\top}\phi_i(s,a) \tag{10}$$

here $\phi_i(s,a) \in \mathbb{R}^{p_i}$ is a fixed feature map that concatenates: (i) normalized KPI snapshots and operational context available at t, and (ii) a one-hot encoding of the agent's action $a \in \mathscr{A}_i$. Features are standardized with fixed normalizers derived from the No-Op reference to ensure consistent scaling across scenarios. During evaluation, agents act greedily:

$$\pi_i(a|s) = 1\left[a \in \underset{a' \in \mathscr{A}_i}{\operatorname{argmax}} Q_i(s,a';\theta_i)\right] \tag{11}$$

Agents are trained as independent learners with a shared global reward: each agent treats the other agents as part of the environment but is updated on the same r_t. The objective is to maximize the discounted return $\mathbb{E}\left[\sum_{t=0}^{H-1}\gamma^t r_t\right]$, with $\gamma \in (0,1]$ chosen close to one to approximate average-reward behavior over the finite horizon.

Training proceeds with TD(0) updates. Let "terminal" indicate $t = H-1$. The Bellman target for agent i is

$$y_i \begin{cases} r_t, \textit{if terminal} \\ r_t + \gamma \max_{a' \in \mathscr{A}_i} Q_i(s_{t+1}, a';\theta_i), \textit{otherwise} \end{cases} \tag{12}$$

The TD error is $\delta_i \leftarrow y_i - Q_i(s_t, a_t^{(i)};\theta_i)$. With linear Q_i, the stochastic-gradient step is:

$$\theta_i \leftarrow \theta_i + \alpha\,\delta_i\,\nabla_{\theta_i} Q_i\left(s_t, a_t^{(i)};\theta_i\right) = \theta_i + \alpha\,\delta_i\,\phi_i\left(s_t, a_t^{(i)}\right) \tag{13}$$

where α is the learning rate. A small $\ell_2 r$ egularizer can be added to stabilize estimates if desired; it was not required given the bounded features and feasibility guards.

Exploration uses ε-greedy action selection with exponential decay per episode e :

$$\varepsilon_e = \max\{\varepsilon_{min},\ \varepsilon_0\,\lambda^{\,e}\}, \tag{14}$$

with $\varepsilon_0 \in (0,1]$, $\lambda \in (0,1)$, and $\varepsilon_{min} \in [0, \varepsilon_0)$. During evaluation and deployment, ε is set to zero.

Training is centralized in the sense that the full shared state s_t and global reward r_t are used to compute updates for all agents. Execution is decentralized: at runtime, each agent computes $a_t^{(i)}$ from the current state stream and its own Q_i, after which the joint action is passed through the feasibility layer described earlier.

The centralized-training, decentralized-execution update procedure is detailed in Algorithm 2.

Algorithm 2: MARL update step

function MARL-Update-Step (Q_{ii}, α, γ)
$s_t \leftarrow env.Get-State(\ )$
$features_t \leftarrow$ env.Get-Features(s_t)
for each agent $i \in M$ *do*
  $a_{i,t} \leftarrow \varepsilon_i - Greedy\left(Q_i\left(features_t^{(i)}, \cdot\right)\right)$
end for
$a_t \leftarrow (a_{1,t}, \cdots, a_{|M|,t})$
$\widetilde{a}_t \leftarrow Project-To-Feasible(a_t; \mathscr{A}^a dm)$
$(s_{t+1}, r_t, KPIs_t) \leftarrow DT-Step\left(s_t, \widetilde{a}_t, \sigma; DT_{params}\right)$
$features_{t+1} \leftarrow$ env.Get-Features(s_{t+1})
$done \leftarrow terminalcondition(e.g., t = H-1)$
for each agent $i \in M$ *do*
  if done then

(continued on next column)

(continued)

Algorithm 2: MARL update step

    $y_i \leftarrow r_t$
  else
    $y_i \leftarrow r_t + \gamma \cdot max_{a \in \mathscr{A}_i} Q_i(s_{t+1}, a; \theta_i)$
  end if
  $\delta_i \leftarrow y_i - Q_i(s_t, a_{i,t}; \theta_i)$
  $\theta_i \leftarrow \theta_i + \alpha \cdot \delta_i \cdot \varphi_i(s_t, a_{i,t})$
  ε_i decay is performed internally by each agent
end for
return(s_{t+1}, θ_{ii})
end function

3.2.3. Deployment protocol (inference, monitoring, fallback)

At deployment, exploration is disabled and only greedy actions are executed. The feasibility (safety) layer remains active, enforcing capacity, non-negativity, service-window, and expedite limits before the DT advances the system. KPIs are streamed from the DT and monitored online. Two supervision hooks are used: (i) Persistent-Guards, which fires when guard activations (e.g., expedite clipping) exceed a preset count within a sliding window; and (ii) KPI-Drift-Detected, which fires when monitored KPIs deviate from their validation bands beyond prespecified tolerances. When any hook remains active for multiple consecutive steps, a conservative heuristic fallback policy is triggered and an offline retraining job is scheduled on the updated data distributions.

From a computational perspective, per-step inference scales with the feature dimension across agents, i.e., $O\left(\sum_i p_i\right)$ for linear Q_i, and is therefore lightweight relative to training. Training is conducted offline; deployment executes only forward passes and feasibility projections.

The production inference loop with monitoring and fallback is detailed in Algorithm 3.

Algorithm 3: Deployed inference step (no exploration) with monitoring & fallback

function Inference-Step $(s_t, Q_{ii}, Guards; \theta_O, budget_{caps})$
for each agent $i \in M$ *do*
  $a_{i,t} \leftarrow argmax_{a \in \mathscr{A}_i} Q_i(s_t, a; \theta_i)$
end for
$a_t \leftarrow (a_{1,t}, \cdots, a_{|M|,t})$
$\widetilde{a}_t \leftarrow Project-To-Feasible\left(a_t; A_{adm}, budget_{caps}, service_{windows}\right)$
if Expedite-Guard $(\Delta\widehat{O}_t, \theta_O) ==$ false then
  $\widetilde{a}_t.exp \leftarrow Reduce-To-Last-Admissible\left(\widetilde{a}_t.exp\right)$
end if
$(s_{t+1}, r_t, KPIs_t) \leftarrow$ DT-Step$(s_t, \widetilde{a}_t, \sigma; DT_{params})$
Log $\left(KPIs_t, guard_{hits}, \widetilde{a}_t\right)$
if Persistent-Guards() or KPI-Drift-Detected() then
  Trigger(Heuristic-Fallback); Schedule(Offline-Retraining)
end if
return $\left(s_{t+1}, \widetilde{a}_t, KPIs_t\right)$
end function

Finally, the research questions translate into the following hypotheses, evaluated under the tuned weight vector ω^* and fixed normalizers:

H1: Relative to No-Op, Random, Heuristic, and Single-agent RL baselines, DT–MARL reduces expected lead time and material waste, increases OTIF, and keeps cost and CO_2 emissions within bounded ranges.

H2: For a fixed weight vector and identical physics and evaluation settings (reward weights, normalizers, seeds, and horizons), the Full-Data configuration improves OTIF and material waste and reduces operational cost and CO_2 emissions relative to the Silo-Data configuration, while a modest increase in lead time is accepted. H3: Under transport, demand, and energy shocks (individually and combined), DT–MARL preserves the signs of the benchmark effects and avoids severe degradations; analogous behavior holds across sector archetypes

without retuning.

H4: Action-family (planning, inventory, logistics, expedite, recycling) and reward-term (cost, CO_2, lead time, etc.) ablations induce distinct, interpretable changes in KPIs, consistent with their causal roles in the learned policy.

4. Experimental setup and baselines

This section details the data-generation pipeline, scenario design, and the training and evaluation protocol used to assess the proposed DT–MARL framework. The methodology was guided by three fundamental principles: (i) Control, ensuring that the DT exposes stressors and decision levers with traceable cause–effect relationships; (ii) Comparability, ensuring that all methods are exposed to identical stochastic realizations, reward weights, and normalizers; and (iii) Reproducibility, by fixing and reporting parameters, seeds, and logging artifacts to enable result replication.

The DT advances in daily steps over a finite horizon H (20 days unless otherwise stated), which defines one reinforcement learning episode. At each step: exogenous variables are sampled, the joint action is applied, the material/information/energy flows are updated, and KPIs are computed from the resulting state.

To isolate decision effects while maintaining generalizability, a single-commodity, multi-echelon network is modeled: a central production facility supplies a downstream consolidation node serving aggregated demand; a reverse loop recovers material subject to recycling capacity. Multi-SKU (Stock Keeping Units) complexity is approximated via demand overdispersion and queueing at finite buffers.

All performance is reported as relative deltas versus No-Op baseline, using fixed normalizers to mitigate dependence on absolute scales. The stochastic primitives and parameter ranges used by the sector-parametric DT are summarized in Table 3, which specifies the distributions, units, and engineering ranges adopted across all experiments.

The distributional families and parameter ranges reported in Tables 3 and 4 were chosen to represent discrete demand arrivals, lead-time uncertainty, cost asymmetries, emission factors, and recovery yields using standard modeling constructs. Demand is represented as an overdispersed count process, implemented as a Poisson baseline with LogNormal dispersion on the rate parameter, so that discrete arrivals and variance inflation due to clustering and batch ordering can be represented. A compound Poisson demand formulation has been analyzed extensively in order-up-to inventory systems and is used as an established reference for count-based demand modeling (Babai et al., 2011). Demand pattern heterogeneity across operating regimes is captured through the sector-varying rate parameter λ_{sector} reported in Table 4, and it is aligned with established demand pattern categories (Syntetos et al., 2005). Processing times and lead times are modeled as LogNormal so that strictly positive, right-skewed durations can be represented (Cobb et al., 2013; Tadikamalla, 1979). Holding cost $h \in [0.5, 1.5]\$/unit{\cdot}day$ and backorder cost $b \in [1.5, 3.0]\$/unit{\cdot}day$ were selected so that backlog penalties exceed holding costs, consistent with planned-backorder inventory modeling practice (Liberopoulos et al., 2010). The electricity emission factor $e_{\mathrm{kWh}} \in [0.35, 0.6] kgCO_2/kWh$ was chosen to represent mixed electricity grids as reported in European life-cycle assessments (Moro & Lonza, 2018). The transport emission factor e_{tr} was set to be consistent with road freight carbon intensity magnitudes discussed in the emissions measurement and green freight literature (Demir et al., 2014; McKinnon & Piecyk, 2009), and is expressed per shipment unit used in the digital twin. The recovery fraction $r \in [0.2, 0.6]$ was determined to represent effective recovery levels from baseline to moderately advanced circular practices as reported for construction and demolition waste contexts (Gálvez-Martos et al., 2018; Rondinel-Oviedo, 2021). Sector-specific ranges in Table 4, including demand rates λ_{sector}, distances d, and transport cost coefficients c_{km}, were selected through iterative operating-point screening so that baseline regimes under the No-Op policy remain non-degenerate and suitable for controlled benchmarking. Robustness was assessed through a $\pm$ 20% parameter perturbation analysis.

The use of a sector parametric digital twin is intended to support controlled stress testing under uncertainty when industrial traces are proprietary and not comparable across sites. Practical relevance is preserved by retaining structural mechanisms that are faced in applied settings, including finite capacity, stochastic lead times, demand uplift and transport delay shocks, expediting trade-offs, and coupled economic and environmental accounting. In addition, results are reported as normalized KPIs and deltas versus No-Op, so that conclusions are driven by relative performance and robustness patterns rather than by absolute scale.

4.1. Sector profiles and scenarios

Four archetypal operating regimes are considered: Construction (project-based operations and complex logistics), Discrete Manufacturing (high-volume assembly), Process Industries (continuous flow, such as chemicals and food), and Waste Management and Recycling (reverse logistics). These labels serve as shorthand profiles for distinct combinations of demand intensity and variability, lead time baseline and dispersion, transport distance and cost, and circularity maturity, rather than as exhaustive representations of entire industries. Thus, heterogeneity within categories, including within discrete manufacturing, is captured by instantiating the digital twin with different parameter vectors, while the same state, action, and reward structure is maintained. Sector and archetype-specific ranges for λ_{sector}, lead time parameters (μ_L, σ_L), distances d, and transport cost coefficients are listed in Table 4.

The proposed framework consists of two layers with different generalization properties. The control architecture, including the five-agent structure, the multi-objective reward formulation, feasibility guards, and the centralized training and decentralized execution protocol, is designed to be domain-agnostic, as the agents correspond to decision functions present in production distribution systems with reverse flows, and the coordination mechanism does not embed sector specific assumptions. The digital twin parameters, including demand intensity, lead time distributions, cost coefficients, and emission factors, are treated as context specific and are intended to be instantiated by selecting values from the archetypal ranges in Tables 3 and 4 for benchmarking, or by re-estimating from site-specific operational data for applied deployments. Construction was selected as the primary benchmark because high demand variability, complex logistics, and circularity potential are combined in that context, enabling stress testing across multiple performance dimensions. Evidence of transferability is provided by the cross-sector experiments reported in Section 5.3, where the policy trained on the construction benchmark is evaluated on the other archetypal regimes without retuning.

Operational stress is introduced via a scenario vector $\sigma = (\phi_{\mathrm{tr}}, \delta_{\mathrm{dem}}, \kappa_{\mathrm{en}})$. We evaluate: (i) transport delay multipliers $\phi_{tr} \in 1.0, 1.2, 1.3$; (ii) persistent demand uplift $\delta_{\mathrm{dem}} \in [0, 1]$ (main value: 0.5); (iii) energy-cost multipliers $\kappa_{\mathrm{en}} \in \{1.0, 1.3\}$; and (iv) a combined shock (1.2,0.4,1.2). These values are selected to induce meaningful but non-degenerate degradations (e.g., transport delays that increase lead times without fully saturating buffers), enabling robust and interpretable performance assessment.

The sector-specific ranges in Table 4 were established through a two-stage procedure. First, demand intensities, transport distances, and lead time parameters were aligned with magnitudes reported in sector-specific supply chain studies: construction supply chains are characterized by project-based demand variability and longer transport distances (Vrijhoef & Koskela, 2000), discrete manufacturing by higher throughput and shorter lead times (Cobb et al., 2013), process industries by continuous flow with lower variability (Fransoo & Rutten, 1994), and waste management by reverse logistics patterns (Govindan, Soleimani, et al., 2015). Second, within these bounds, operating-point screening

was conducted to ensure that baseline service levels under No-Op fall within the 85 to 95 percent range and capacity utilization within the 70 to 85 percent range, thereby avoiding degenerate regimes and enabling controlled benchmarking.

The archetypes in Table 4 are defined broadly to enable robustness assessment across regimes. Within each archetype, a specific subsector is represented through a consistent selection of parameter values across Tables 3 and 4, including λ_{sector}, μ_L, σ_L, d, and c_{km}, and when needed, cost and circularity parameters such as b, h, and r. For discrete manufacturing, a high-volume assembly context can be modeled through higher arrival intensity and shorter baseline lead times, while a low-volume heavy equipment context can be represented through lower arrival intensity and longer, more variable lead times, together with stronger backlog penalties relative to holding costs. No change is required in the DT–MARL control architecture for such instantiations, since only the digital twin parameters are adjusted to reflect the intended operating context.

4.2. Reward, normalizers, and weight selection

The scalar reward is the weighted sum in Eq. (7), with one normalizer per KPI and sector, computed from No-Op rollouts (averaged across seeds and periods) and subsequently fixed across methods, scenarios, and ablations. Weight selection follows a two-stage procedure: (i) train a lead-time–only controller to estimate the attainable gain ΔL^{max};(ii) perform a grid search over ω and select sets that retain $\geq 70\%$ of ΔL^{max} while minimizing $\Delta Cost\% + \Delta CO_2\%$ (ties are resolved by $\Delta Inventory\%$). The resulting $\omega^\star$ is held fixed across all experiments (benchmarking, shocks, cross-sector comparisons, VoD, ablations).

Hard bounds have not been imposed on cost or emissions, since the implemented formulation relies on joint optimization through weighted scalarization, with the weight selection procedure encoding practical priority toward service improvements. In the reported experiments, lead time and service performance were treated as primary operational improvement targets, while cost and CO_2 measures were controlled through the selected objective weights and normalization, and inventory and backlog exposure were monitored to ensure stable flow behavior, and expediting intensity and capacity utilization were reported to support operational interpretability of the learned policies.

4.3. Training protocol, seeds, and logging

Agents are trained under CTDE using Linear-DQN with online TD(0), Hyperparameters are kept within narrow ranges for stability: learning rate $\alpha \in [3 \times 10^{-4}, 3 \times 10^{-3}]$, discount factor $\gamma \in [0.95, 1)$, and ε –*greedy* exploration with $\varepsilon_0 \approx 0.2, \lambda \approx 0.999$, and $\varepsilon_{min} \in [0.02, 0.1]$.

Each block (benchmark, shocks, cross-sector, VoD, ablations) uses a fixed horizon $H = 20$ days and pre-declared seeds and episodes to support estimation of means and 95% confidence intervals: benchmark (construction) uses 30 seeds × 110 episodes; stress tests and cross-sector comparisons 20 seeds × 90 episodes; and ablations 15 seeds × 70 episodes. The same seed table is reused across methods within each block to ensure matched stochastic realizations (comparability).

Convergence is monitored with two diagnostics computed over the final 10% of episodes: (i) late-episode slope (least-squares slope of per-episode returns; target ≈0); and (ii) tail variability (interquartile range, IQR, of those returns; target: low).

All runs log the artifacts required for reproducibility and VoD analysis: per-sector normalizers, the selected weight vector $\omega^\star$, observation masks m (Full vs. Silo), guard thresholds (e.g., θ_O for expedite and any budget/energy caps), and the seed dictionary with episode indices.

4.4. Baselines

To isolate the contributions of decentralization, learning, and data integration, the following baselines were evaluated and are summarized in Table 5. The training and evaluation budget per block (benchmark, stress tests, cross-sector comparisons, ablations, and VoD) is specified in Section 4.3; seeds are reused across methods within each block.

4.5. Evaluation protocol and statistics

Per-episode KPIs are averaged over the H daily steps and then across seeds. Results are reported as means with 95% confidence intervals, computed from the seed-level averages. For lower-is-better metrics (lead time, cost, CO_2, waste, inventory, CPU, EPU), percentage changes relative to the No-Op baseline are reported; for higher-is-better metrics (service level, OTIF), changes in percentage points versus No-Op are shown.

Table 5
Baseline policies and control strategies.

Baseline	Strategy Description
No-Op	The joint action is fixed to a do-nothing policy $a_t \equiv a^{\varnothing}$: nominal release rates, default consolidation/dispatch cadence, normal transport mode, no expediting, and fixed recycling settings. This baseline provides the absolute reference and the per-sector normalizers used in the reward.
Random	At each decision epoch, each agent i samples uniformly from its discrete action set, $a_{i,t} \sim \mathrm{Unif}(\mathscr{A}_i)$; the joint action $a_t = (a_{1,t}, \cdots, a_{\lvert M\rvert,t})$ is then projected by the feasibility layer before being applied. This baseline preserves action availability and constraints while removing any learned structure.
Heuristic (rule-based)	A hand-crafted policy implements standard operations rules across the five decision domains: – Inventory (per echelon): (s, S) *control* The reorder point uses a normal-approximation safety stock: $s = \mu_D L^{pred} + z\sigma_D\sqrt{L^{pred}}$, where μ_D, σ_D are the mean and standard deviation of daily demand, and L^{pred} is the predicted lead time. The parameter z sets the target cycle-service level (*e.g.*, $z \approx 1.28$ for approximately 90% service). The order-up-to level is $S = s + Q$, with an EOQ-style increment $Q = \sqrt{2KD/h}$(setup/fixed cost K, demand rate D, holding cost h). Orders are placed when the inventory position $< s$, up to S, and capped by capacity. – Planning release rate (WIP control): A lot-sizing factor $\ell \in \{0.9, 1.0, 1.1\}$ scales nominal releases, subject to capacity c and a WIP capacity W^{max}. $release\ rate = \min\{\ell\mu_D, c, W^{max} - WIP_t\}$ – Logistics: The accumulated quantity Q_t is shipped when $Q_t \geq Q^{\mathrm{ship}}$ or a time-since-last-dispatch threshold τ_{disp} is reached. The default shipping mode is normal. – Expedite: If a predicted lateness indicator exceeds a threshold (e.g., $\widehat{T}_t > \tau^{\mathrm{late}}$) or an OTIF shortfall trigger (as defined in section 3.2.1), the mode switches to expedite. This reduces the mean and variance of transit time by a factor $\rho_{exp} > 1$, and incurs a surcharge Δc_{tr}. Budget and energy caps, as well as feasibility guards, apply. – Recycling: A fixed recovery intensity r_{fix} is applied up to recycling capacity. Recovered material is fed back into upstream stages.
Single-Agent RL	A monolithic Linear-DQN controls the full Cartesian $\mathscr{A} = \prod_i \mathscr{A}_i$, using the same reward and hyperparameters as MARL. This setup quantifies the benefits of specialization and decentralized execution.

When the baseline exhibits nonzero variance, effect sizes are reported using Glass's Δ, based on the baseline standard deviation; when the baseline variance is ∼ 0, the statistic is left undefined (—).

Robustness is assessed by rerunning the benchmark protocol under the predefined shock scenarios. Cross-sector generalization is assessed by applying the same protocol to the remaining sector archetypes.

4.6. Value of Data protocol

VoD is quantified by contrasting the Full-Data and Silo-Data observation regimes defined in Section 3.2.1, using matched seeds, a fixed horizon, and identical reward weights, normalizers, and scenarios. For the VoD, performance differences are reported as Full minus Silo. For lower-is-better KPIs, negative values indicate improved performance under Full; for service metrics (e.g., OTIF, service level), positive percentage-point differences indicate improved performance under Full. All VoD estimates are presented as means with 95% confidence intervals across seeds.

4.7. Validity, scalability, and sensitivity

The single-commodity, two-echelon setup was used to keep experiments statistically tight while preserving the core trade-offs of interest. Scalability to multi-product or multi-site settings relies on concise observation summaries and action factorization (with parameter sharing) without altering the evaluation protocol. Sensitivity checks based on ± 20% rescaling of key rates (demand, process times, transport/energy factors) are conducted to confirm that qualitative conclusions (signs and ordering across methods) remain stable.

5. Results

The performance of all evaluated policies is reported in terms of relative deviations from the No-Op baseline. For KPIs where lower values are preferable, results are expressed as percentage changes. For service-level metrics, outcomes are reported as differences in percentage points. The No-Op policy serves as a consistent anchor across all scenarios, ensuring that reported scales are comparable and performance differences are transitive. Statistical significance is assessed using 95% confidence intervals, computed on the absolute mean values of the KPIs.

5.1. Benchmark in the construction setting

This section establishes the reference behavior of the system under nominal conditions, in the absence of any exogenous shocks. The primary objective is to evaluate whether the proposed DT–MARL enhances service, specifically timeliness and reliability, while maintaining bounded economic and environmental externalities.

Under these baseline conditions, the tuned DT–MARL policy exhibited a balanced improvement across multiple objectives. Relative to the No-Op baseline, it achieved a 4.5% reduction in average lead time and a 1.8 percentage-point increase in OTIF deliveries. These service gains were accompanied by a moderate 8.8% increase in operational cost and a 4.0% rise in CO_2 emissions, while material waste was concurrently reduced by 1.6%.

Table 6 summarizes the deltas with respect to No-Op across all methods. For clarity, negative deltas indicate improvements in lower-is-better metrics (lead time, cost, CO_2, material waste, unit cost, and energy per unit), whereas positive deltas reflect improvements in higher-is-better metrics (service level and OTIF).

Fig. 1 depicts the benchmark KPIs for the construction setting, based on sample means across runs and seeds. Horizontal bars denote 95% confidence intervals when variance estimates are available; for single-run estimates, bars are omitted.

DT–MARL reduces lead time and improves OTIF performance relative to the No-Op baseline, while keeping increases in cost and CO_2 emissions bounded. In contrast, the Heuristic achieves high service levels at the expense of substantially higher cost and environmental impact.

To more precisely position DT–MARL on the trade-off surface, Table 7 reports pairwise contrasts against the best non-MARL baseline. For each indicator, the "best baseline" refers to the non-MARL method with the most favorable mean (minimum for lower-is-better; maximum for higher-is-better).

Oriented contrasts are signed so that positive values favor DT–MARL: for higher-is-better metrics, the contrast is calculated as $\mu_{\text{DT-MARL}} - \mu_{\text{best}}$, for lower-is-better metrics, it is $\mu_{\text{best}} - \mu_{\text{DT-MARL}}$. Glass's Δ is computed using the standard deviation of the best baseline; when that baseline is (near-)deterministic, Δ is not reported.

The contrasts reported in Table 7 are computed relative to the per-metric best among non-MARL baselines; consequently, negative values are expected for objectives dominated by corner solutions. The shortest lead time and highest OTIF are attained by the Heuristic baseline; accordingly, DT–MARL falls behind by 5.46 days in lead time and by 12.1 percentage points in OTIF. However, the Heuristic baseline also incurs the highest economic and environmental externalities (Table 6), which is consistent with a corner solution on the Pareto frontier.

For cost- and sustainability-oriented indicators, the best baseline is typically No-Op, which exhibits essentially zero variance; therefore, Glass's Δ is undefined, and the apparent per-metric advantages reflect inaction rather than superior performance. The only metrics with evaluable effect sizes are service level (a slight disadvantage for DT–MARL, $\Delta \approx -0.30$) and energy per unit (a moderate disadvantage relative to Random, $\Delta \approx -0.87$).

Taken together with Table 6, these results indicate that DT–MARL does not maximize any single objective. Instead, it occupies a balanced region of the trade-off surface, delivering concurrent improvements in timeliness and waste reduction while keeping cost and CO_2 emissions within narrow, predictable bounds, consistent with the fixed multi-objective reward structure, feasibility safeguards and deployment requirements.

A comparison with single-agent reinforcement learning was used to clarify the positioning of DT–MARL on the trade-off surface. Relative to the No-Op baseline, a larger lead-time reduction was obtained by single-agent reinforcement learning (−9.6% vs − 4.5%), together with a higher OTIF gain (+3.2p.p. vs + 1.8p.p.), and a larger reduction in material waste (−2.4% vs − 1.6%). However, larger increases were also incurred in operational cost (+10.0% vs + 8.8%) and CO_2 emissions (+4.5% vs + 4.0%). In absolute terms, a 1.2percentage-point smaller cost increase and a 0.5 percentage-point smaller CO_2 increase were incurred by DT–MARL, corresponding to approximately 12% and 11% reductions in the magnitude of the externality increases relative to single-agent

Table 6
Benchmark deltas vs No-Op by algorithm.

Algorithm	Lead time (%)	OTIF (p.p.)	Service (p.p.)	Cost (%)	CO_2 (%)	Material waste (%)	Cost per unit (%)	Energy per unit (%)
No-Op	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Random	−8.5	2.5	0.2	10.1	4.9	−1.6	7.6	−2.6
Heuristic	−79.0	13.8	0.0	126.4	48.3	−9.6	126.4	0.0
Single-agent RL	−9.6	3.2	0.4	10.0	4.5	−2.4	8.5	−1.5
DT–MARL (tuned)	−4.5	1.8	0.2	8.8	4.0	−1.6	8.8	−0.0

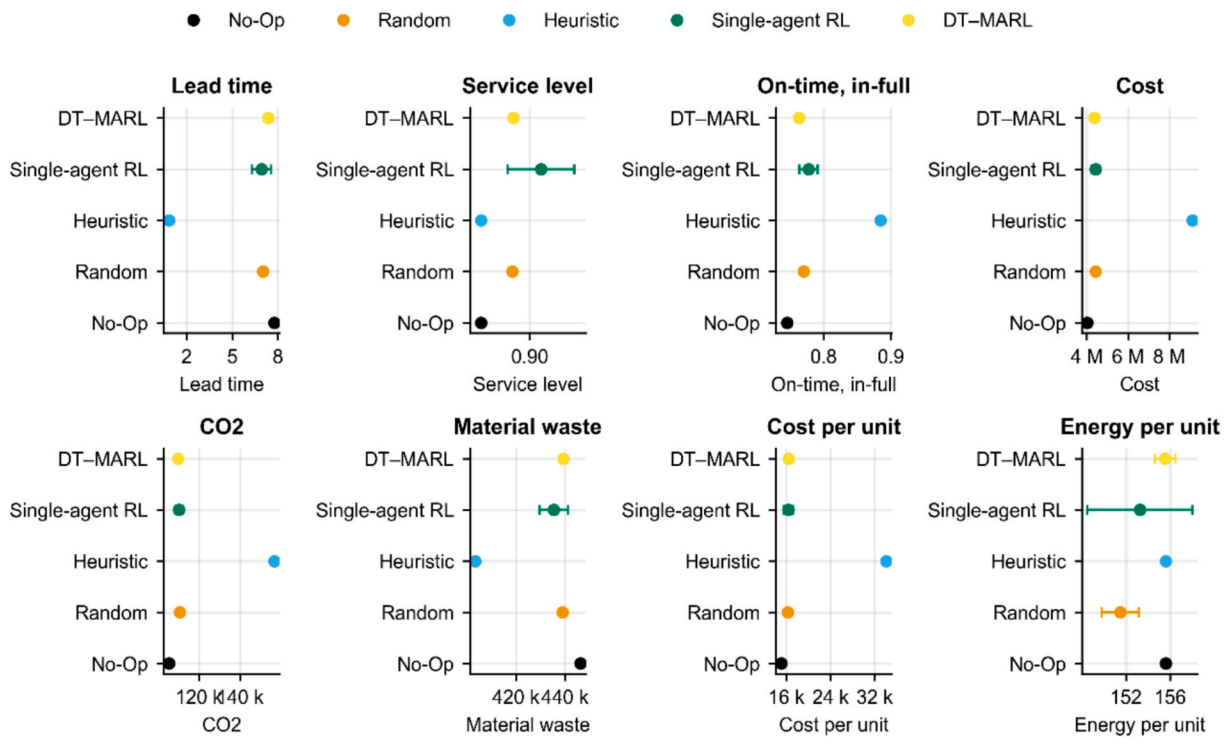

Fig. 1. Benchmark KPIs in the construction setting.

Table 7
Pairwise contrasts at the benchmark: DT–MARL vs the best non-MARL baseline *("+" favors DT–MARL; Glass's Δ uses the baseline standard deviation.)*.

Indicator	Best non-MARL baseline	Oriented contrast (+favors DT–MARL)	Baseline SD	Glass's Δ
Lead time (days, ↓)	Heuristic	−5.455	0.000	—
OTIF (0–1, ↑)	Heuristic	−12.1p.p.	0.000	—
Service level (0–1, ↑)	Single-agent RL	−0.2p.p.	0.006	−0.300
Operational cost (currency, ↓)	No-Op	−353,274.771	0.000	—
CO_2 emissions (kg, ↓)	No-Op	−4,273.083	0.000	—
Material waste (kg, ↓)	Heuristic	−35,973.301	0.000	—
Cost per unit (currency, ↓)	No-Op	−1,327.571	0.000	—
Energy per unit, EPU (kWh, ↓)	Random	−4.064	4.671	−0.870

reinforcement learning.

Despite identical objective definitions and evaluation conditions, different operating points were obtained. A more service-prioritizing profile was produced by single-agent reinforcement learning, while a more externality-bounded profile was created by DT–MARL. It was shown that the preferred operating point is context-dependent, and it should be selected according to the relative priority placed on service versus externalities.

Beyond benchmark positioning, three architecture-level motivations supported the multi-agent formulation. First, scalability under action-space growth was emphasized: while centralized control remains tractable in the present setting, combinatorial growth is expected in multi-commodity, multi-site, or deeper multi-echelon extensions, and action factorization with parameter sharing was identified as the intended pathway for scaling. Second, modularity was supported because the five-agent decomposition aligned with standard Industrial Engineering decision functions, enabling targeted diagnostics and mechanism-aware tuning. Third, interpretability via attribution was supported through agent-level ablations, where distinct and interpretable KPI shifts were observed under controlled removals (see Table 11); such attribution was more difficult to obtain from a monolithic policy selecting all actions jointly.

The operational significance of the benchmark deltas was interpreted in conjunction with uncertainty under the controlled DT protocol. Performance was aggregated at the episode level over a fixed 20-day horizon and then averaged across matched seeds and episodes, with 95% confidence intervals reported from seed-level averages (Fig. 1 and Table 6). Consequently, the variability induced by stochastic demand, lead times, and the DT physics was made explicit, and the stability of the observed mean improvements could be assessed without relying on single-run outcomes. At the same time, it was noted that whether a + 1.8 percentage-point OTIF gain and a − 4.5% lead-time reduction justify an + 8.8% cost increase is context dependent and should be evaluated against sector-specific service valuation and penalty structures.

5.2. Policy robustness to operational shocks

To assess the resilience of DT–MARL, its learned policy was tested against four scenarios of exogenous environmental change. These disruptions, detailed in Table 8, include a 13% increase in transport time (Transport), a 50% surge in demand (Demand), a 13% rise in energy prices (Energy), and the simultaneous application of all three stressors ("Combined").

The central finding from this analysis is the policy's high degree of robustness. Across all tested scenarios, DT–MARL's qualitative behavior remained remarkably stable. The directionality (sign) of all performance deltas relative to the No-Op baseline was consistently preserved: lead time and material waste were reliably reduced, while on-time-in-full deliveries improved.

Furthermore, the magnitudes of these outcomes remained within a narrow and predictable range. For instance, lead time reduction was sustained between 4.2% and 5.2%, and the associated cost increase was

Table 8
Shock experiments: tuned DT–MARL (Δ vs No-Op).

Scenario	Lead time (%)	OTIF (p.p.)	Service (p.p.)	Cost (%)	CO_2 (%)	Material waste (%)	Cost per unit (%)	Energy per unit (%)
Transport	−5.2	1.8	0.2	9.4	4.3	−1.5	9.4	−0.1
Demand	−4.4	1.8	0.2	9.3	4.3	−1.6	9.3	0.3
Energy	−4.9	1.8	0.2	8.9	4.1	−1.6	8.4	−0.6
Combined	−4.2	1.7	0.2	9.3	4.3	−1.5	9.3	−0.1

contained between 8.9% and 9.4%. This stability demonstrates that the policy effectively mitigates the impact of shocks without significant degradation in performance.

A notable instance of adaptive optimization was observed under the 13% energy price increase. In response to this specific cost signal, DT–MARL not only preserved service levels but also actively reduced energy consumption per unit by 0.6%. This economically rational adjustment provides strong evidence that the policy has learned to dynamically manage resource intensity rather than follow a static rule set.

In summary, the policy exhibits a coherent and adaptive strategy under stress. As corroborated by the visual data in Figs. 2–5, DT–MARL maintains shorter lead times and higher delivery reliability while keeping economic and environmental externalities contained, confirming its robustness under adverse operating conditions.

5.3. Cross-sector generalization

To evaluate the generalizability of the learned policy, DT–MARL, which was tuned exclusively on the construction benchmark, was deployed without any retraining across three distinct industrial archetypes: discrete manufacturing, process industries, and waste management.

This experiment was designed to assess whether transferable coordination principles could be captured beyond the construction benchmark, rather than to demonstrate performance in a single application domain.

The results, presented in Table 9, demonstrate that the policy's fundamental benefits are indeed generalizable. Two key performance improvements were consistently observed across all tested sectors: a significant reduction in average lead time (ranging from −3.8% to −5.2%) and a notable decrease in material waste (−1.4% to −1.8%). These findings suggest that the policy has learned a robust and transferable logic for enhancing flow efficiency and circularity.

However, the impact on OTIF delivery has been shown to depend on sector-specific dynamics. While a substantial 5.0 percentage-point gain in OTIF was obtained in the process industries sector, a marginal degradation of − 0.3 percentage points was observed in discrete manufacturing despite the concurrent 3.8% lead-time reduction. This decoupling between average cycle time and delivery-window compliance has been documented in production planning settings where due-date performance is mediated by variability, release mechanisms, and due-date tightness, such that improvements in mean flow time can coexist with unchanged or worsened delivery-window compliance under heterogeneous operating conditions (Stevenson et al., 2005; Thürer et al., 2011). In addition, differences in order release and sequencing constraints have been shown to influence tardiness outcomes even when average flow times are reduced (Ragatz & Mabert, 1988).

A boundary condition for zero-shot transfer has therefore been indicated. While lead-time reduction and material-waste improvement have generalized across all tested sectors, OTIF performance has been shown to be more sensitive to sector-specific timing and compliance logic. For deployment in discrete manufacturing contexts, modest

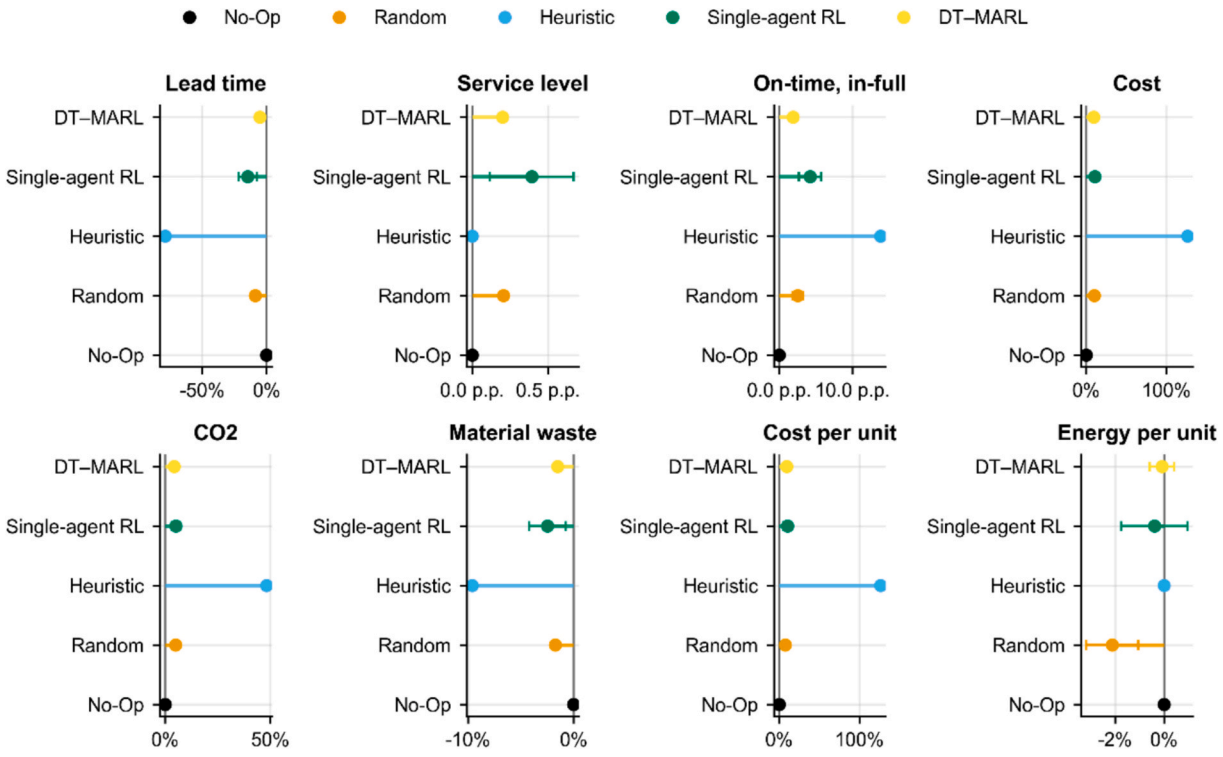

Fig. 2. DT–MARL performance under transportation shocks.

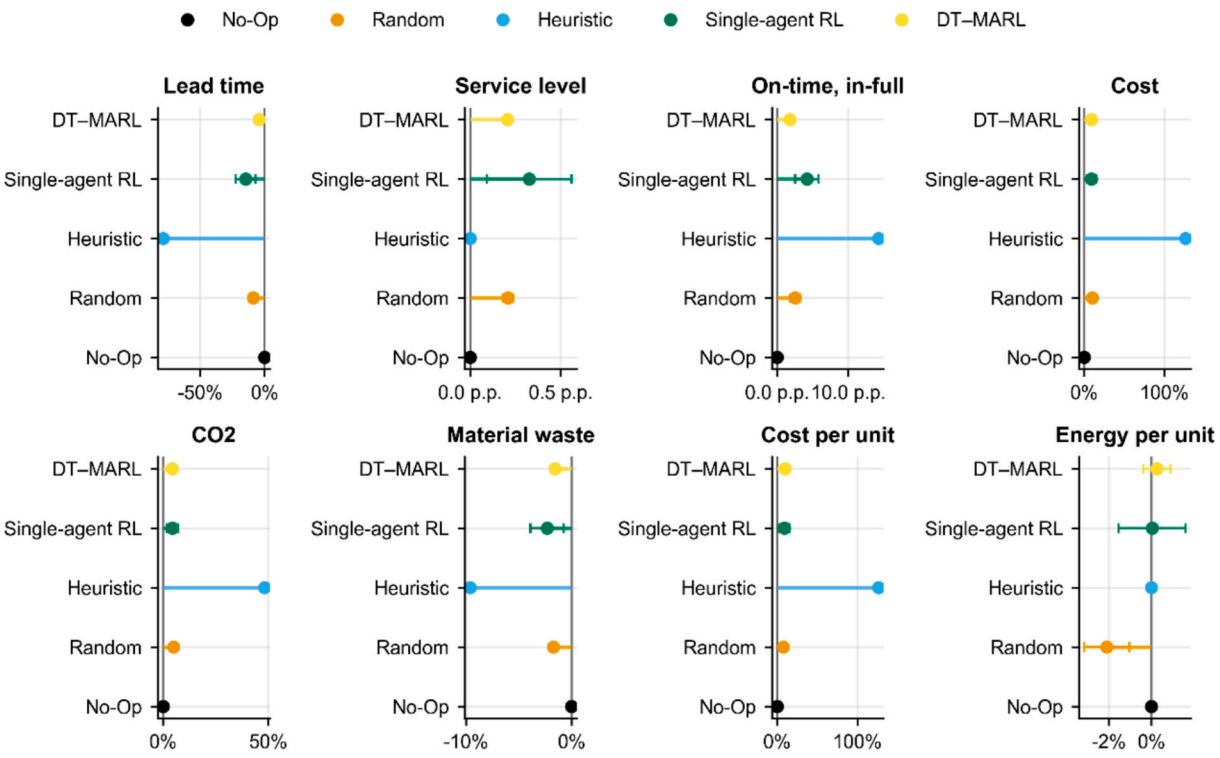

Fig. 3. DT–MARL performance under demand shocks.

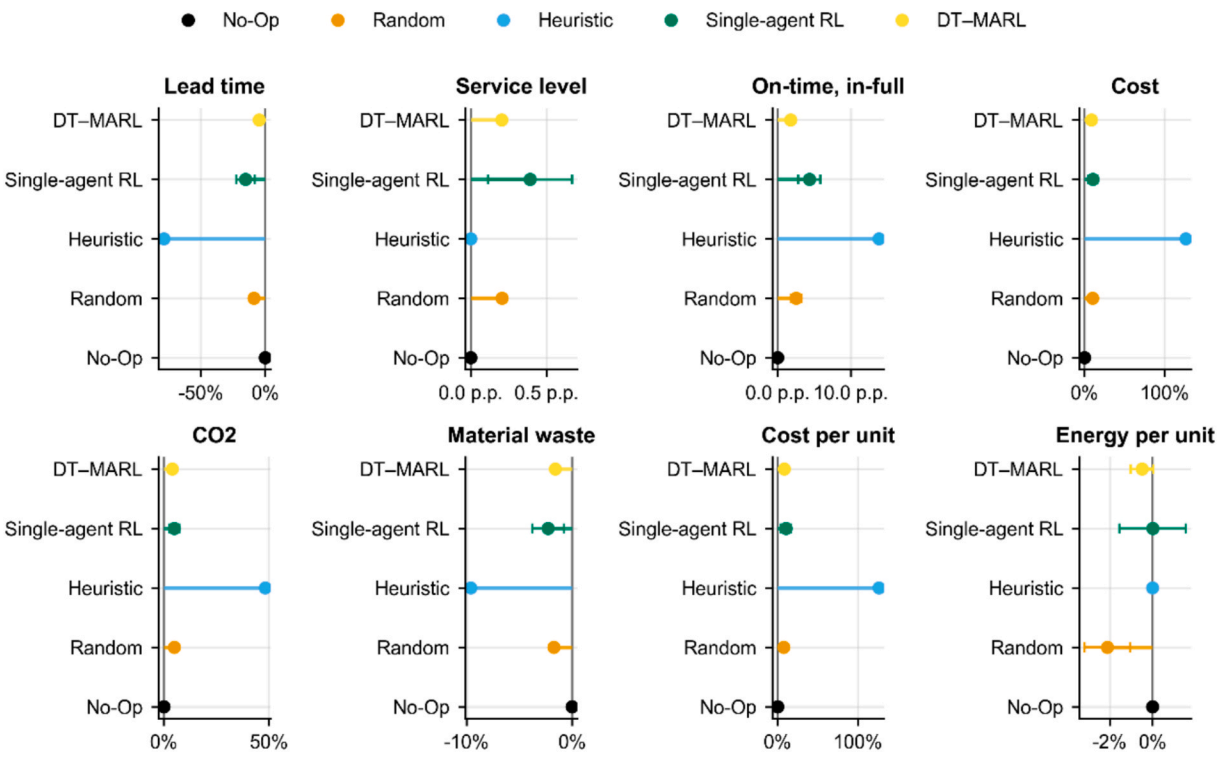

Fig. 4. DT–MARL performance under energy shocks.

additional training or reward-weight adjustment targeting OTIF variance could be considered to recover schedule-adherence performance without sacrificing the observed flow-time and waste benefits.

Despite the variable impact on OTIF, the overarching conclusion is that the policy successfully generalizes its core performance profile. As illustrated in Fig. 6, it consistently improves operational speed and reduces waste, while keeping cost and emission increases within bounded limits across diverse settings. The policy is therefore not narrowly confined to its training domain, but can serve as an effective baseline controller for a broad range of circular supply chain applications.

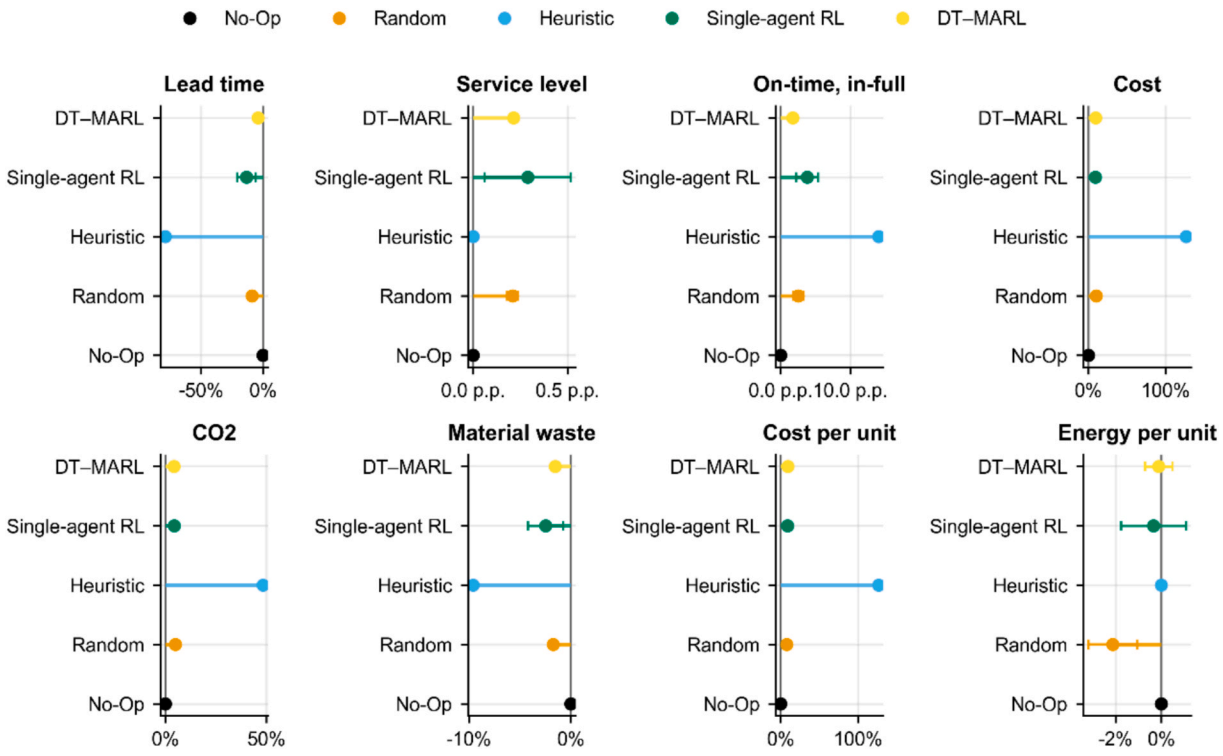

Fig. 5. DT–MARL performance under combined shocks.

Table 9
Cross-sector experiments: tuned policy Δ vs No-Op.

Sector	Lead time (%)	OTIF (p.p.)	Service (p.p.)	Cost (%)	CO_2 (%)	Material waste (%)	Cost per unit (%)	Energy per unit (%)
Discrete manufacturing	−3.8	−0.3	0.2	12.6	5.4	−1.8	14.7	2.0
Process industries	−5.2	5.0	0.2	11.2	5.0	−1.5	11.2	−0.1
Waste management	−4.9	0.2	0.2	10.7	4.6	−1.4	11.3	0.5

5.4. VoD (Full versus Silo)

To quantify the marginal impact of integrated information, this analysis compares two distinct observability configurations. *Full Data* refers to the integrated-information setting, where DT–MARL operates with the complete set of signals defined in Section 3.2.1, (including KPI snapshots, stage loads and work-in-process, on-hand inventory by echelon, backlog flags, shipment queues, lane-level estimated-time-of-arrival statistics, recycling-capacity utilization, and exogenous indicators). Conversely, *Silo Data* corresponds to the masked observation model in Section 4.6, which applies a binary mask to the same vector while keeping the digital-twin physics, reward weights, normalizers, scenarios, seeds, and horizons identical. For the results presented below, performance differences are computed as "Full minus Silo", using percentage changes for lower-is-better indicators and percentage points for service measures.

The results, shown in Table 10, reveal that the primary value of integrated data lies in enabling the controller to identify and exploit a more efficient operational trade-off. Access to comprehensive data allows the policy to prioritize economic and environmental performance, accepting a modest 3.4% increase in lead time in exchange for significant, concurrent reductions in operational cost (−3.7%), CO_2 emissions (−1.9%), and material waste (−1.2%).

A key insight from the VoD analysis was that integrated information did not uniformly improve all KPIs. Instead, it enabled a rebalancing of the operating point along the trade-off surface under fixed reward weights. Under Full-Data observability, cost and emission reduction opportunities were identified through cross-functional visibility that had not been available under siloed observations. The observed 3.4% increase in lead time was interpreted as a deliberate policy shift rather than as a degradation, as expediting intensity has been reduced and shipment consolidation was increased when feasibility was supported by integrated observability.

This behavior was consistent with the tuned reward weights, under which cost and CO_2 penalties outweighed marginal lead-time increases within the service-feasible region. Crucially, the VoD protocol held DT physics, reward weights, normalizers, seeds, and horizons constant across Full and Silo conditions, so the lead-time increase was attributable to policy choice under richer information rather than to environmental changes. Under Full-Data, substantial reductions were obtained in cost (−3.7%), CO_2 emissions (−1.9%), and material waste (−1.2%), while OTIF was marginally improved (+0.7p.p.).

Managerially, it was indicated that data integration should not be expected to accelerate all metrics simultaneously. Value was instead created by enabling navigation of the multi-objective landscape, where small degradations in lower-priority KPIs were accepted to reach a more desirable operating point under the stated objective structure (Galbraith, 1974; Lee et al., 1997).

5.5. Ablation studies for causal attribution

To attribute the observed performance to specific model components, two sets of ablation studies were conducted. First, each of the five action families (e.g., expediting, recycling) was individually deactivated. Second, key terms from the reward function (e.g., cost, lead time) were individually zeroed out. The resulting performance **Δ**, reported

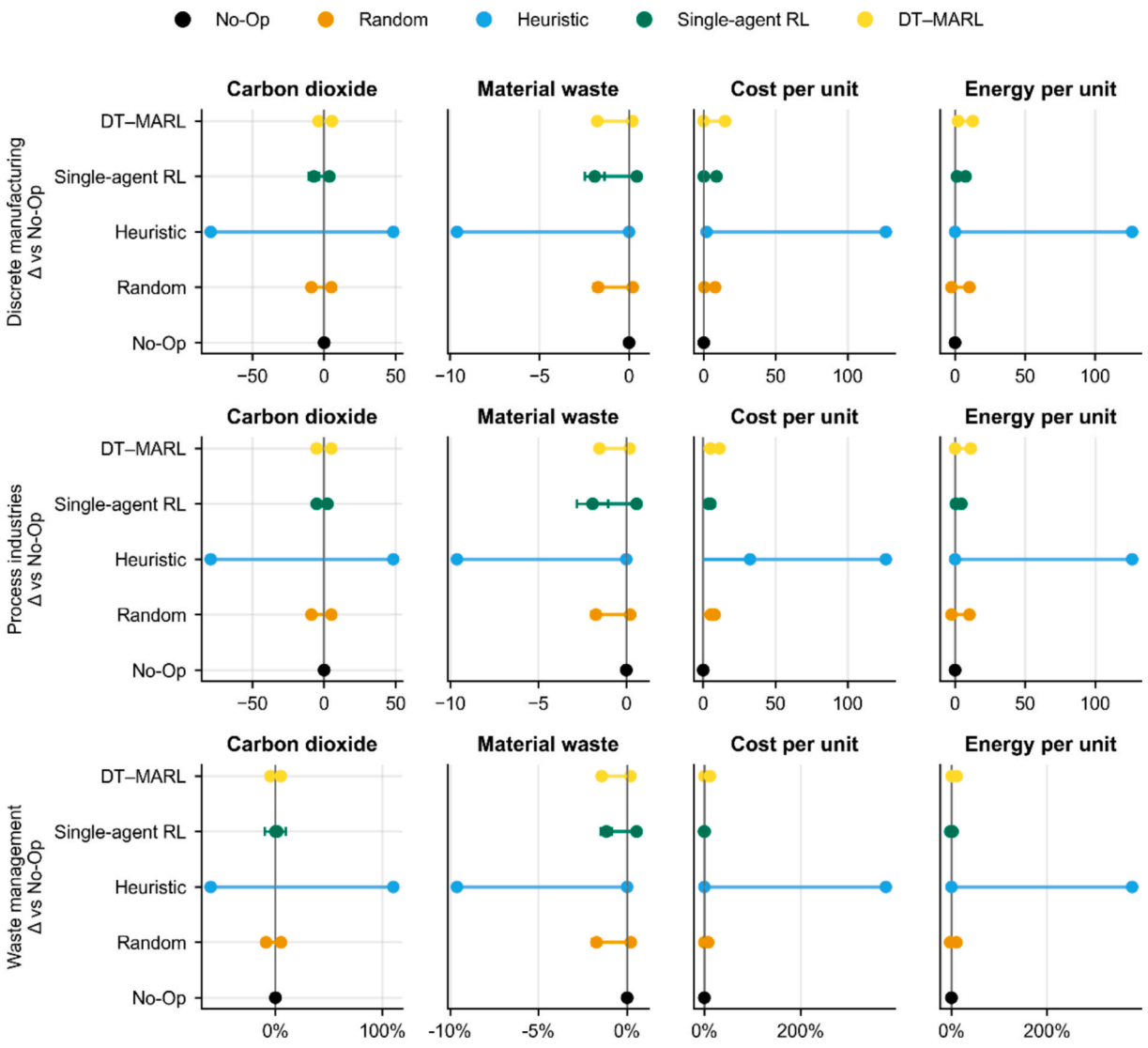

Fig. 6. Analysis of the DT–MARL model's generalization across sectors.

Table 10
Full Data minus Silo Data under the tuned policy.

Indicator	Full − Silo
Lead time (%)	+3.354
Cost (%)	−3.745
CO_2 emissions (%)	−1.889
Material waste (%)	−1.184
Service (p.p.)	+0.007
OTIF (p.p.)	+0.715

relative to the full, tuned policy, are presented in Table 11.

The analysis reveals the distinct causal role of each component. For instance, disabling the expediting function was identified as the primary driver of service velocity, resulting in a 7.3% increase in lead time and a 1.9 percentage-point drop in OTIF (see Fig. 8). This degradation was accompanied by a corresponding decrease in cost (−6.6%) and emissions (−3.1%), thereby quantifying the core speed-externality trade-off.

The impact of other actions was more specialized: deactivating recycling directly increased material waste by 1.6%, while disabling the scheduling/planning lever led to a marginal reduction in lead time (−1.9%), suggesting a degree of redundancy with other levers. Similarly, the reward term ablations confirmed that the policy's behavior is logically aligned with its incentives. Removing the penalty for cost or CO_2 prompted the agent to prioritize speed, leading to faster delivery times but higher costs and emissions. Conversely, removing the lead-time term eliminated the direct incentive for speed, resulting in a slight increase in delivery times.

Collectively, these studies successfully deconstruct the learned behavior of DT–MARL, isolating the specific mechanisms responsible for balancing the multi-objective requirements of the CSCs.

Table 11
Ablations: changes vs the full-tuned policy.

Configuration	Lead time (%)	Cost (%)	CO_2 (%)	Material waste (%)	Service (p.p.)	OTIF (p.p.)
No expediting	+7.31	−6.63	−3.09	+0.01	+0.00	−1.86
No scheduling/ planning	−1.86	−0.12	−0.07	+0.04	+0.01	+0.27
No logistics	−0.76	−1.25	−0.79	−0.02	−0.19	+0.02
No inventory	−0.56	+0.51	+0.21	−0.09	+0.00	−0.04
No recycling	+0.21	−0.15	−0.08	+1.59	+0.00	−0.06
No cost term	−0.52	+0.50	+0.22	−0.01	+0.00	+0.15
No CO2 term	−0.43	+0.41	+0.16	+0.01	+0.01	+0.14
No lead-time term	+0.16	+0.05	+0.00	−0.09	+0.01	+0.00

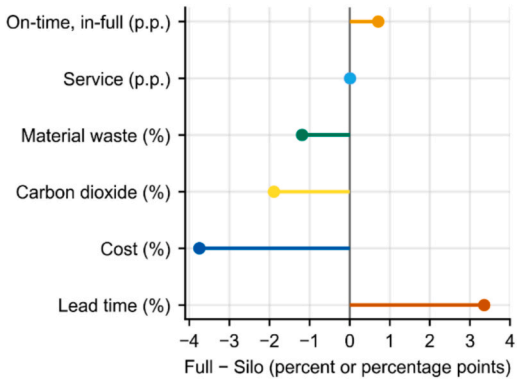

Fig. 7. Quantifying the VoD: Full vs. Silo configurations.

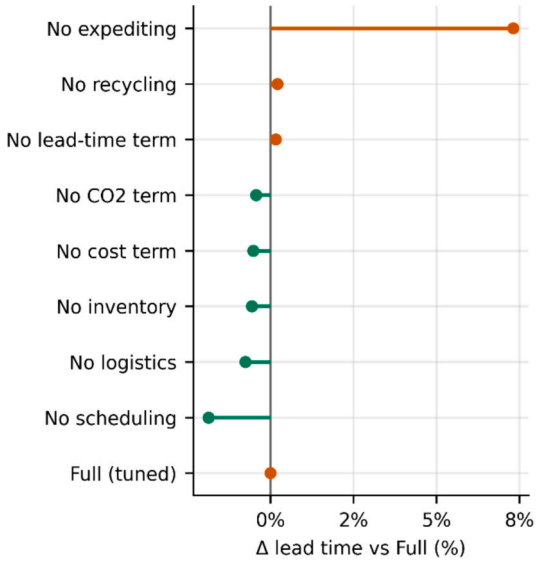

Fig. 8. Ablations (lead-time change vs Full).

5.6. Stability

Within-episode stability was evaluated using stepwise series of average lead time. DT–MARL was assessed against two baselines: a No-Op policy and a Random policy. For each method, the stepwise mean across seeds was computed, and dispersion was summarized by the IQR, the coefficient of variation (CV), and the share of steps falling within the band $\mu \pm 2\sigma$ (where μ denotes the mean and σ the standard deviation). Fig. 9 displays the trajectories and the $\mu \pm 2\sigma$ bands (dispersion across seeds) for the compared controllers.

During the final 10% of the horizon (steps 18–20), DT–MARL operated within a narrow, stationary band: mean lead time was 6.568 days, with IQR = 0.027, CV = 0.42%, and an ordinary least squares linear trend of − 0.027 days per step. Across the full episode, 95.0% of DT–MARL steps fell within $\mu \pm 2\sigma$, indicating tight dispersion relative to its own mean. As a reference, No-Op remained flat at 7.319 days (zero dispersion by design), while Random exhibited a lower trajectory coupled with a wider $\mu \pm 2\sigma$ band and a non-negligible drift.

Taken together, these diagnostics support within-episode stability of DT–MARL: after the initial transients, lead time fluctuated inside a bounded envelope and showed negligible drift in the closing segment of the episode.

5.7. Discussion and managerial implications

Taken together, the evidence demonstrates that the DT–MARL framework improves timeliness and delivery reliability while keeping economic and environmental externalities within a narrow and predictable range. In the benchmark scenario without shocks, DT–MARL shortens average lead time by 4.5% and increases OTIF by 1.8 percentage points, with material waste reduced by 1.6% and energy per unit essentially unchanged. Cost and CO_2 increases remain bounded at 8.8% and 4.0%, respectively (see Fig. 1 and Table 6).

Managerial significance is treated as context dependent. Although improvements are reported with confidence intervals under matched seeds and fixed horizons, the translation of a +1.8 percentage-point OTIF gain and a −4.5% lead-time reduction into economic value depends on service-level penalties, revenue sensitivity to reliability, and the extent of the simulation-to-reality gap. In addition, the reported +8.8% cost increase is measured relative to the No-Op baseline, which represents inaction rather than an optimized operating point. Therefore, the results are positioned as demonstrating reproducible trade-off shaping under uncertainty, while site-specific calibration and validation are recommended before deployment.

Under transport, demand, energy, and combined shocks, the sign of

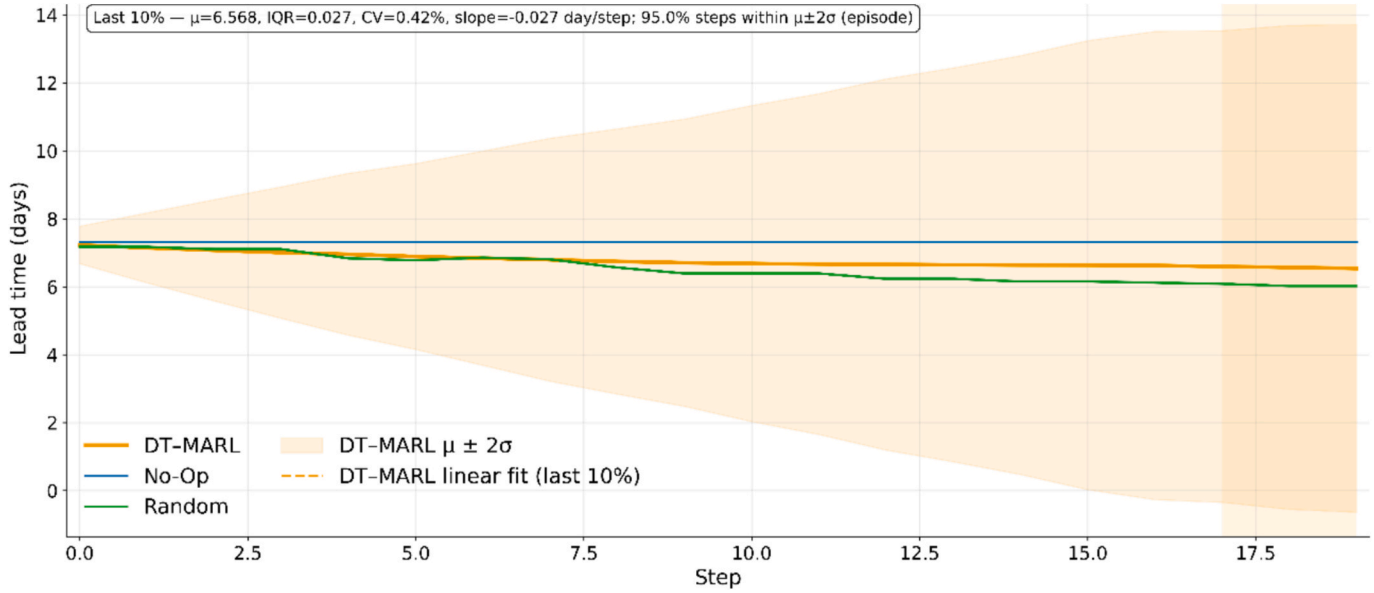

Fig. 9. Lead-time stability: trajectory and $\mu \pm 2\sigma$ band (DT–MARL vs baselines).

these effects is preserved and the magnitudes remain close. As a result, lead time improves by 4.2–5.2%, OTIF increases by 1.7–1.8 percentage points, material waste decreases by approximately 1.5–1.6%, and cost and CO_2 fall within a tight band of 8.9–9.4% and 4.1–4.3% band (Table 8; Figs. 2–5).

DT–MARL also transfers across sectors without retuning. Discrete manufacturing, process industries, and waste management all exhibit shorter lead times (−3.8% to −5.2%) and lower material waste (−1.4% to −1.8%); process industries show the largest gain in OTIF (+5.0p.p.), while discrete manufacturing trades a slight decrease in OTIF (−0.3p.p.) for faster flow (see Table 9 and Fig. 6).

Compared to realistic baselines, the benchmark contrasts clarify DT–MARL's position on the service–cost–emissions surface. The Heuristic attains the largest nominal gains in timeliness and OTIF but does so at the expense of very large increases in cost and CO_2 (e.g., +126% and +48% in the benchmark), whereas DT–MARL delivers smaller, targeted gains in timeliness and reliability, along with reduced material waste and bounded increases in cost and emissions. Larger improvements in timeliness and OTIF are obtained by Random and single-agent RL, and material waste is also reduced, yet slightly higher cost and CO_2 increases are incurred, whereas smaller, targeted gains are achieved by DT–MARL with comparable waste reduction and more conservative externality growth. Pairwise contrasts and effect sizes computed from the exported variances confirm that the apparent dominance of Heuristic in delivery reliability is inseparable from its extreme economic and environmental penalties. DT–MARL occupies a balanced region where the three-criteria is better aligned with the tuned reward (Tables 5 and 6).

Relative to single-agent RL, smaller service gains are obtained by DT–MARL in the present benchmark, while smaller increases in cost and CO_2 are incurred, reflecting the externality-conscious tuning of the reward weights and the intended positioning on the trade-off surface.

These findings address the study's research questions and support the hypotheses outlined in the experimental plan.

With respect to effectiveness (H1), DT–MARL improves lead time and OTIF relative to the No-Op baseline, while reducing material waste and keeping cost, CO_2 emissions, and energy per unit tightly bounded. Relative to the Heuristic baseline, a more balanced outcome is delivered, and the extreme externality increases induced by that baseline are avoided, while waste reductions that are comparable to the other learning baselines are maintained.

Regarding robustness (H3), the sign of DT–MARL's improvements is preserved under transport, demand, and energy shocks, as well as their combination, with only modest changes in margins. Under energy shock, the energy-per-unit indicator decreases slightly (−0.6%), which is consistent with reduced energy intensity in response to rising prices.

For generalization (H1/H3), DT–MARL transfers across sectors without retuning: all three sectors display shorter lead times and lower waste, with sector-specific differences in delivery reliability that are consistent with known routing and batching effects.

As for the VoD (H2), using the integrated observation model (Full-Data) instead of a siloed model (Silo-Data) is associated with lower cost (−3.7%), lower CO_2 emissions (−1.9%), and reduced material waste (−1.2%), a slight improvement in OTIF (+0.7p.p.), no change in service level, and a modest increase in lead time (+3.4%) (see Table 10 and Fig. 7). This pattern reflects a shift toward economic and environmental efficiency when richer signals are available, with a small, explicit cost in cycle time.

Notably, the lead-time increase observed under Full-Data constitutes a non-trivial finding: integrated information does not universally accelerate operations, but instead enables trade-offs that are inaccessible under siloed observability. Data integration initiatives in circular supply chains should therefore be evaluated against multi-dimensional value metrics rather than against single-KPI expectations, consistent with the view that information value in complex systems is mediated by decision architecture and objective structure (Galbraith, 1974).

Finally, concerning mechanism (H4), the ablation studies provide a direct causal mapping between policy components and outcomes: removing expediting increases lead time by 7.3% and reduces OTIF by 1.9 percentage points, while lowering cost and CO_2 emissions (−6.6% and − 3.1%), identifying expediting as the dominant speed lever with transparent trade-offs. Removing logistics lowers cost and CO_2 (−1.2% and − 0.8%) but also reduces service (−0.2p.p.), consistent with logistics acting as a service-protection buffer. Removing planning and scheduling slightly improves lead time (−1.9%) with negligible spillovers, suggesting redundancy with other actions in this environment. Removing recycling increases material waste (+1.6%) with minimal cross-effects. Reward-term ablations shift behavior as expected: zeroing the cost or CO_2 term pushes DT–MARL toward speed, with small increases in the corresponding externalities (see Table 11 and Fig. 8).

From a deployment standpoint, DT–MARL is particularly suitable for organizations seeking shorter lead times and improved delivery reliability, while accepting modest and predictable increases in cost and CO_2 and valuing reductions in material waste. If an extreme pursuit of OTIF is mandated regardless of externalities, a heuristic similar to the one evaluated here will achieve greater reliability gains but at disproportionately high economic and environmental cost; in this regard, the benchmark and shock results quantify these penalties. In settings with volatile energy prices, the energy-shock response indicates that DT–MARL adapts by slightly lowering energy intensity, which is desirable in cost-exposed environments. Where data integration across silos is feasible, the VoD results recommend using the integrated observation model to reduce cost and CO_2 and improve OTIF, recognizing the small increase in cycle time as a deliberate and transparent shift along the trade-off surface. The ablation evidence provides practical guidance for tuning: (i) maintain expediting as the primary lever for cycle-time control, but regulate its intensity to cap cost and CO_2 impacts; (ii) preserve logistics actions when service protection is paramount; (iii) keep recycling decisions active to ensure waste reductions; and (iv) treat explicit scheduling as a secondary lever in this environment, given its limited incremental value.

To ground these recommendations in a concrete decision context, consider a process-industry plant operating a circular line in which post-consumer material is recovered and reprocessed, facing contractual penalties tied to OTIF compliance and a corporate mandate to reduce landfill-bound waste. A manager evaluating DT–MARL would begin with the cross-sector evidence (Table 9): under zero-shot transfer of the construction-tuned policy, the process-industry archetype exhibits the largest OTIF gain among the tested sectors (+5.0p.p.), together with a 5.2% lead-time reduction and a 1.5% decrease in material waste, at cost and CO_2 increases of approximately 11% and 5%, respectively. To assess robustness, the manager could then consult the shock results obtained on the benchmark sector (Table 8), where the qualitative direction of improvement is preserved under transport, demand, and energy perturbations, and effect magnitudes remain within a narrow range. Finally, the ablation map (Table 11) would inform lever configuration: expediting as the primary speed lever, recycling to sustain waste diversion, and the cost reward term to bound externality growth. This sequence illustrates how the framework's structured evidence base can be assembled into a deployment brief that makes trade-offs explicit and actionable, while site-specific calibration remains necessary before operational commitment, as discussed in Section 6.

6. Conclusions and future work

This study introduces a DT–MARL framework for circular supply chains and evaluates its performance under a fixed multi-objective reward across a benchmark setting, exogenous shocks, and three industrial sectors. A single neutral anchor (the no-action policy) was used throughout, allowing percentage and percentage-point effects to remain comparable across indicators and scenarios. The results demonstrate that DT–MARL consistently improves timeliness and delivery reliability while keeping cost and CO_2 impacts within a narrow, predictable range

and reducing material waste. These patterns hold under transport, demand, and energy shocks (both individually and in combination) and transfer without retuning to discrete manufacturing, process industries, and waste management.

Experiments on information design show that integrated data shift the operating point toward lower cost and lower CO_2 without compromising compliance, and ablation studies provide a clear causal map from levers to outcomes: expediting emerges as the primary driver of cycle-time reduction, logistics acts as a service buffer, and recycling serves as the direct control on waste.

Beyond the quantitative evidence, the study contributes an evaluation protocol that preserves transitivity of effects across settings, reports absolute levels with 95% confidence intervals to make trade-offs observable, and separates algorithmic performance from informational conditions by contrasting integrated versus siloed observations. Together, these features enable reproducible comparisons and transparent interpretation of speed–service–emissions trade-offs in circular supply chains.

A key limitation of the study is its external validity and data realism: all evaluations were conducted in a DT environment rather than on actual production traces. Although the DT reproduces process logic and preserves common normalizations, seeds, and horizons across scenarios, any structural misspecification (for instance, unmodeled operational constraints, policy interventions, or exogenous shock distributions) can introduce a simulation-to-reality gap. Therefore, the reported gains should be interpreted as model-consistent "what-if" improvements. Relative comparisons remain informative, since all algorithms face the same physics and information sets, but absolute magnitudes may shift in deployment.

To mitigate this risk, future work will calibrate the DT using historical event logs and re-estimate stochastic primitives, run shadow-mode or A/B pilots to measure field deltas using the same key performance indicators, stress-test with domain randomization and sensitivity analyses around demand seasonality, transport and energy shocks, and parameter uncertainty, and inject measurement noise and non-stationary regimes to align the observation process with real-world data.

A further limitation of the current evaluation is that the sector categories are implemented as archetypal regimes and the network topology is intentionally stylized as a single-commodity, two-echelon configuration. This design choice enables controlled experimentation and valid statistical comparisons, but interactions present in more complex CSC settings are not represented, including multi-site coordination, multi-commodity flows with material substitution, and shared capacities across products or sites. The primary purpose of the stylized scope is to demonstrate robust cross-regime control behavior while coordination effects are isolated from confounding structural variations. In applied deployments, site-specific DT parameters should be re-estimated from operational data, and extensions to multi-site and multi-commodity configurations should be pursued through state augmentation, agent replication with parameter sharing, and, where needed, hierarchical coordination.

Notwithstanding these limitations, practical applicability is supported by the fact that the decision levers and constraints represented in the DT can be instantiated in site-specific settings, so that the cooperative control architecture can be evaluated across subsector contexts without changes in the DT–MARL formulation, while validation remains required under the targeted structural and operational constraints.

Future research directions also include the following: First, record full training trajectories across episodes to quantify between-episode convergence rates and tail variability under alternative reward weightings and observation designs. Second, move from a single fixed set of weights to preference-consistent tuning (via structured scalarization schedules or preference learning) to trace attainable frontiers among timeliness, delivery reliability, cost, CO_2, and waste, and to estimate the price of moving along those frontiers. Third, expand the shock space (in terms of magnitude, duration, and composition) to delineate the boundaries of the sign-preservation property observed in this study and to estimate the cost of robustness. Fourth, develop sector-specific variants that retain common normalization and horizon settings but allow for modest retuning, testing whether sectoral routing and batching effects can be leveraged for additional gains without enlarging externalities. Fifth, translate ablation insights into mechanism-aware policy shaping, such as budgeted expediting or explicit carbon budgets enforced as constraints during learning, so that bounded-externality behavior is guaranteed rather than emergent. Sixth, deepen the information-design analysis by attributing performance changes to specific signals, identifying minimal sufficient observation sets that retain most benefits of integration at lower data-engineering cost, and validating these choices in shadow-mode pilots.

In summary, DT–MARL emerges as a practical candidate for deployment in CSC settings where predictable trade-offs, robustness to exogenous change, and transparency in operational levers are required. DT–MARL's consistent improvements in timeliness and delivery reliability, along with bounded cost and CO_2 impacts and reduced material waste, provide a solid foundation for pilot adoption. Meanwhile, the proposed roadmap targets the remaining steps toward scalable, preference-aligned deployment in real-world operations.

CRediT authorship contribution statement

Eduardo Guzmán: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Beatriz Andrés:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Marta Torres-Polo:** Writing – original draft, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by the project titled "*Gestión Integral de Materiales y Residuos en la Industria de la Construcción: Fomentando la Economía Circular mediante la Adopción de Inteligencia Artificial y Sistemas Inteligentes*" (ref. SI4/PJI/2024-00211). The project is supported by the *Comunidad de Madrid* through a direct grant agreement aimed at fostering and promoting research and technology transfer at the *Universidad Autónoma de Madrid*. Additionally, the research leading to these results received funding from the European Union Horizon Europe Programme with grant agreement No. 101147855 Intelligent and Sustainable Building Management powered by Cross-Sectoral Lifecycle (DATAWiSE).

References

Anwar, R., Kwon, J.-W., & Kim, W.-T. (2025). A deep reinforcement learning-based concurrency control of federated digital twin for software-defined manufacturing systems. *Applied Sciences, 15*(15), 8245. https://doi.org/10.3390/app15158245

Babai, M. Z., Jemai, Z., & Dallery, Y. (2011). Analysis of order-up-to-level inventory systems with compound Poisson demand. *European Journal of Operational Research, 210*(3), 552–558. https://doi.org/10.1016/j.ejor.2010.10.004

Badakhshan, E., Mustafee, N., & Bahadori, R. (2024). Application of simulation and machine learning in supply chain management: A synthesis of the literature using the Sim-ML literature classification framework. *Computers and Industrial Engineering, 198*(October), Article 110649. https://doi.org/10.1016/j.cie.2024.110649

Bakhshi, S., Ghaffarianhoseini, A., Ghaffarianhoseini, A., Najafi, M., Rahimian, F., Park, C., & Lee, D. (2024). Digital twin applications for overcoming construction supply chain challenges. *Automation in Construction, 167*(August), Article 105679. https://doi.org/10.1016/j.autcon.2024.105679

Brandenburg, M., Govindan, K., Sarkis, J., & Seuring, S. (2014). Quantitative models for sustainable supply chain management: Developments and directions. *European Journal of Operational Research, 233*(2), 299–312. https://doi.org/10.1016/j.ejor.2013.09.032

Burgos, D., & Ivanov, D. (2021). Food retail supply chain resilience and the COVID-19 pandemic : A digital twin-based impact analysis and improvement directions. *Transportation Research Part E, 152*(June), Article 102412. https://doi.org/10.1016/j.tre.2021.102412

Busoniu, L., Babuska, R., & De Schutter, B. (2008). A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(2), 156–172. https://doi.org/10.1109/TSMCC.2007.913919.

Chaharmahali, G., Ghandalipour, D., & Jasemi, M. (2022). Modified metaheuristic algorithms to design a closed-loop supply chain network considering quantity discount and fixed-charge transportation. *Expert Systems With Applications, 202* (April), Article 117364. https://doi.org/10.1016/j.eswa.2022.117364

Chen, L., Dong, T., Peng, J., & Ralescu, D. (2023). Uncertainty analysis and optimization modeling with application to supply chain management. A systematic review. *Mathematics, 11*(11). https://doi.org/10.3390/math11112530

Ciano, M. P., Peron, M., Panza, L., & Pozzi, R. (2025). Industry 4.0 technologies in support of circular Economy: A 10R-based integration framework. *Computers and Industrial Engineering, 201*(January), Article 110867. https://doi.org/10.1016/j.cie.2025.110867

Cobb, B. R., Rumı, R., & Salmero, A. (2013). Inventory management with log-normal demand per unit time. *Computers & Operations Research, 40*, 1842–1851. https://doi.org/10.1016/j.cor.2013.01.017

de Lima, F. A., Seuring, S., & Sauer, P. C. (2021). A systematic literature review exploring uncertainty management and sustainability outcomes in circular supply chains. *International Journal of Production Research, 60*(19), 6013–6046. https://doi.org/10.1080/00207543.2021.1976859

del Real Torres, A., Andreiana, D. S., Ojeda Roldán, Á., Hernández Bustos, A., & Acevedo Galicia, L. E. (2022). A Review of Deep Reinforcement Learning Approaches for Smart Manufacturing in Industry 4.0 and 5.0 Framework. *Applied Sciences (Switzerland), 12*(23). https://doi.org/10.3390/app122312377

Demir, E., Bektas, T., & Laporte, G. (2014). A review of recent research on green road freight transportation. *European Journal of Operational Research, 237*, 775–793. https://doi.org/10.1016/j.ejor.2013.12.033

Devika, K., Jafarian, A., & Nourbakhsh, V. (2014). Designing a sustainable closed-loop supply chain network based on triple bottom line approach : A comparison of metaheuristics hybridization techniques. *European Journal of Operational Research, 235*(3), 594–615. https://doi.org/10.1016/j.ejor.2013.12.032

Dey, B. K., Yilmaz, I., & Seok, H. (2022). A sustainable supply chain integrated with autonomated inspection. *Flexible Eco-Production, and Smart Transportation. Processes, 10*(9). https://doi.org/10.3390/pr10091775

Dolgui, A., Ivanov, D., & Sokolov, B. (2017). Ripple effect in the supply chain : An analysis and recent literature. *International Journal of Production Research, 7543* (October), 1–17. https://doi.org/10.1080/00207543.2017.1387680

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2017). Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v32i1.11794

Fransoo, J. C., & Rutten, W. G. M. M. (1994). A Typology of Production Control Situations in Process Industries. *International Journal of Operations & Production Management, 14*(12), 47–57. https://doi.org/10.1108/01443579410072382.

Galbraith, J. R. (1974). Organization design: An information processing view. *Interfaces, 4*(3), 28–36. https://doi.org/10.1287/inte.4.3.28

Gálvez-Martos, J.-L., Styles, D., Schoenberger, H., & Zeschmar-Lahl, B. (2018). Construction and demolition waste best management practice in Europe. *Resources, Conservation and Recycling, 136*, 166–178. https://doi.org/10.1016/j.resconrec.2018.04.016

Geng, S., Huang, S., Guo, Y., Qian, W., Fang, W., Zhang, L., & Wang, S. (2025). Digital twin driven dynamic scheduling of discrete manufacturing workshop with transportation resource constraint using multi-agent deep reinforcement learning. *Robotics and Computer-Integrated Manufacturing, 95*(March), Article 103042. https://doi.org/10.1016/j.rcim.2025.103042

Govindan, K., Jafarian, A., & Nourbakhsh, V. (2015a). Bi-objective integrating sustainable order allocation and sustainable supply chain network strategic design with stochastic demand using a novel robust hybrid multi-objective metaheuristic. *Computers and Operations Research, 62*, 112–130. https://doi.org/10.1016/j.cor.2014.12.014

Govindan, K., Soleimani, H., & Kannan, D. (2015b). Reverse logistics and closed-loop supply chain: A comprehensive review to explore the future. *European Journal of Operational Research, 240*(3), 603–626. https://doi.org/10.1016/j.ejor.2014.07.012

Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review, 55*(2), 895–943. https://doi.org/10.1007/s10462-021-09996-w

Gu, W., Liu, S., Guo, Z., Yuan, M., & Pei, F. (2024). Dynamic scheduling mechanism for intelligent workshop with deep reinforcement learning method based on multi-agent system architecture. *Computers and Industrial Engineering, 191*(February), Article 110155. https://doi.org/10.1016/j.cie.2024.110155

Guide, D., & Van Wassenhove, L. (2009). The evolution of closed-loop supply chain research. *Operations Research, 57*, 10–18. https://doi.org/10.1287/opre.1080.0628

Hassini, E., Surti, C., & Searcy, C. (2012). A literature review and a case study of sustainable supply chains with a focus on metrics. *International Journal of Production Economics, 140*(1), 69–82. https://doi.org/10.1016/j.ijpe.2012.01.042

Ivanov, D. (2023). Intelligent digital twin (iDT) for supply chain stress-testing, resilience, and viability. *International Journal of Production Economics, 263*(June), Article 108938. https://doi.org/10.1016/j.ijpe.2023.108938

Ivanov, D., & Dolgui, A. (2020). A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0. *Production Planning & Control*, 1–14. https://doi.org/10.1080/09537287.2020.1768450

Krenczyk, D. (2024). Deep reinforcement learning and discrete simulation-based digital twin for cyber–physical production systems. *Applied Sciences (Switzerland), 14*(12). https://doi.org/10.3390/app14125208

Kreuzer, T., Papapetrou, P., & Zdravkovic, J. (2024). Artificial intelligence in digital twins—A systematic literature review. *Data and Knowledge Engineering, 151* (December 2023), 102304. https://doi.org/10.1016/j.datak.2024.102304.

Kuo, H. A., Hong, T. Y., & Chien, C. F. (2025). A deep reinforcement learning based digital twin framework for resilient production planning under demand uncertainty and an empirical study in semiconductor wafer fabrication. *Computers and Industrial Engineering, 208*(July), Article 111389. https://doi.org/10.1016/j.cie.2025.111389

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43*(4), 546–558. https://doi.org/10.1287/mnsc.43.4.546

Li, X., Krivtsov, V., Pan, C., Nassehi, A., Gao, R. X., & Ivanov, D. (2024). End-to-end supply chain resilience management using deep learning, survival analysis, and explainable artificial intelligence. *International Journal of Production Research, 63*(3), 1174–1202. https://doi.org/10.1080/00207543.2024.2367685

Liberopoulos, G., Tsikis, I., & Delikouras, S. (2010). Backorder penalty cost coefficient "'b"': What could it be? *Intern. Journal of Production Economics, 123*(1), 166–178. https://doi.org/10.1016/j.ijpe.2009.07.015

Liu, M., Fang, S., Dong, H., & Xu, C. (2021). Review of digital twin about concepts, technologies, and industrial applications. *Journal of Manufacturing Systems, 58*(PB), 346–361. https://doi.org/10.1016/j.jmsy.2020.06.017

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems* (p. 30).

Madani, Z., Goodarzian, F., Navaei, A., & Ali, I. (2026). Optimization modelling for a sustainable closed - loop supply chain network using IoT : multiobjective metaheuristic algorithms. In *Central European Journal of Operations Research* (Vol. 34, Issue 1). Springer Berlin Heidelberg. https://doi.org/10.1007/s10100-024-00942-z.

McKinnon, A. C., & Piecyk, M. I. (2009). Measurement of CO2 emissions from road freight transport : A review of UK experience. *Energy Policy, 37*(10), 3733–3742. https://doi.org/10.1016/j.enpol.2009.07.007

Mingorance, R., Pereira, D. C., Uribetxebarria, J., & Leturiondo, U. (2025). A methodology leveraging digital twins to enhance the operational strategy of manufacturing plants in unexpected scenarios. *Results in Engineering, 27*(June), Article 106761. https://doi.org/10.1016/j.rineng.2025.106761

Moro, A., & Lonza, L. (2018). Electricity carbon intensity in European Member States : Impacts on GHG emissions of electric vehicles. *Transportation Research Part D, 64* (July 2017), 5–14. https://doi.org/10.1016/j.trd.2017.07.012

Ngwu, C., Liu, Y., & Wu, R. (2025). Reinforcement learning in dynamic job shop scheduling: A comprehensive review of AI-driven approaches in modern manufacturing. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-025-02585-6

Oroojlooy, A., & Hajinezhad, D. (2023). A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence, 53*(11), 13677–13722. https://doi.org/10.1007/s10489-022-04105-y

Ouahabi, N., Chebak, A., Kamach, O., & Zegrari, M. (2025). Dynamic production scheduling and maintenance planning under opportunistic grouping. *Computers and Industrial Engineering, 199*(February 2024), 110646. https://doi.org/10.1016/j.cie.2024.110646.

Pan, J., Zhong, R., Hu, B., Feng, Y., Zhang, Z., & Tan, J. (2025). Smart scheduling of hanging workshop via digital twin and deep reinforcement learning. *Flexible Services and Manufacturing Journal, 37*(1), 157–178. https://doi.org/10.1007/s10696-024-09543-z

Pan, Y. H., Qu, T., Wu, N. Q., Khalgui, M., & Huang, G. Q. (2021). Digital Twin Based Real-time Production Logistics Synchronization System in a Multi-level Computing Architecture. *Journal of Manufacturing Systems*, *58*(November 2020), 246–260. https://doi.org/10.1016/j.jmsy.2020.10.015.

Peng, B. (2024). Navigating green horizons: An empirical exploration of business practices aligned with environmental goals in the era of sustainable economy. *Managerial and Decision Economics, 45*(7), 4732–4752. https://doi.org/10.1002/mde.4284

Peng, H., Shen, N., Liao, H., Xue, H., & Wang, Q. (2020). Uncertainty factors, methods, and solutions of closed-loop supply chain — a review for current situation and future prospects. *Journal of Cleaner Production, 254*, Article 120032. https://doi.org/10.1016/j.jclepro.2020.120032

Pires, F., Leitão, P., Moreira, A. P., & Ahmad, B. (2023). Reinforcement learning based trustworthy recommendation model for digital twin-driven decision-support in manufacturing systems. *Computers in Industry, 148*(February), 2. https://doi.org/10.1016/j.compind.2023.103884

Pishvaee, M. S., Farahani, R. Z., & Dullaert, W. (2010). A memetic algorithm for bi-objective integrated forward / reverse logistics network design. *Computers and Operations Research, 37*(6), 1100–1112. https://doi.org/10.1016/j.cor.2009.09.018

Ragatz, G. L., & Mabert, V. A. (1988). An evaluation of order release mechanisms in a job-shop environment. *Decision Sciences, 19*(1), 167–189. https://doi.org/10.1111/j.1540-5915.1988.tb00260.x

Rajwar, K., Deep, K., & Das, S. (2023). An exhaustive review of the metaheuristic algorithms for search and optimization: Taxonomy, applications, and open

challenges. *Artificial Intelligence Review, 56*(11), 13187–13257. https://doi.org/10.1007/s10462-023-10470-y

Rondinel-Oviedo, D. R. (2021). Construction and demolition waste management in developing countries: A diagnosis from 265 construction sites in the Lima Metropolitan Area. *International Journal of Construction Management*, 1–12. https://doi.org/10.1080/15623599.2021.1874677

Sajadieh, S. M. M., & Noh, S. D. (2025). A review of digital twin integration in circular manufacturing for sustainable industry transition. *Sustainability (Switzerland), 17*(16). https://doi.org/10.3390/su17167316

Schroer, K., Ahadi, R., Ketter, W., & Lee, T. Y. (2025). Data-driven planning of large-scale electric vehicle charging hubs using deep reinforcement learning. *Transportation Research Part C: Emerging Technologies, 177*(April), Article 105126. https://doi.org/10.1016/j.trc.2025.105126

Siatras, V., Bakopoulos, E., Mavrothalassitis, P., Nikolakis, N., & Alexopoulos, K. (2024). *Production Scheduling Based on a Multi-Agent System and Digital Twin: A Bicycle Industry Case, 15*. https://doi.org/10.3390/info15060337

Simard, V., Rönnqvist, M., LeBel, L., & Lehoux, N. (2023). Improving the decision-making process by considering supply uncertainty – a case study in the forest value chain. *International Journal of Production Research, 62*(3), 665–684. https://doi.org/10.1080/00207543.2023.2169382

Stevenson, M., Hendry, L. C., & Kingsman, B. G. (2005). A review of production planning and control: The applicability of key concepts to the make-to-order industry. *International Journal of Production Research, 43*(5), 869–898. https://doi.org/10.1080/0020754042000298520

Syntetos, A. A., Boylan, J. E., & Croston, J. D. (2005). On the categorization of demand patterns. *Journal of the Operational Research Society, 56*(5), 495–503. https://doi.org/10.1057/palgrave.jors.2601841

Tadikamalla, P. R. (1979). The lognormal approximation to the lead time demand in inventory control. *Omega, 7*(6), 553–556. https://doi.org/10.1016/0305-0483(79)90074-4

Talla, A., & McIlwaine, S. (2024). Industry 4.0 and the circular economy: Using design-stage digital technology to reduce construction waste. *Smart and Sustainable Built Environment, 13*(1), 179–198. https://doi.org/10.1108/SASBE-03-2022-0050

Tang, H., Cheng, M., Bhatti, U. A., Xu, B., Zhou, N., Guo, R., & Wei, B. (2025). Digital twin-driven reinforcement learning-based operational management for customized manufacturing. *Engineering Applications of Artificial Intelligence, 159*(July). https://doi.org/10.1016/j.engappai.2025.111754

Thürer, M., Stevenson, M., & Silva, C. (2011). Three decades of workload control research: A systematic review of the literature. *International Journal of Production Research, 49*(23), 6905–6935. https://doi.org/10.1080/00207543.2010.519000

Timperi, M., Kokkonen, K., & Hannola, L. (2024). Digital twins for environmentally sustainable and circular manufacturing sector: Visions from industry professionals. *Production and Manufacturing Research, 12*(1). https://doi.org/10.1080/21693277.2024.2428249

Vrijhoef, R., & Koskela, L. (2000). The four roles of supply chain management in construction. *European Journal of Purchasing & Supply Management, 6*(3–4), 169–178. https://doi.org/10.1016/S0969-7012(00)00013-7

Xiao, J., Zhang, Z., Terzi, S., Tao, F., Anwer, N., & Eynard, B. (2025). Multi-scenario digital twin-driven human-robot collaboration multi-task disassembly process planning based on dynamic time petri-net and heterogeneous multi-agent double deep Q-learning network. *Journal of Manufacturing Systems, 83*(September), 284–305. https://doi.org/10.1016/j.jmsy.2025.09.011

Xu, C., Tang, Z., Yu, H., Zeng, P., & Kong, L. (2023). Digital Twin-Driven Collaborative Scheduling for Heterogeneous Task and Edge-End Resource via Multi-Agent Deep Reinforcement Learning. *IEEE Journal on Selected Areas in Communications, 41*, 3056–3069. https://doi.org/10.1109/JSAC.2023.3310066

Yan, Q., Wang, H., & Wu, F. (2022). Digital twin-enabled dynamic scheduling with preventive maintenance using a double-layer Q-learning algorithm. *Computers and Operations Research*, *144*(July 2021), 105823. https://doi.org/10.1016/j.cor.2022.105823.

Yang, C., Yu, H., Zheng, Y., Feng, L., Ala-Laurinaho, R., & Tammi, K. (2025). A digital twin-driven industrial context-aware system: A case study of overhead crane operation. *Journal of Manufacturing Systems*, *78*(August 2024), 394–409. https://doi.org/10.1016/j.jmsy.2024.12.006.

Yuan, M., Zhang, Z., Mao, K., Ye, Y., & Pei, F. (2025). Digital-twin-based dynamic flexible job shop scheduling problem via multi-agent proximal policy optimisation. *Digital Twin, 2*(2). https://doi.org/10.1080/27525783.2025.2507007

Zhang, W., Peng, Z., Zhao, F., Feng, B., & Mei, X. (2026). A novel deep reinforcement learning framework based on digital twins for dynamic job shop scheduling problems. *Expert Systems with Applications, 296*(PA), Article 128708. https://doi.org/10.1016/j.eswa.2025.128708

Zhu, M., Calderon, C., Ford, A., Robson, C., & Jin, J. (2025). Digital Twin for resilience and sustainability assessment of port facility. *Sustainable and Resilient Infrastructure, 00*(00), 1–34. https://doi.org/10.1080/23789689.2025.2526928