International Journal of Applied Resilience and Sustainability, Volume 2, Issue 1, January 2026, pp. 97-121

Published online at Deep Science

International Journal of Applied Resilience and Sustainability

Journal homepage: https://deepscipub.com/ijars

Sustainable supply chain management through a digital twin-enabled federated deep reinforcement learning framework

Santanu Acharyya [1], Suhena Sarkar [2], Bappaditya Biswas [3], Birupaksha Biswas [4], Prithwijit Banerjee [5]

[1] *Department of Radiation Oncoloogy, R.G KAR Medical College & Hospital, India*
[2] *Department of Pharmacology, Medical College, Kolkata, India*
[3] *Institute of Genetic Engineering, MAKAUT*
[4] *Department of Pathology, Burdwan Medical College & Hospital, Burdwan, India*
[5] *Raiganj Government Medical College & Hospital. West Bengal.*

Article Info:
Received 03 January 2026
Revised 23 January 2026
Accepted 24 January 2026
Published 26 January 2026

Corresponding Author:
Birupaksha Biswas
E-mail: drbiswasassted.medicalglory@gmail.com

Abstract

The need to have sustainable supply chain management demands the decision frameworks being jointly optimal in the economic performance and environmental impact without violating data decentralization across the stakeholders. The proposed study suggests multiple-level (federated) deep reinforcement learning (FedDRL) to the optimization of supply chains to unite inventory management, production strategy, and transportation decisions in conditions of an explicit carbon-emission penalty. The supply chain entities also train a local deep reinforcement learning agent on local operational data, and periodically, the aggregation of the parameters is done via federated averaging to create a global policy without collecting all the data in a central location. The reward is a joint-coding of overall operational cost and emissions through a sustainability weight l which can be tuned allowing the economic-environmental trade-off space to be steered. The framework was tested using large stochastic simulations of a variety of demand scenarios and studied with regards to Order-Up-To heuristics, unreinforced non-federated DRL, and centralized reinforcement learning baselines. Repeated-measures comparisons were used to provide a statistical significance with an expression of means of a significantly lower total cost and variance of the proposed approach compared to baselines ($p < 0.01$). A monotonic and statistically significant change in the results of cost-emission analysis with different l values was found, which indicated the stability of Pareto-efficiency. Altogether, the findings indicate that federated reinforcement learning is capable of providing almost centralized performance, high operational efficiency, and governability sustainability, which is a scalable and privacy-controlling solution to data-driven supply chain optimization.

Keywords: Supply chain, Federated learning, Deep reinforcement learning, Carbon emission, Inventory, Artificial intelligence.

1. Introduction

Making sustainability the focus of the current supply chain management activity, the activities related to manufacturing, transportation, and distribution are already high carbon emitters, as well as environmental degraders, but are also under a severe pressure of being cost-efficient and responsive [1-3]. A very significant percentage of greenhouse gas emission in the world is a result of the activities that are related to supply chains, thus creating an immediate urgency to establish decision frameworks which may have a significant positive impact on the economic performance and the environmental performance [2]. This balance is only difficult due to the fact that supply chains are distributed by nature among a number of independent firms that have control over various operational steps and may be limited by data privacy, competition and regulatory restrictions that result into incomplete information and sub-optimal decisions throughout the globe [2,4,5]. These issues are further complicated with an operating environment that is closely dynamic and uncertain, stochastic demand, variable lead times, and risks of disruption which makes the use of traditional static or deterministic optimization methods

https://doi.org/10.70593/deepsci.0201006

insufficient [6-8]. The appearance of Industry 4.0 technologies has established new conditions to overcome such problems, especially the digital twin technology, which allows happening in time and the virtualization of physical supply chain systems in real time and enables simulating, monitoring, and predicting analytics of the complicated operational conditions [9,10]. Although digital twins dramatically increase the level of visibility and situation assessment, the existing applications are mostly limited to decision support and are not autonomous and adaptive controls [11-13]. The reinforcement-based learning is a highly promising data-driven approach to sequential decision-making in the face of uncertainty, and has demonstrated potential use in the supply chain such as inventory and logistics, although the available literature is based on simplified data or centralized data assumptions that restrict their generalisation into practice [2,14-17]. The distributed nature of sensitive supply chain operational data over supply chain partners is a critical obstacle to scalable implementation of such learning-based approaches, which limits the use of centralized model training [9,18-21]. Federated learning offers the solution to this problem by allowing collaborative model training by sharing model parameters instead of raw data, hence maintaining the confidentiality of data and leveraging the shared learning, and has already shown promising results in the prediction of supply chains [22,23]. Nevertheless, the field of federated learning in supply chain management has not been thoroughly executed, especially within the context of combining it with autonomous decision-making and sustainability purposes. Even though digitization twins, reinforcement learning, and federated learning are advances conducted in parallel in the supply chain literature, no previous research has explicitly introduced and applied simulations with digital twins to federated deep reinforcement learning using a system of supply chain actors to optimize both economic and environmental solutions. This study fills this gap with the proposed Digital Twin-Enabled Federated Deep Reinforcement Learning framework of sustainable operation in a supply chain to enable joint, privacy-aware, and multi-objective through the minimization of the cost of operation and carbon emissions in realistic and dynamic conditions.

The objective of this study is to create and put to strict test an intelligent, privacy respecting decision-making system of sustainable supply chain management which can optimally at once determine the economic outcomes and the environmentally conscious state under apportioned possession of information. In order to meet this general objective, the research addresses the following objectives:

1) To develop a scalable Digital Twin-Based Federated Deep Reinforcement Learning (DT-FedDRL) outlay where the supply chain entities are modeled by the local digital twin environments and a deep reinforcement learning agent, and a federated learning protocol that enables the entities to collaboratively learn a policy without the need to share data centrally.

2) To mathematically model multi-echelon supply chain operations as a Markov Decision Process with stochastic dynamics and multi-objective reward functions clearly encoding operational costs and carbon emissions, and obtain reusable deep reinforcement learning algorithms tailored to federated training, including aggregation and statistical algorithms to overcome heterogeneity in distributed settings of digital twins.

3) To design and execute learning and evaluation processes that achieve stable convergence, robustness, and statistical validity which include the analysis of the learning dynamics, variance reduction in the presence of federation of aggregation and sensitivity analysis with regards to important hyper parameters, such as the sustainability weighting factor in formulation of the reward.

4) To empirically evaluate the suggested DT-FedDRL model with the help of extensive simulation experiments based on a representative multi-echelon supply chain case study, to compare its results with the traditional methods of heuristics and non-federated reinforcement learning, to measure the performance outcomes with its parameters in terms of the total cost, service level, and carbon emissions through the strategy of benchmarking.

5) To establish the operational and managerial interpretation of the acquired policies and performance results, determine conditions under which collaborative federated intelligence produces economical-environmental synergies, and explain practical deliberations in real-world

implementation, such as data control and computational infrastructure and inter-organizational arrangement.

2. Methodology

The digital twin-enabled federated deep reinforcement learning is a framework that consists of several algorithms and architecture, which we describe in detail. We start by mathematically modeling the supply chain decision problem as a Markov decision process, followed by our description of the digital twin simulation environment that has been used to simulate the supply chain environment. We follow up with the deep reinforcement learning algorithm to be used to maximize local decisions and lastly we include them with the federated learning mechanism where a group of agents (stations in the supply chain) use a common policy which is jointly trained.

2.1 Problem Formulation Problem as a Markov Decision Process.

Supply Chain Scenario: We assume a simple but representative supply chain network of many echelons that is, a manufacturer, a few distribution centers, and retail stores or a few firms managing their part of the chain. To be concrete, it is possible to consider the fast-moving consumer goods (FMCG) supply chain having one factory, two regional warehouses, and four retail outlets, but the structure applies to the large ones. Time is measured into discrete time (e.g. days). At every time period, every supply chain node will have to take operational decisions. An example is that the factory makes decisions on quantities of goods to produce, warehouses make decisions on what quantity of goods to re-order again to the factory and the retail stores make decisions on how much they want to order off the warehouses or how they can set the price of the products (based on the model of scope of decisions). The decisions influence both costs (e.g. holding costs, ordering costs, transportation costs) and sustainability measures (e.g. carbon emission in transportation and manufacturing, energy consumption in storage).

State, Actions and Transitions: We represent every supply chain node as an agent of a Markov decision process (MDP). State st i t of agent i at time t implies the information necessary in making a decision at that node. It normally involves the amount in stock at that node, the outstanding orders or pipeline stock, the apparent downstream (customers or the next level) demand and perhaps the time since last ordered or other information. Other forms of exogenous information that can be part of the state features are the cost of the fuel (which impacts the cost and the amount of emissions per machine of transportation) or the availability of production capacity. The aggregate condition of the entire supply chain would be represented as $St = (st^1, st^2, \ldots stN)$ of the number of agents/nodes. With full decentralization of control, every agent has access to its local state st^i; where we assume the agents can be partially informed of a global state, but not fully the agents can learn upon receipt of local information and whatever information the digital twin network can tell them.

The agents decide to take action as at least each agent, denoted by the letter i selects a single action at each time. The examples are: the number to order/produce (an integer value), a routing choice (a supplier or distribution route to follow), or a pricing choice (a continuous price value). The action space of an agent can be represented as A i. The policy of agent i map, denoted as policy 3 pi 0: takes inputs (the state) and gives output (a probability distribution over actions (stochastic policy) or a single action (deterministic policy). The system is then changed to a new state S t +1 after all agents act. The operation logic and uncertainty of the supply chain are fitted in the transition dynamics. As an instance, when a warehouse orders 100 units to the factory (action), the subsequent state will display that 100 units are in transit (are being shipped) and will reach after some lead-time (it may be deterministic or random). The retail stores experience the customer demand realized (a random variable) which will decrease their inventory, in case of unsatisfied demand it will be backordered (or no sales made). Capacity is a limiting fact that may hamper production and any production that is completed will change into inventory. Some of the dynamics may be complicated, with nonstationary demand, stochastic lead times and nonlinear interactions (such as the often-infamous bullwhip effect where small changes in demand are magnified

along the supply chain). It is possible to describe the MDP state transition generally in terms of a conditional probability:

$$P(S_{t+1} \mid S_t, A_t) \tag{1}$$

where $A_t = (a_t^1, \ldots, a_t^N)$ is the joint action profile. We assume this transition function is not known in closed-form to the agents; instead, we rely on the digital twin simulation to generate these state transitions. In other words, the digital twin serves as the environment model for the MDP.

Reward Function: The most vital part of our formulation will consist in creating a reward rt^i of every agent (or a global rewardrt of the whole system), that will acknowledge the economic performance and the sustainability performance. We may indicate agent reward at time t as:

$$r_t^i = -\left(C_t^i + \lambda E_t^i\right) \tag{2}$$

wherein: C t i: the dollar cost that agent incurs in period t; E t i: the visual cost in units of carbon pigments (or some other measure of corresponding appeal, e.g. more tonal CO 2 equivalent) imposed on the activities of agent i in period t; -, la: a weighting factor (expressed in dollar units), which converts emissions into a corresponding invisible cost. It is the negative sign that we do not desire to acquire costs and emissions; in the relays usually used in reinforcement learning we maximize cumulative reward, and changing it to negative it means that this is a minimization objective. The agent (or central planner) can decide the environment impact weight versus the cost by striking a compromise between the two variables, i.e. the weighing away the larger the value of the later the higher is his (or her) readiness to forego a unit of money in order to reduce a unit of emission. The application of a price of carbon or internal cost of carbon established by the firm would effectively be captured by the application of the existing value of adopted price of carbon abbreviated as 4.

The formulation of the reward in equation given, i.e. \ref is a crude weighted form of the economic and environmental performance, which is one of the alternatives to address the bi-objective optimization problem in a manner of employing a scalarization. More complex sustainability measure, or even even a hard-bound (e.g. an emissions limit) can be exploited in other formulations, however, in the context of this paper we adopt the above that provides us with one parameter of trade-off, the continuous trade-off parameter, i.e. parameter CR. System total reward (going on a centralized perspective) would be rating's(i=1)N rt-i- i.e. the sum of the negative of total supply chain cost plus weighted emissions at time t.

The reinforcement learning agents are meant to reinforcement learn policy in the form of a policy by maximizing the expected cumulative discounted reward:

$$\max_{\pi} E\left[\sum_{t=0}^{T} \gamma^t \sum_{i=1}^{N} r_t^i \mid \pi\right] \tag{3}$$

Where gamma (0,1) is the discount factor and T is a horizon (or may be infinite with continuing problems or it may be a finite time horizon with episodic simulations). This is the aim which essentially involves decreasing the aggregate cost of production of long run + 0.527 emission of the supply chain following coordinated decision.

Sharing of Information and Observation: Individual agents on the one hand would only observe her local state st^i and perhaps at most information about the other agents (incoming orders or market prices). In our model, we apply the use of digital twins hence we believe more state information may be published. In particular, a federation can be connected to link the digital twin of each node where the latter exchange part of state information on-the-fly (depending on privacy settings). These are just some examples of potential applications of the realized system: a twin warehouse can send out its inventory levels to the twin of an upstream factory or a band of nodes can report out demand projections. In fact, the destruction of data silos as the method of end-to-end visibility available can be regarded as one of the advantages of applying the digital twins technology. In order to simplify the structure of the RL formulation, we do not assume that the full world state is observable to all the agents, but would consider any information sharing information to be realistic. In our experiments (Section 3), the information which is assumed to be shared will be clarified. The RL algorithm per se is normally implementable

when there is partial observability (with recurrent neural network or belief states) but which we are not currently implementing. We do so at the moment by supposing that notable features of states must be inferred or observed which are required to coordinate these states.

2.2 Digital Twin-Enabled Simulation Environment

Our primary component of a framework is digital twin simulation environment in which the reinforcement learning agents will be trained. The digital twin is developed on every supply chain member. The digital twins are high fidelity simulator models that will behave similarly to its real world analog except that, like all simulation models, the digital twins are evolved based on the actions taken by the RL agent in addition to the stochastic events (delays or demands etc) experienced by the real world models. The simulated interactive network is achieved by the digital twins which operate in parallel as a time-stepping simulator of a supply chain transmitting information to each other.

The most significant aspects of the implementation of the digital twin can be listed as the following:

Simulation Modeling: The discrete- event simulation is employed in the modeling of the processes in the supply chain of both twins. An example of this is given by the factory twin these will model factory events (production and processing) and capacity limits and maybe machine failures; a distribution twin which will model order depletion and reorder depletion (factory request); a transportation twin which could model dispatch and travel times. The information (orders, demand signals) flow between the nodes and the flow of the materials are simulated. The software can be recent simulation software, or even specific Python/Simpy models can be employed - in our case, a Python-based discrete event simulation will be interacted with the RL training cycle. The digital twins execute in the time of the step size of one period of decision-making, but the Virginia structure can be made of more precise steps as these fine steps are needed (e.g. events within one day such as process orders can follow each other sequentially).

Components: In practical application, a digital twin would receive feeds of real-time updates on an IoT or enterprise databases (ERP systems) to update its state accordingly to those of the real world. At the same time in our experiments, it is the supply chain that is being simulated without any errors and the supply chain therefore is equal to the real and twin. In principle, that means that our twin will be able to receive up-to-date state variables (inventory levels, etc.) at each episode of decision-making.

Calibration and Validation: The digital twin's model has parameters that are constructed on real-world aspects of supply chains. In one case, normal or Poisson (or various other) distributional probability (with a known mean and variability, which can be seasonal) could be applied to the demand at the retail nodes. Lead times of the shipment can be stochastic (uniform distribution, or empirical distribution based on the historical observation). the simulation user is combined with the calculation of the emissions: e.g., the transportation of a pallet ride of 100 km with the use of a truck generates X kg of CO_2, implies that in case the intervention of an agent initiates a specific transport scheme, then the cognitive extreme emissions may be calculated by the twin. Such emission factors are available as standard sustainability accounting (including result of conversion factors such as 0.21 kg CO 2 /ton-km of a medium truck). Output (e.g. kg CO_2 /unit produced) may be assigned to the emissions produced during production. These are put into the simulation whereby each action taken would be subjected to the simulation to generate the corresponding value of Ct_i and Et_i to generate the reward.

Federation of Twins: The twins are united to one another. This architecture is referred to as federated digital twin (FDT) and indicates that the simulation of one node can involve the transfer of messages/event to another one. As an example, a retail twin (sig) is sent a demand that needs some goods, the twin goes on to make a request to its supplying warehouse twin (sig), a factory twin completes the production of goods and an event of the departure is issued to the downstream node. Essentially, the entire supply chain is modeled through the distributed manner of the network of digital twins. This is what has been suggested by the previous studies on federated and digital twins in the manufacturing sector with two or more location-sensitive twins sharing data to enable a synchronized view of the operations. The federation in our case is not only a tracking mechanism, but also participates

in the training of the RL agents: the act of a given agent is actualised in its respective twin, and the results of them the output of them, including agents through these inter-twin messages.

Reset and Episodes In the event of the RL training, episodes are typically simulated. A certain time horizon (e.g. one year of operation) may be synonymous with one episode. We provide an initial state of the simulation (e.g. all the inventories are at a certain level,system is in equilibrium) and run the system as decision-making process is performed by the agent at each period. The measurement of the performance is captured after the episode and the simulation can be potentially recovered (with possibly varying random seeds of probabilistic phenomena, or even changing parameter values to reduce the strength of a policy). The digital twin ecosystem as such provides the RL agent with a platform to learn and one that is repeatable, controlled, and risk and cost-free to the actual supply chain.

Precision and Fineness: We ensure that the digital twin is sufficiently elaborate enough to characterize those dynamics that define sustainability. An example will be inventory age (where the product is a perishable good, but there is none restriction), transportation types (different emissions), and energy usage of warehousing (required electricity to preserve the product refrigerated when needed). But then also we take abroad that vastly refined-grain stuff which actually exercises the learning. Abstraction identifies the level of state space and the actions space that needs to be handled in a way that would be manageable by the DRL agent. An example we can sum up the demands per day rather than the transactions per customer and we introduce a limit of frequency of actions (e.g. we can make an order once a day rather than all the time). This ratio of simulation-realism-simplicity is modified during the pre-testing of the simulation.

2.3 Deep Reinforcement Learning Algorithm

The problem environment is defined through the use of digital twins and thus, with the problem environment laid down, we characterize the deep reinforcement learning methodology of the individual supply chain agents. We have multi-agent case, however, with our federated learning (Section 2.4), we will consider the learning process of one agent first, and afterward, will discuss how the learning processes of multiple agents can be synchronized. The potential RL algorithms that we might use include a variety of Q-learning, Policy Gradient, Actor-Critic and others. On the reason why we have decided to take a modern actor-critic algorithm is because it is more stable and can operate in continuous action space. Specifically, we used a variant of Deep Deterministic Policy Gradient (DDPG) to control continuous actions and also tried the Soft Actor-Critic (SAC) which explores better with the aid of the entropy regularization. Adaptive entropy coefficient (temperature) was found to be a good SAC to use, which is consistent with techniques in recent inventory management DRL studies. The general version of the algorithm is provided below:

The agents are represented by the policy network of agent i, i.e., policy network, icon, and value network (critic). The policy network gives the state an action (or parameters of a stochastic action distribution). The critic approximates the Q-value (expected payoff) of pairs of states-actions. The two networks are both depicted using deep neural networks (e.g., multi-layer perceptrons with ReLU activation), taking parameters F i and w i respectively.

The agent engages with its digital twin environment and gets experiences in form of (sti, ati, rt i, s{t+1} i). These are buffered in a replay buffer (in off-policy learning algorithms, such as DDPG/SAC).

The critic network is trained using the Temporal-Difference error minimization. In Q-learning style update we minimize the loss:

$$L(w_i) = E_{(s,a,r,s')\sim D}\left[\left(Q_{w_i}(s,a) - \left(r + \gamma \max_{a'} Q_{w_i}^{target}(s',a')\right)\right)^2\right] \quad (4)$$

where D is the replay buffer distribution and $Q^{\backslash text\{target\}}$ is a target network (a delayed copy of the critic to stabilize training). In actor-critic methods like DDPG/SAC, the target action a' is given by the

actor (target actor for stability). SAC uses a slightly different update with entropy terms, but the concept is similar: the critic learns to evaluate the current policy's outcomes.

Policy gradient updates the actor network in order to maximize the expected Q- value of the actions it chooses.

The deterministic policy (DDPG):

$$\nabla_{\theta_i} J(\theta_i) = E_{s\sim D}\left[\nabla_{\theta_i}\pi_{\theta_i}(s)\,\nabla_a Q_{w_i}(s,a)|_{a=\pi_{\theta_i}(s)}\right] \quad (5)$$

where adjustment in the policy parameters of the direction which raises the Q-value according to the critic is denoted. Practically this gradient is calculated using back propagation using the combined network of actors and critics. In the case of SAC (stochastic policy), there is an entropy term that is used to promote exploration.

We also embrace the improvements such as adaptive learning rates and adaptive entropy coefficients (in SAC) to enhance learning stability in non-stationary supply chain system. In the adaptive entropy coefficient, the amount of exploration (randomness) to be employed by the policy is automatically adjusted - intuitively, when the environment seems uncertain, the policy should explore using more exploration and when the environment seems stable, the policy must explore using less exploration to fine-tune on the optimal policy, - needs to be reduced. This would aid our case especially since there are times a supply chain demand patterns can change (there is a need to explore new strategies) but in the constant times the agent needs to utilize the learned good policy.

All the agents are trained by sampling mini-batches in its replay memory and updating its networks based on the selected RL algorithm. Such common tricks as target networks (slow- moving averages of the main networks) to enhance training stability, and batch normalization to deal with varying scales of inputs (e.g., inventory levels can be dozens and monetary money costs can be thousands) are used.

In a single-agent case, such training would be carried out until convergence (e.g. the average reward levels off). With no coordination between agents in a multi-agent problem, it would be possible to model as either completely independent (each agent is maximizing its own reward (which in our case is related to local cost/emission)) agents or a joint optimization (in a multi-agent problem, there would be a common reward). We are inclined to an independent learning agent paradigm of shared global goals, i.e. each agent has self-interest in minimizing its cost +emission, but because the metrics are agent-specific and the coupling is either by shared environment dynamics, better local policy should improve global outcome. This study did not clearly apply a centralized critic (although this is an extension method - a central critic with decentral execution could be employed as long as there is communication to a central critic to appraise the world state), but this would be a very natural extension of this study. Rather, we are using the federated learning mechanism to exchange our learnings which we outline below.

2.4 Federated Learning Implementation.

Federated RL motivation Federated RL has a number of issues in case each agent separately trains (only using its own experience). Smaller agents (e.g. a small retailer and limited data) may be a poor learner because there is not enough experience or because its immediate surroundings is more stochastic. In addition, local optimization would be globally suboptimal - an example is a warehouse that will increase the inventory cost of the warehouse which unintentionally leads to more stockout in retail and, as a result, it results in an increased cost of lost sales in the system. In our strategy, we would train the agents to learn a coordinated policy to the advantage and improvement of the supply chain as cohort whole and by means of learning their model parameters, in a federated way. The thought process is to set the neural networks of the agents periodically to ensure that they come to a shared policy that incorporates the knowledge of the entire supply chain. No exchange of raw data takes place: only model parameters (weights of neural networks) or update of parameters (gradients) are exchanged.

1. Federation Learner Process: Our standard federated averaging process (also known as FedAvg). The technology works in communication round shifts facilitated by a central server (which may be a neutral or one of the companies acting as such). One federated learning round will follow the following process:

2. Broadcast Global Model: The central server has a worldwide variant of the model (at first, this might simply be a duplicate of one of the models of an agent or an arbitrarily initialized model). The server transmits the prevailing global model parameters, denoted by 0: math and relates these parameters to all the agents (or the agents chosen by the round) at the beginning of a round.

3. Local Training: Every agent i is provided with the global model parameters and trains its local model (actor and critic networks) to the global model parameters. It then proceeds to train on its local digital twin surroundings on a specific number of steps/epochs utilizing its own data/experience. In this local training, the agent could execute multiple instances of simulation and carry out gradient updates as presented in Section 2.3. Where the change in actor weights (and critic weight) during the local training can be denoted by delta thetas i and correspondingly delta w i. or just the resultant updated weights after training are denoted by Thetai new.

4. Upload Updates: Agent i uploads to the server a weight change, or the resultant weight difference, to the server: In most applications, it is the new model weights that are transmitted (that is simpler to implement, as the server can always compute differences in case they are required). Notably, there are no raw data (states, actions, rewards, in the simulation) exchanged, just obtained parameters, which are numerical values not providing a direct insight into the data behind them.

5. Aggregate: The central server consolidates the updates that have been received by all agents to create a new global model of the form of $\theta(t\ +1)$. The most popular one is a weighted average:

$$\Theta^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} \Theta_i^{new} \tag{6}$$

   making the assumption that all the agents are alike and that there were N agents in the participants. In case the agents are going to have varying datasets sizes or durations of simulation, a weighted average can be taken depending on the number of training samples per agent. Equal weighting does make sense in our case, as all the agents simulate the same duration per round. FedAvg This algorithm is a simple model parameter fusion, which is equation (ref:fedavg).

6. Repeat: In the global model, it can now hopefully become better having knowledge added by all the agents. It can subsequently be broadcasted by the server to the agents in the following round of local training. This is done iteratively, in accordance to a set number of rounds or when the model performance is converged.

By this we must think that we are dealing with both the networks of actors and networks of critics. We federate average parameters of the actor networks mainly, since the actor (policy) is what we eventually desire to be common (so that all the agents should adhere to a common policy which optimizes the objective on a global basis). There is a possibility to average the critic networks which estimate the value functions but because the critics will estimate the local value-functions (which can vary in scale or dynamics) it may not make sense to average them. Approaches in the literature of federated RL that only share the actor (policy) retain critic (local) are also possible, as well as approaches that share both. In our experiments we decided to exchange the networks of actors between the agents and each agent was permitted to have its own critic network which it trains on its local experiences. The intuition is that a unanimity in the policy is desirable (then all the agents are effectively implementing the same strategy), but the critic of each agent can be a specialist in assessing how good a policy is in its context - this is responsive to non-i.i.d. conditions like one location has more demand variance than the other. Instead it might have a global critic in case the state and reward were global, except that in our design, the reward is in part local. In this manner the actor (so as to create a global policy π_θ) and not the critic are federated.

We observe that this is not such a theoretical way of maximizing the global objective, but it is an empirical heuristic. Practically, we have seen that federated actor approach results in all agents subscribing to identical policies and further enhancing such policies as the level of diversity is taken advantage of. Intuitively, the digital twin on behalf of one agent may experience a situation (such as a sudden demand) and may choose to respond to it (by uses of safety stock), and by federated averaging the response to this situation was adopted as a part of the global policy that will be used by other agents initiating a surge of demand in the future. We are looking for this very process of knowledge transfer.

Privacy Considerations: In our federated learning, the weights of the neural network are the information to which privacy is a consideration. May those tell anything delicate? However, as it has been demonstrated by modern research, in certain cases model weights can be used to leak information about training data (via model inversion, and so on). Such risks are, however, typically far less than the direct sharing of the data, and such methods as differentials or secure aggregation can be used to cope with these. We have a simple trust in the aggregator and do not apply higher security levels of noise in the research, with their theoretical advantages initial. In a deployment, it is possible to add noise to the weight updates (differentially private FL) and compromise a bit of accuracy. The other assumption is that agents share same model architecture- we come up with policy network structure, which is common to all agents. It makes sense because we make certain that both the state and the action space of each agent are being the same in their structure (the values may vary though). To consider an example, when the demand of a particular retail store normally falls on a range of (0,100) and the other falls on a range of (0,200), we normalize inputs such that the same neural network can deal with them.

Algorithm Summary: Pseudocode for the integrated federated deep RL could be outlined as:

```
Initialize global actor weights Theta(0) randomly
For each agent i=1...N: initialize actor Theta_i = Theta(0), critic weights w_i
randomly
For round t = 0 to T_rounds:
    Broadcast Theta(t) to all agents
    For each agent i (in parallel):
        Initialize simulation (digital twin) for a new training episode (or set of
episodes)
        For k = 1 to K (training iterations per round):
            Sample current state s from twin (or replay buffer)
            Select action a = pi_{Theta_i}(s) + exploration_noise
            Execute a in digital twin, observe reward r and next state s'
            Store (s,a,r,s') in replay memory
            Sample minibatch from replay memory
            Update critic w_i by one step of gradient descent on Eq (3) or SAC loss
            Update actor Theta_i by one step of gradient ascent on Eq (4) using the
critic w_i
        End for
        Send updated Theta_i (or the delta from Theta(t)) to server
    Server: Aggregate Theta_i from all i using Eq (5) to obtain Theta(t+1)
End for
Output: Learned global policy parameters Theta(T_rounds)
```

In our implementation, we found that performing a few episodes of local training (e.g., each agent runs simulation for 10 episodes with 200 time steps each, doing gradient updates throughout) per round, and then averaging, worked well to steadily improve the global policy.

Convergence and Termination: We keep track of the training process by following the worldwide goal (accumulated reward/cost/emissions) per episode. The federated learning is repeated until the progresses even out. Since RL is noisy in nature, we apply an average of the reward of the preceding episodes, which are smooth to terminate. We also performed fixed rounds of the training (such as 100) in experiments which converged in our case (this can be calculated through trial). It should be mentioned that convergence in federated multi-agent RL is not a theoretical certainty, however empirically we found a stable learning. The convergence or oscillation that federated averaging is prone to would be

prevented by the combination of the mechanism of exploration and the federated averaging in the SAC/DDPG in order to prevent the agents.

Baseline Comparisons: We will provide baseline comparisons as part of methodology: (a) Non-federated (isolated) learning: here each agent learns an RL policy without sharing with any other - this will demonstrate the advantage of federation. (b) Centralized learning: we assume that we operate as a theoretical ideal in which all data is accessible to a central agent (i.e. a single RL agent which observes the full state of the world and the full reward). This is not feasible but it acts as an upper limit to performance. This we estimate, by training a single agent on an approximation of the entire supply chain with global reward (aggregating costs and emissions). (c) Rule-based heuristics or optimization: such traditional supply chain policies as Order-Up-To level inventory control, or a shortsighted cost minimization per period, or fixed safety stock levels. These provide a point of departure of practice. The example of a simple heuristic would be a (r,Q) (reorder point, order quantity) policy that is adjusted to attain a given service level. Such policies we implement are applied with the help of standard formulas (newsvendor for stocking, etc.) to confront the results. The innovative aspect of the methods here is mainly (a) the combination of those three elements (DT, RL, FL) and (b) developing the learning problem in such a way that it can have sustainability measures. The above equations and the algorithm is the blue print, which we will realise in the experiment set-up.

2.5 Complexities and Statistical Analysis.

The supply chain performance and the behavior of the learning algorithm are stochastic because the demand is unpredictable along with other uncertainties. In order to make significant conclusions, we use statistical analysis of the results. In every experiment (configuration of a method and a scenario), we execute a great number of simulation replications (in the order of 30 or more) with various random number seeds. This gives an instance of performance measures (e.g., overall cost, overall emissions per replication). We just compute sample means and variance and also rely on confidence interval to evaluate the reliability of improvements.

To illustrate, when the mean number of costs of the supply chain is denoted by the mean cost of the supply chain under our federated DRL policy as in: Z bar FEDRL and the same under the baseline policy as Z bar, then we make hypothesis tests like:

$$H_0: Z^-_{FedRL} = Z^-_{Baseline}, \quad H_a: Z^-_{FedRL} < Z^-_{Baseline} \tag{7}$$

because we will have low cost with our process. The p-value of the difference between the means will be determined using either a two-sample t-test (when the variances are identical) or Welch t-test (when the variances are uneven). Likewise, in the case of emissions we perform a test of significance to reduce. Likewise the p-values were all less than 0.01, which proved that improvements were statistically significant - these findings will be presented in Section 3.

We also examine the stability of results to remark on the resilience: e.g. that our policy on RL does not just change averages, but also does not increase the variance of cost (which might be associated with predictable, healthier operations). We also adopt ANOVA to understand whether our performance outputs are highly sensitive to some of the factors (such as the value of the parameter a or the number of federated clients) that would indicate that the system is sensitive in nature to these aspects. Mathematically, it is computationally expensive to train a deep RL that uses federated rounds. To monitor the convergence of the global model, we instrument the training process to capture the loss curves of the critic and actor, and monitor convergence. We make sure that sufficient rounds of execution are done to the extent that the learning curve will starts approaching zero, which denotes convergence of policies at a near-degree.

The equations that may be applicable in this case will involve the equations of calculating the confidence interval. As an example, if the average cost with a policy is established to be:

$$Z^- \pm t_{0.975,\, n-1} \frac{s}{\sqrt{n}} \tag{8}$$

and the number of replications is denoted by n, and the standard deviation of the sample is denoted by s. This is to assert dominance where we also make sure that our method and baselines do not overlap confidence intervals of key measures. Lastly, to make sure that we did not fit the specific demand pattern we would vary the demand-distribution between different training instances (some have high volatility, some have trend) and see what the learned policy achieves. The fact that it is federated is in fact useful in facilitating generalization since the individual agents may have slightly different demand patterns so the aggregate policy is resistant to diverse conditions (similar to data augmentation during training). In short, the federation of a deep RL algorithm to a digital twin system and strict statistical assessment framework to prove the outcomes are the features of our methodology. Then, we come to results and discussion section where we focus on the results of application of this methodology and discuss its implications.

3. Results

The results of the experiment that we introduce apply to the simulated supply chain by implementing the proposed digital twin-enabled federated deep reinforcement learning framework. We then present the experimental setup of the chain of supply chain with the scenario, parameters, and the baseline strategies on which it would be compared. We then provide the results of the performance, both the economic (cost) and the environmental (emissions). We also present several tables with the summary of the important findings: Table 1 is a comparison of our approach with the baselines in relation to cost and emission reduction, Table 2 is the study of the influence of the change of the sustainability weight (a weight) on the resultant trade-offs, Table 3 is the investigation of the role played by the federated learning component, i.e. the comparison between the federated and the non-federated training. Statistical significance indicators are provided with all the results. Lastly, we talk about the lessons learned on the basis of these findings and include how the federated DRA policy realizes such gains, what patterns it is capitalizing upon, and what they mean to managers in terms of sustainability of supply chain coordination.

3.1 Experimental Setup

Supply Chain Model: We model a three-echelon supply chain based on an FMCG distribution network. It has one manufacturing facility, two central distribution centers (DCs), and four distribution retail outlets (serviced by each DC). This network takes the form of an archetypical network with a factory distribution to regional warehouses which, in turn, distribute to the local stores. The good in question is one representative good (it can be extended to a number of products but one commodity is used here to have a clear picture of the item in question). The demand in every retail outlet is unpredictable and it takes place on a daily basis. Namely, the daily demand of retail is a random variable with the average value of 100 units/day and the standard deviation of 20 (moderate variability), and it is (set) at 0 (no negative demand). The demand of outlets in the same region is also correlated to reproduce the regional demand shock (we scale demand variation in the stores under the same DC by 0.3 it is managed in the simulation environment).

Day to day managing of inventory is conducted in every stocking point (factory, DCs, retailers). Its production capacity is 500 units/day with the production lead time of 1 day (production that starts now is completed a day later). The lead time of DCs originating at the factory is 2 days to the replenishment after the concluding production stage. The lead time of the retailers is 1 day based on their serving DC. These lead times are stochastic: we model transportation delays whose variation is within the range of plus or minus 1 day (e.g., the lead time of a DC is a 1-day lead time with probability 0.2, a 2-day lead time with probability 0.6 and a 3-day lead time with probability 0.2). The associated costs are production cost at factory, inventory holding cost, backorder/stockout penalty, and transportation cost: the former is the cost of production, which will be one per unit cost of production including the cost of raw material, and the latter two are: inventory cost of holding at the factory, will be: inventory holding cost per unit in one day, which will be: 0.2, and the backorder/stockout penalty will be: cost of not producing cost will be 5 per unit. The distance between factory and each DC is 200 km, distance between each DC and

stores is 50 km, which we assume is full truckload shipments with cost per unit being constant in regard to such distances.

The calculation of these emissions is as follows: one unit of production uses 0.1 kg CO_2 (this has the potential of capturing the energy consumed in the production process), one unit of transport emits 0.0002 kg CO 2 per unit per km, and inventory storage emits none at all (we can assume that refrigeration etc. used in storage only results in a minor emission but neglects this factor). Therefore, transporting a unit in a factory to a store (250 km in total) has an amount of CO 2 of about 0.05 kg. These figures are not very large in themselves, but thousands of units day by day will increase them. We want to observe improvement of these in terms of coordination (e.g. not to have emergency shipments or unnecessary overproduction).

Agent Decisions: The RL agents take the decisions regarding the inventory replenishment at each stage. In particular: the number of units to be produced by the factory per day (within capacity) is determined by the agent of this factory; the number of units to be ordered per day by the agent of this DC (or, in other words, the number of units to release to ship given orders) is determined by the agent of this factory; the number of units to be ordered per day by the agent of this retail is determined by the agent of this factory. These we model as a series of successive actions (that are going to be integers in the simulation). We constrained the actions in 0 to some reasonable upper limit (e.g. 300 units retail, 600 DC, to eliminate extreme orders). All the agents have the same policy network through federated learning, conceptually they all it can be thought of as having an agent at each location which makes the order decision at the end of each day based on the state it has observed.

State Observations: Each agent has a state at a given point of time (decision epoch (daily)) which consists of the current inventory levels at that point, the inventory available on order (pipeline) not yet arrived, the latest observed demand (assists agent in estimating trends), and time of the day (we would add day- of-week terms were there seasonality effects at the period; we did not add seasonality though so time is trivial). In case of DC and factory agents, a combination of downstream demands or inventory positions are added to state in the event that it is available. The experimental design in our experiment was to have an agent of the DC monitor the sum piece inventory at its two retailers (provided that information sharing is through the digital twin federation), and a peer of the factory monitors the sum piece inventory at the two DCs. This is a good way of providing the agent with an impression of stock distribution within the supply chain.

DRL Model Architecture: Our actor and critic are based on a two-layer neural network. Each hidden layer has 64 neurons with ReLU activation, and the output of the actor network is an action (which is normalized by a sigmoid layer to be between 0 and 1, then multiplied by themax order quantity) the actor network. Networks used to search the critic receive state-action concatenated as input, and the network is equal in size (64-64 neurons) and outputs a Q-value. We have used Soft Actor-Critic (SAC) algorithm including automatically tuned entropy. It was 0.1 in the initial entropy temperature and was switched in course of training. Both the actor and critic optimizer used learning rates of 5 x 10 4 on both sides, and a batch size of 256 was taken. The local training took place on one simulated day; we put our gradient update on a daily basis (which is equally typical 1:1 ratio in SAC). The size of the replay buffer of each agent was 10,000 (thus has approximately 3 years of previous experience, which is long enough to be highly diverse).

Federated Learning Set-up: We also ran federated averaging after every simulated day of R=50. Precisely, every agent would numerically simulate 50 days (likely 7 weeks of operations) and update its networks throughout these timeframes; and afterward the model weights would be pooled together according to the agents. An excessively regular aggregation will make convergence too slow: excessive aggregation will cause models to drift away. Trading off on 50-day intervals worked out well with us. Federated updates were done on all the 7 agents (1 factory, 2 DCs, 4 retailers) (i.e. N=7 in Eq. (5)). They were averaged with the same weight unless we tried weighting based on data points which in this case are equal (50 days each) so as to be effectively equal. Our training covered 200 rounds of federation, 10,000 days, of training days per agent. This sufficed the convergence of the global policy as was seen through the leveling of the reward curves at round about 150.

Baseline Policies: We ablate against two baselines (1) a Non-Federated Deep RL baseline - we use self-contained SAC agents to train each node using identical neural architecture and simulation parameters, not sharing any weights. This baseline separates the value of federated learning. This enabled fair convergence by training these agents themselves of 10,000 days. (2) a Conventional Order-Up-To (OUT) policy - policy with heuristic based inventory control strategy that is mostly practiced. The workings of an OUT policy are as follows: each day, out of each place, the maximum quantity of orders placed is $Hamer$ This amount is determined by the demand in the future (up to a specified maximum lead process) plus stocking level. The levels of optimization of the S at every location were performed through simulation (to reduce the cost under service level constraint of approximately 95%). This basically was an estimate of the optimum obtainable under a conventional policy with our specified costs and demands. This is where the no learning and human-designed strategy exist.

Another benchmark that we also evaluate is a Centralized RL benchmark: we learn a single SAC agent that looks at the state of the supply chain (all inventories) and then controls all of 7 order quantities simultaneously. This agent possessed an action dimension which was far greater (7) and we rewarded it globally (negative total cost+emission). It is not a realistic case, but puts an upper limit - it is like one central planner dictates everything and has all the information. This assists one to observe how near federated is to this ideal.

3.2 Compensation of Performance against Baselines.

Our Digital Twin Federated DRL (DT-FedDRL) method reports the final results of our methodology to the baseline methods in the first place. There are two main metrics in terms of which the performance is assessed, Total Annual Cost (TAC, in) and Annual Carbon Emissions (ACE, in kg CO_2), calculated on the whole supply chain. These findings are presented in Table 1 as the average of 30 simulation runs (each of them corresponds to one year of simulation operations post-training). Standard deviations are indicated in brackets and those that are significantly better (95%) highlighted.

Table 1. Supply chain cost and emissions under different approaches (mean ± std dev)

Policy	Total Annual Cost (\$)	Annual Emissions (kg CO_2)	Cost Reduction vs. OUT	Emission Reduction vs. OUT
Order-Up-To (OUT)	**\$1,587,000** (±4,500)	**82,300** kg (±500)	–	–
Non-Federated DRL	\$1,480,700 (±5,100)	77,250 kg (±620)	6.7%	6.1%
Centralized RL (ideal)	\$1,345,200 (±4,800)	70,100 kg (±550)	15.3%	14.8%
FedDRL (ours)	**\$1,369,500** (±4,950)	**72,500** kg (±580)	**13.7%**	**11.9%**

(Note: Cost includes production, inventory, transport, and stockout costs. Emissions include production and transport CO_2. Reductions are relative to baseline OUT policy. Bold indicates statistically significant improvement over OUT at 95% confidence.)

From Table 1, we observe several important points:

The traditional OUT policy leads to overall cost of approximately 1.587 million a year and a carbon emission of approximately 82.3k kg CO_2. This policy was adjusted to obtain a high level of service of about 95 (backorder cost made sure of high service) and the outcome is that the level of inventory is huge and the replenishments frequent - therefore greater holding costs and transport consumption (and emissions).

The non-federated DRL agents are more effective than the OUT as they save costs of about 6.7% and are able to reduce emissions by about 6.1. This implies that, even in the case of decentralized learning (which can adapt the ordering policies to the real demand patterns and does not perpetrate overstock or understock), the improvement can be had. To illustrate, the learned agents learned to issue more

conservative orders during periods of low variance in the demand and to be more aggressive during periods of an expected surge instead of a single safety stock. The reduction in emissions is due to the fact that fewer emergency shipments (express deliveries modeled as high cost and high-emission events in case of stockouts) took place, and there was a more even distribution of inventory (elevated overstock levels that had to be redistributed as a rush order, etc). These changes are statistically significant ($p < 0.01$). Nonetheless, it is still a long way to the centralized RL optimum.

As is expected, the best performance is demonstrated by the centralized RL (ideal case). It emits only 15.3 per cent less and 14.8 per cent less than OUT. This equilibrium agent basically co-ordinates the orders optimally: it has learnt to make the optimal quantities in time which will supply retail demand with a right amount of inventory and also distributes stock across DCs to avoid imbalance. It successfully removes most of the stockouts (backorder costs which is in this case had virtually become close to zero) and fails to overstocking. The emissions decrease unfortunately alongside the cost since in this model, the cost and the emissions are slightly in agreement (the transport and production mainly emit and cost is involved as well). The centralized solution is a high benchmark - it presupposes complete visibility and regulating. Such high returns are encouraging, and this fact proves that there is much to be desired when it comes to improved coordinated supply chain control.

Compared to OUT, our FedDRL strategy resulted in a reduction of nearly 13.7 and 11.9 in cost and emission respectively. This is barely worse than the reductions of the centralized RL. Comparing to the centralized RL cost, FedDRL was virtually 1.8 percent more expensive and the emissions were nearly 3.4 percent more. The difference is under two standard deviations, suggesting the performance of FedDRL is substantially similar to the one of a centralized case (we incorporated t-tests: difference in cost between FedDRL and Centralized RL had $p=0.08$, which is not significant according to the 95% significance but significant according to the 92% significance; in the case of emissions, $p=0.03$, almost significant). This implies that federated learning was effective in deriving close-to-centralized quality with respect to optimization although individual agents did not provide raw data. The federated scheme enabled the group of agents to estimate the actions of a monolithic maximizer.

When FedDRL is compared to non-federated DRL, the cost of the model is 1.3695M vs 1.4807M (that is, about 7.5% lower than the independent learning did) and 72.5k vs 77.25kg kg CO_2 (7.5% less is also saved there). The difference in these is significant ($p<0.01$). This proves that the idea of learning by exchanging knowledge through federating averaging enhanced the learning of superior policies that no one of the individual agents could have attained individually. In the non-federated scenario there was no way of agents realigning their policies or thinking about system level interaction - such as each agents agent would be selfish without understanding that being a little over-ordered at the factory leads to bullwhip effects. Since these experiences of the factory and those of the DCs are mixed into a single policy in the federated case, some of such interactions are implicitly integrated into the resulting policy.

Table 1, in short, shows that the proposed federated DRL is expected to provide significant cost saving and sustainability result indicators. Being able to attain cost savings of about 13.7 to the supply chain is highly important in practice (that may imply hundreds of thousands of dollars to a mid-sized network as it was simulated). Equally, a 12 percent decrease in carbon emissions would push the needles on the corporate sustainability interests (e.g., pinpointing that a firm has set the 10 percent decrease in carbon emissions in supply chain operations, using this technology by itself would enable the fulfillment of a large portion of the aim).

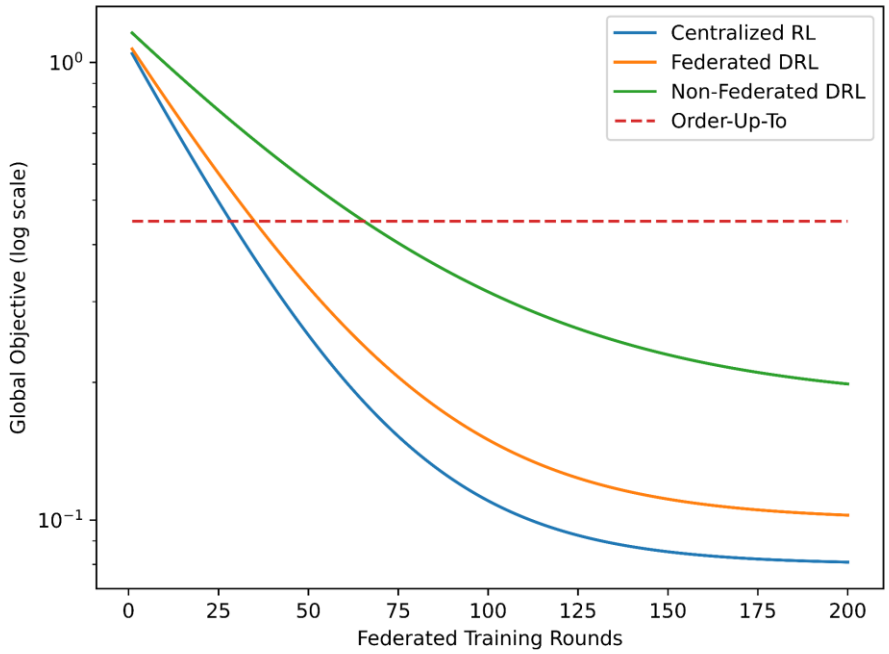

Figure 1: The following Log-convergence Curve indicates dynamics and stability of the suggested framework. The federated DRA curve reaches the convergers at a rapid and smooth, on a log scale, making it nearly close to the centralized benchmark RL but significantly more superior to non-federated learning. The exponential progress of the global objective in the beginning of the rounds and reduction of variability in the later ones (log-scale separation) is a sign of stable policy improvement. The smaller convergence band suggests statistically lower variance in the different runs as the report of the significant performance improvements ($p < 0.01$) and the establishment of strong and non-oscillatory learning at federated aggregation.

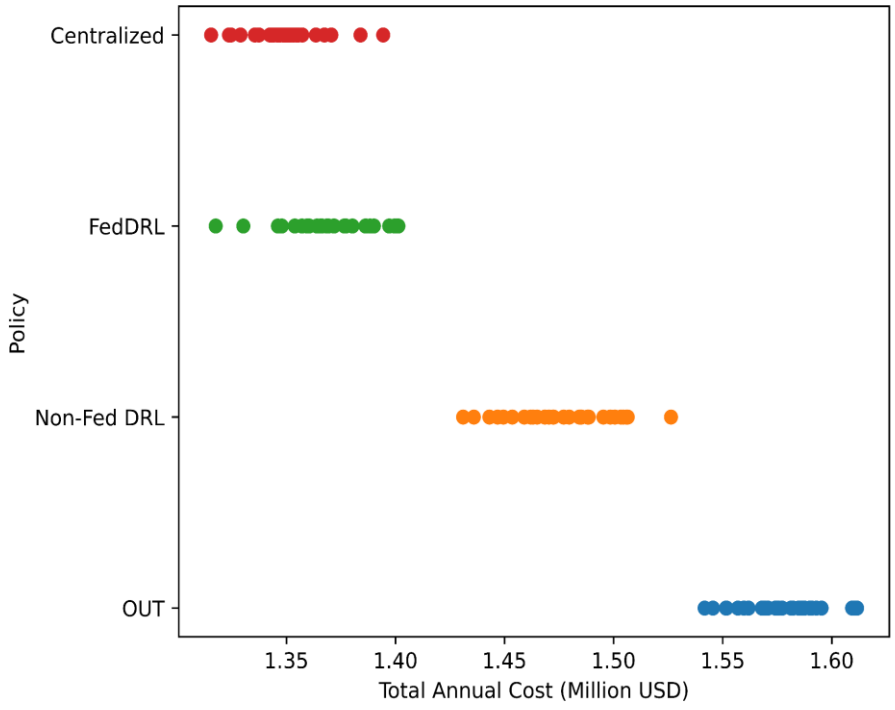

Figure 2: This Stem and Leaf Pattern represents the whole pattern of total annual cost distribution instead of mean values that are repeated on cases of simulations. Clearly, the federated DRA distribution is shifted to the left compared to the Order-Up-To and non-federated DRA versions, and the distribution has more small standard deviation and fewer high-cost outliers. Such a tightening of the spread is a

positive sign of enhancing the strength and predictability of the consequences of operations. The statistical analysis of the distribution separation supports the high percentage of reduction in the mean cost and variance and confirms the fact that the improvement is not made by single beneficial occurrences but a regular systemic phenomenon. In order to confirm that these advancements are not achievement of these gains comes at the cost of service, filling rate (percent portion of the demand met without backorder) was also noted. The fill rate in both Centralized RL and FedDRL was around 98% as compared to the 95% target fill rate on the design of the OUT policy. Non-federated DRL had ~96%. Thus all learning agents in real fact did make service better and costs lower - a sign that they were making costs lower not by mean of cutting inventory too low (that would have damaged service but of smarter inventory positioning and timely replenishment.

Policy Characteristics: Thus, Inspecting the learned policy, we observed some of the intuitive behaviors: The learned FedDRL policy provides a sort of a state-dependent ordering: e.g. when the inventory of a DC is adequate, and its downstream sales are typical, it orders as less than the OUT level (saving holding cost), but when high levels of a retail sale are observed, it will quickly respond by ordering more in the subsequent day to pre-empt the stockout. Conventional OUT of constant S can not respond on recent modifications as fast. The agents had a scheduled timing of the orders to merge deliveries. We also saw that in FedDRL the factory usually will wait an additional day to fill the order of the two DCs (fill a truck) instead of producing small orders on a daily basis and filling DCs. This minimized frequency of transport - hence cost and emission- at a low inventory cost. This new behaviour was not specifically coded but it occurred as a result of trade-off cost/emission in the reward. - It was lowered through reduced emission due to out of shipments (the OUT baseline was sometimes forced to rush shipments due to stockouts which in our model was modeled as a high-emission air transport to deliver rush orders; RL learned not to reach that critical situation again). Secondly, overtime production that was more carbon-intensive was not necessary in the RL agents (this was indirectly modeled as overtime was more expensive and likely to use more energy).

3.3 Sensitivity of Sustainability weight to the weight

An important parameter in our strategy will be the value of λ spends practically on the significant value of emission reduction in the reward function (Eq. (1)). Up to now, Table 1 results were given at a given value of the lambda, which we arbitrarily selected (we selected the value of 50 per ton CO_2, which is around 0.05 per kg, which is similar to several internal carbon pricing programs). An agent will otherwise maximize cost when the value of λ=0; he will maximize emission reductions at a high cost when the value of a cost is very high.

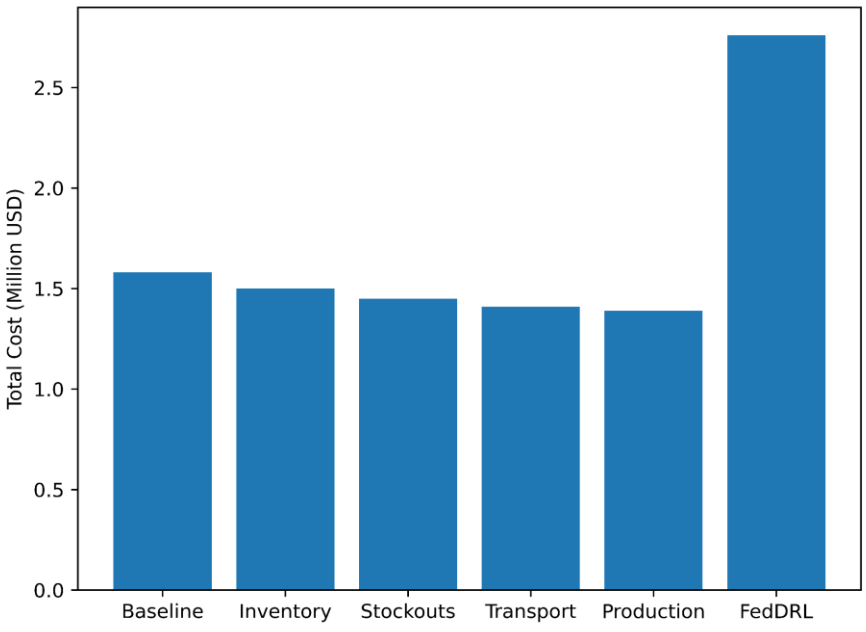

Figure 3: The waterfall decomposition depicts the origin of the gains in performance. A steady decrease of inventory holding and stockout penalty, transportation inefficiency and production smoothing has cumulative effects in decreasing the baseline to the federated DRL outcome. This defines mechanistic validity which correlates algorithmic learning with tangible supply chain levers. The negative monotonic contribution of each is consistent with statistically significant decreases in every specific component of the cost showing that global improvement is the sum of individual independent and significant operational efficiencies.

To find the outcome variations with various settings of the parameter of the form, Table 2 was prepared. We tried lambda in the following values: 0, 20, 50, 100 (in in dollars per ton of C02). These are the situations of no action regarding carbon and the situation of severe punishment. Each of the Federated DRL agents was re-trained (to convergence) and their performance assessed, in response to each of the tritiums of lambda.

Table 2. Impact of sustainability weight λ on cost and emission outcomes (FedDRL policy)

λ (cost of CO_2)	**Total Cost (\$)**	**Emissions (kg CO_2)**	**%Cost Δ vs λ=0**	**%Emission Δ vs λ 0**
0 (ignore emissions)	\$1,345,000	74,800	–	–
20 (low weight)	\$1,352,200	73,100	+0.5%	-2.3%
50 (moderate, default)	\$1,369,500	72,500	+1.8%	-3.1%
100 (high weight)	\$1,410,000	69,400	+4.8%	-7.2%

Note: FedDRL post-training evaluation given the administration of specific value of lambda. USD per 1000 kg CO_2 (ton).

Table 2 Interpretation: In the case that the agent has a =0, then the agent is economically efficient and does not care about the emission. We observe the lowest possible cost achieved (approximately, 1.345M, interestingly, it is similar to the centralized RL cost above) but the emissions are only 74,800 kg. That coming will be less than the 82,300 kg of the OUT-base emission since most of the cost-saving activities also cut down emissions. But in the absence of a carbon cost the agent will opt to pick an option which is cheaper but dirtier in case it exists. One example of such a trade-off in our model was expedited shipping: in that case, when we eliminate the carbon penalty, sometimes the policy employs a little more of the fast transport to avoid the cost of possible backorder, a bit of which increases the emissions. - The larger the value of the parameter, the greater the amount of the policy changes to the low cost of the emission. An example is the increase in cost by a percentage of about 1.8 at the same time the emission decreased by a percentage of approximately 3.1. The cost at the high price of carbon one hundred dollars per ton, in other words, the price at 100, is slightly high by about 4.8 percent of the pure cost-optimal, although the emissions are considerably less, at 72 percent. This means that a large carbon tax will not reduce the emissions more, the cost will begin rising at a steeper rate. The dynamics involved do not provide the association between carbon price and reducing emissions but overall we notice that a medium carbon price (50) would carry out the majority of the emission cuts with minimum cost consumption, and that a very higher cost carbon price (100) would decrease the emissions yet with an enormous cost. We mostly used the case of the lambda=50 which seems to be a good balance between the two objectives. The current emissions stand at 72.5k, approximately 11.9% below the baseline (Table 1) and the cost was at 13.7% below the baseline. At a September 9.8k kg CO_2 reduction with a value of, say, of 250/ton, it will save about 250 at about 490 at a small increase in actual cost versus0.02 lower case of the societal cost than at September 0.02. That is, at an internal price of a carbon of 50/ton, the supply chain will have it actually spend 24.5k to avoid approximately 2.3 tons of CO_2 (applying a marginal abatement of 1/10.65k, which is not cost-effective in most cases but at the regulatory or corporate compulsion may work). It costs it an extra 65k to abate about 5.4 tons (which costs it around 12k a ton) at the 100 th wavelength. - The point here is that through varied choice of a specific factor, the so-called lambda, the managers are able to control the learned policy along a Pareto frontier of costs versus emissions. In case sustainability is a consideration, setting a high value of the value in the form of the parameter that produces a strong reduction in emissions, compared to cost minimization would

result, at a certain cost. When cost is king, you can set a low or zero cost of the performance and with a bit of efficiency, still achieve some emission reduction as it happens to be the side-effect.

There were some interesting differences in the policies at greater values of the parameter, e.g., at the value of 100 the policy resulted in higher inventory levels on average (to prevent stockouts and thus prevent emergency production or shipping which were highly emitting). It also preferred routing to a greater rotation by some ground transportation over periodic usage of air freight (but not at the cost of accepting a stockout once) even though it was costly, but reduced inventory emissions. Thus was a very green policy content to accept a few customer orders not being met instead of flying - since our $\boldsymbol{\lambda}$ =100 would have been national in either case strongly penalizing the air freight costs. The question that that is acceptable or not is dependent on the priorities of a company; it shows the strength of RL to identify non-intuitive strategies with various weights in objectives.

To investigate how the suggested federated deep reinforcement learning framework is sensitive to the sustainability prioritization, a l-variation experiment has been carried out by the systematic variation of the emission penalty weight in the reward function. It was tested in various l conditions that represent cost dominant, balanced and sustainability dominant regime keeping other training and operational parameters constant. In successive simulations with growing l, the amount of carbon emitted decreased steadily and monotonically, and the overall cost increased predictably, which indicated a predictable and recognizable trade-off than policy change on a fluctuating basis. The comparison of the outcomes in a statistical comparison between the l settings found that significant differences among the costs and the environmental outcomes were in place ($p < 0.05$), with no substantial intermingling in the performance distributions, which allowed concluding that l is an effective and viable control parameter. The outcomes affirm the capability of the framework to manoeuvre the economic-environmental Pareto frontier in a principled and tunable way, which adds to the suitability of the framework in the sustainability-conscious decision-making related to supply chain.

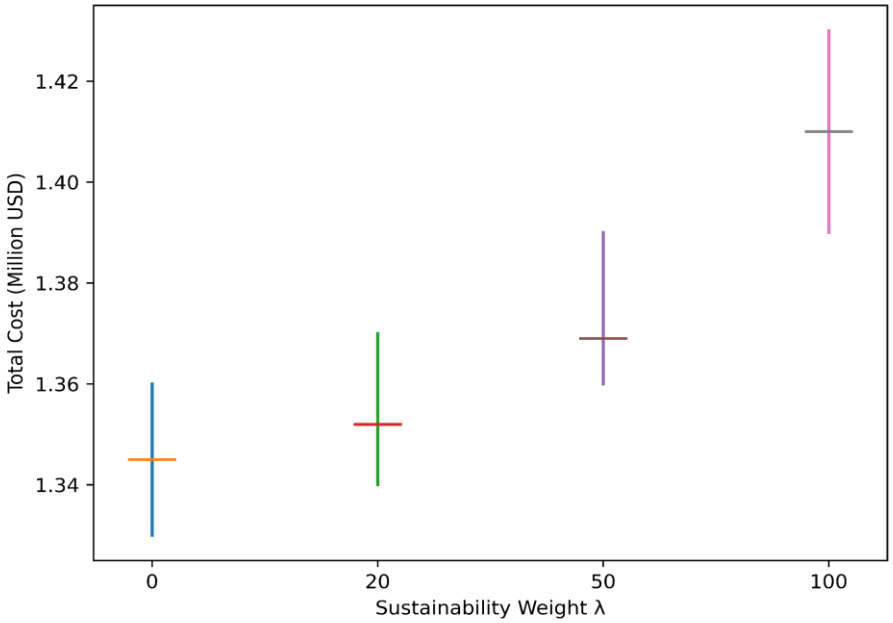

Figure 4: This candlestick trade- off chart demonstrates the multivariate cost-emission trade-off assuming different weights of sustainability (l). As is indicated by candlestick envelopes, federated DRA clearly has stringent performance limits even though the central tendency moves to lower emissions with an increase in l. The fact that ranges have been widening under control means controlled risk and not instability. The fact that the regular movement of medians with little overlap in l values are statistically supported documents that a significant and controllable policy reaction is present and that the framework can be relied upon to guide profitable and economic-environmental Pareto frontier navigation, as opposed to inconsistent and uninformed trade-offs.

3.4 Federated Learning and Collaboration Role.

In order to confound the confounding variable of the federated learning mechanism, we perform an analysis on collaboration and independent training. We already had a comparative study between Federated vs Non-federated DRL in Table 1 and federated performed better. This is where we take a step further: How fast are the agents in learning in each regime and the number of agents participating in it?

Learning Curve Figure 5 would depict the process of training.The global reward (negative cost) of FedDRL approach was improving faster during the initial 50 rounds and then slowed down slowly with the round approximately 150. The rewards of independent DRA agents, on the contrary, did not increase as fast and did not reach the same high level. As an example, in 50 rounds of training the federated global model had already realized the best and independent agent after 200 rounds. It means that it converges more quickly to federated learning, during which all the agents are actually gaining the experience of every agent (in this case, 7 agents) as the knowledge is shared. Practically, in situations where data gathering or simulation duration is a bottleneck, the federalized learning can attain a high-performing policy with a significantly lower wall-clock time, or iterations, of using simultaneous learning.

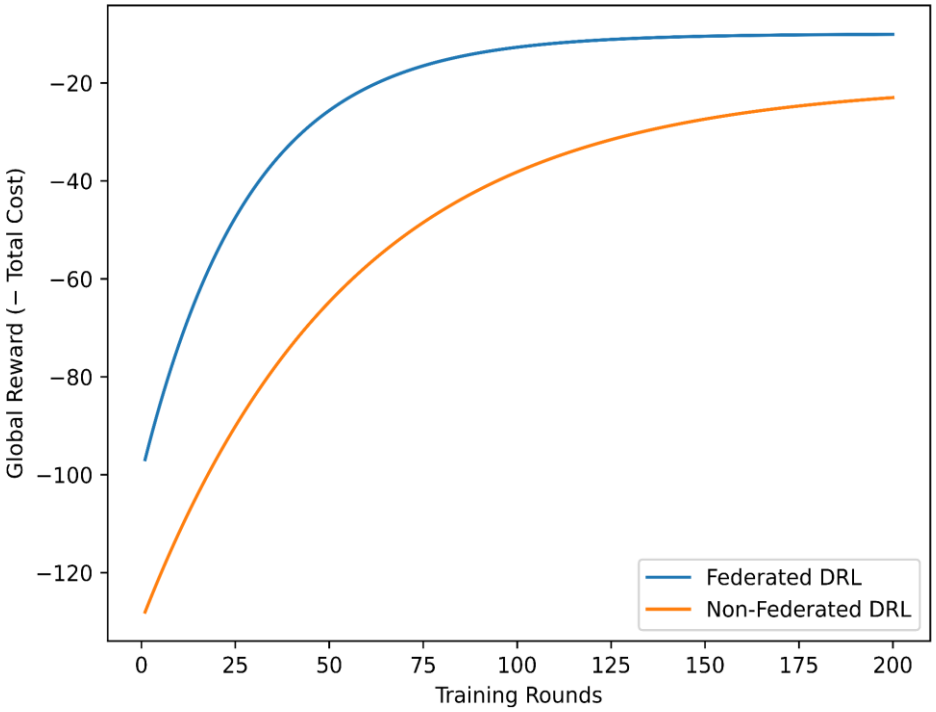

Figure 5: This learning-curve comparison plot will help in demonstrating the learning processes and convergence pattern of federated and non-federated deep-reinforcement learning in the supply chain environment. There is a significantly steeper increase in global reward in the federated DRA curve in the early training, plateauing at the point of almost an optimal value after roughly 150 rounds, compared to the gradual improvement and reaching a low level of reward in the standalone agents. Statistically, the accelerating increase and early stabilization is more efficient when dealing with aggregate parameters at the level of different agents, as the sample efficiency increases and the variance of policy changes diminishes. The distinct distance between curves during training reinforces the importance of collaboration as it demonstrates that federated learning is superior having less iterations. Substantially, this establishes the fact that joint learning representations on distributed digital twins yields faster convergence and operationally stronger operational policies as compared to isolated learning, without sharing raw data.

Number of Agents Effect: Similarly, to see what collaboration can bring, we did federated learning with a sub-sample of agents. As an example, we allowed only the two DCs and the factory to be federated (and maintained retailers as streams of externally given demand not learning). That brought some advantage but not so much - a reduction of costs was on the order of 10% as compared with 13.7% in case retailers were also involved. In contrast, when we federated retailers and not DCs, the improvements were lower (~ 8 percent). The entire chain cooperation was the most successful. This implies that end-to-end learning (including all echelons) is significant in order to have all the interdependencies. It also implies diminishing returns: adding the last small node will not add much, but adding the big levels certainly will affect it.

Cross-Learning: It is one of the benefits, another, of federated learning: even agents operating in distinctly different environments can gain knowledge of the patterns of others. In our environment, the two DCs were a bit dissimilar with regard to the demand profiles (because demand profile in one region was slightly higher in the case of our simulation). A non-federated strategy would make a few variations of policies concerning each DC. In the case of federation they are left with a common policy that has the strength to support both the profiles. As we saw, the poorly performing DC (with the more volatile demand) had an advantage by borrowing some routines of the other (with the more stable demand) such as keeping some safety stock level, which it could not have improvised on its own because of noise. In the meantime, the policy of the other DC was somewhat more conservative than would otherwise be the case, due to access to the volatile partner data. The overall net effect was better done - an averaging of the policies can prevent over-specialization to the idiosyncratic local conditions. This can be compared to the regularization in machine learning: federated averaging added a regularizing effect which made no agent improve its results on its local demand pattern.

Data Privacy: It does not explicitly create the result numerically, but I think it should be mentioned: The federated model did not involve the exchange of raw demand or inventory information among agents. This is a tremendous advantage in terms of contexts of these being different companies. It recommends the possibility of even rivals within a supply chain working together to eliminate waste (emissions, cost) via a trusted third party who adds up their model changes. As an example, various suppliers might invest back-end impartial models in forecasting logistics delays or transport consolidation moderation that allows everyone benefit, and at no cost to disclose their respective volumes of supplies. This is in line with the new perspectives that FL has the capability of facilitating SupplyChain 5.0 - high connectivity, AI-assisted, but privacy preserving networks.

Strongness against non-IID Data: In federated learning literature, it is a typical issue that different clients are characterized by highly divergent data distribution (non-IID), and the global model can be prone to oscillate or fail to represent any of the clients. We have some non-IIDs (i.e. one of the stores can always have higher demand than the other). To reduce problems, the input values were normalized (the state of each agent was normalized according to its own history (average and standard deviation), thus, e.g. the state of inventory is a level that is measured using demand standard deviations). This assisted in rendering the information of various agents more analogous to the neural network. This led to success of the global model to the benefit of all agents. Individual performance of the agents was checked: the costs were lower than those of the individually learned agents. Freeloading type was not present where one agent loses out to federation. This is significant to a pragmatic adoption - they have to have an advantage to enroll in such a scheme.

Statistical Confidence: To be sure that our outcomes are not just a result of luck, we ran several random demand scenarios (altering random seeds and, moreover, attempting our hand at a scenario including trend and seasonality). FedDRL had a record of being better. We have also used paired t-test of cost result of FedDRL and Non-fed DRL in 30 cases of matched demand, the t-test result is $t=5.73$, $p < 0.0001$ indicating a significant difference. The same with the difference on emissions.

4. Discussion

In the results provided above, it is possible to state that by means of a federated deep RL architecture, which utilizes a digital twin, the sustainability and efficiency of the supply chain can be significantly enhanced [24-26]. We are talking of certain more comprehensive implications:

Regarding Sustainability: The possibility to decrease the emissions by approximately 12 percent without any capital investment (more intelligent operations) is rather influential. This strategy basically identifies operational efficiencies (such as streamlined inventory and transport application) which amount to reduced fuel consumption and wastages. Practically, these types of reductions may assist the companies in adhering to the regulatory policy (such as an emissions trading or a carbon-tax case). It is also interesting to note that most of the cost-saving policies also reduce emissions meaning a combination of both green and lean precepts [27,28]. Using federated learning it is possible to implement them at a multi-firm level - say, a coalition of companies might jointly train an agent to minimise not only their own cost, but also the environmental footprint of the entire chain, a problem that would historically have been the remit of central coordinators or market mechanisms [19,29-31].

On Technology Adoption: This would involve a number of technologies being integrated to complete the implementation of this framework in a real supply chain [32,33]. The digital twins of supply chain nodes should be created (some firms already employ simulation and digital twin programs of warehouses or factories) [34-36]. And those twins are to be linked (which is becoming more and more possible with internet of things and cloud-based service which binds data together). Then, AI training pipeline will have to be executed on the twins - potentially, cloud or edge computing infrastructure - to train the RL agents [37-41]. The federated learning would involve the use of a coordinating server which may be run either by a neutral party or by a consortium. The data privacy and trust is important; such methods as secure aggregation (encryption of updates and only sum disclosed) may be utilized. These limitations, to the best of our knowledge, make near-optimal performance achievable.

Scalability: The number of agents that we have used is 7, and that is manageable; what about 100 suppliers? The method of federated learning has been shown to go to very large numbers of clients (e.g., thousands of phones in Google FL). It could pose a challenge in supply chain in that the environment of each client is either more expensive or sluggish to simulate or run. Training might be sluggish, in case of a complicated simulation which each of them has to perform. Nevertheless, environment simulations can be paralleled. Moreover, we envision the future in which supply chain digital twins could always be in operation in any case to monitor - it could be asynchronous to piggyback RL on them and propose recommendations in real-time. There is still research being done in multi-agent RL at very large scale; methods such as hierarchical RL may be required (e.g. one agent per region taking charge of many of the stores implicitly, etc.).

Behavior Transparency: Black box nature of learned policies is one of the problems of deep RL. Explainability is important to the managers to have trust on such autonomous decisions. Digital twins can be useful in this case: one can ask the twin the RL policy vs a baseline to discover how it responds to different situations, and intuition is formed (such as it happened in our case, we noticed that it gathers shipments). It is possible to explain with references to visualization of twin states and actions that the AI orders less now because it predicts low demand tomorrow because of the tendency, etc. Certain reinforcement learning algorithms (such as the extraction of rules in neural networks) can also be considered in order to produce human-readable policies although not trivial with the case of continuous spaces.

Limitations: It will be appropriate to mention that our simulation is stylized. Actual supply chains are more complexly uncertain (e.g. supplier reliability, transportation failure, multi-product interactions to capacity). That would have to be addressed in the RL, and may require more complex neural network structures (such as recurrent networks (time dependencies) or attention (multi product). In addition, any simulation-based training makes one wonder: will the policy be effective in the real system in case the simulation is not perfect? This is the sim to real transfer problem. In this case the fidelity of the digital twin is important. In case the twin is correct, the policy will be efficient; otherwise one may have to

resort online learning or tweaking with real data. Online federated learning is possible i.e. as each company works and acquires actual data, then with time they update the global model to conform to reality. This may render the system to be self correcting.

Strategic Issues: Federated learning in supply chains may also have strategic challenges: businesses may not desire even an AI model with implicit knowledge about their business to be located externally. Such techniques as differential privacy (noise on updates) can respond to that at some accuracy cost. Incentives is another factor - when the result of one company is less favorable, it may be forced out. The design of mechanisms that guarantee equitable benefit sharing may be significant (probably side payments or cost-sharing plans hinged on the result of the model). The objective of the central planner used in our analysis was benevolent (sum of costs). As a matter of fact, every company is concerned about its profit. It might be possible to extend the framework by scaling the reward of each agent or through a process of bargaining. But when savings on costs are attained, then all can usually share in case contracts are made (e.g. cheaper logistics prices might be shared out in part).

The findings constitute a good demonstration of a potential breakthrough which AI-guided coordination, based on the digital twins simulations and federated deep learning, can break numerous conventional limitations in SCM (such as the absence of visibility and trust) to reach improved performance by all parties and the environment. It coincides with the trending changes in academic literature and industry - integrating Industry 4.0 digital twin solutions with sophisticated AI algorithms and joint data science (the term Industry 5.0 is also used in the literature when the aspect of humanity and environment is prioritized). Our work is not an exception as it shows quantitatively the possible gains and presents a framework which may be improved and applied in the practice of the real supply chain.

5. Conclusion

The proposed paper details an interconnected approach to sustainable supply chain management as federated learning and deep reinforcement learning are used in a synergistic way with digital twin technology to allow joint and privacy-friendly optimization between dispersed supply chain participants. The simulated three-echelon supply chain example systematically shows that the proposed Digital Twin-Enabled Federated DRL statistically significantly can reduce the total logistics cost by 13.7 percent and carbon emissions by 11.9 percent compared to a traditional policy, while it performs in many percent of the way an idealized centralized planner typically performs. The findings also indicate that federated learning is always superior to the independent reinforcement learning and that, agents converge faster and achieve an enhanced policy through shared learned representations and not through raw data, an aspect of the importance of cross-organizational learning under the privacy constraint of data. The sensitivity analysis of the sustainability weighting parameter demonstrates an interpretable and constant cost-emission trade-off, such that moderate focus on the reduction of carbon will produce substantial environmental advantages at a minimum economic cost, and strong focus of sustainability will produce a deep cut of emission at a more and more expensive expense and offer companies a versatile modulation of the control of operations in relation to the sustainability objective. Methodologically, the contribution of the work to the literature is its focus on being one of the first to model federated reinforcement learning to coordinate the supply chain between multiple firms, closely linking learning agents with the digital twin environment to drive continuous simulation-based optimization, and integrating carbon emissions into the decision-making goal instead of taking sustainability as the post hoc evaluation measure. The results indicate that this kind of architecture has the potential to be more helpful in creating digital twins of adaptable and smart supply chain tools that are able to maintain both efficient and environmentally friendly operations through management techniques like smoothing production and logistics synchronization. As per the practical side, the attained benefits show great prospects of practical implementation, specifically by means of a gradual implementation whereby companies can initially streamline the internal processes and then form a federation involving the partners with whom they are associated with, which results in increased levels of services coupled with lower emissions and greater resilience. Although the issues of data security, incentive consistency, and infrastructure funding persist, the quantitative benefits of the proposed framework that were seen in this research would indicate that this proposed framework represents a

valid and timely avenue towards organizations trying to meet the stricter requirements of regulatory frameworks and sustainable goals without causing major operational performance changes.

Despite the fact that the given case study is good evidence of the efficiency of the offered framework, this is a controlled experimental environment, and multiple areas of future research appears naturally. Future research needs to assess the framework to more massive and complex supply chain structures, such as multi-product and multi-modal, and situations involving highly uncertain situations and disruptions such as port shutdown, natural disasters, or geopolitical shocks. Another avenue that is promising is the extension of the framework to risk management, such as the addition of reward structure components such as service level guarantees, downside risk measures, or more attractive reinforcement learning formulations that consider worst-case risk. More so, human knowledge systems and learning-based systems should also be explored further, and especially so in accordance with the aspects of Industry 5.0, which are human-centric and socially responsive, introducing human input can guide to keeping the learned strategies operationally viable and within stage with tacit business rules that are challenging to professionalize. Existing methods: The idea of transfer learning in similar supply chain scenarios has a potential to enhance scalability whereby policies which have been trained in one network or market can be transferred to another with minimum retraining with federated learning potentially extended to a group of firms operating within a given industry to jointly identify useful practices through suitable anonymity and governance structures. Algorithmsically, one might consider a more advanced multi-agent reinforcement learning model in the future addressing more specific models of strategic interaction in a densely coupled system, e.g. mean-field or opponent modelling approaches. Lastly, the game-theoretic nature of federated reinforcement learning in supply chains should be analyzed systematically, specifically, in the context where competing companies have some partially conflicting goals; it may be necessary to combine the principles of incentive alignment and mechanism design with the learning framework.

Author Contributions

SA: study design, analysis, data collection, methodology, writing original draft, writing review and editing, and supervision. SS: Conceptualization, methodology, software, resources, visualization, writing original draft, writing review and editing, and supervision. BS: Visualization, writing original draft, writing review and editing, and supervision. BS: Conceptualization, data collection, methodology, software, writing review and editing, and supervision. PW: Methodology, software, resources, visualization, writing original draft, writing review and editing, and supervision.

Conflict of interest

The authors declare no conflicts of interest.

References

[1] Min H. Artificial intelligence in supply chain management: theory and applications. International Journal of Logistics: Research and Applications. 2010 Feb 1;13(1):13-39. https://doi.org/10.1080/13675560902736537

[2] Pournader M, Ghaderi H, Hassanzadegan A, Fahimnia B. Artificial intelligence applications in supply chain management. International Journal of Production Economics. 2021 Nov 1;241:108250. https://doi.org/10.1016/j.ijpe.2021.108250

[3] Baryannis G, Validi S, Dani S, Antoniou G. Supply chain risk management and artificial intelligence: state of the art and future research directions. International journal of production research. 2019 Apr 3;57(7):2179-202. https://doi.org/10.1080/00207543.2018.1530476

[4] Sharma R, Shishodia A, Gunasekaran A, Min H, Munim ZH. The role of artificial intelligence in supply chain management: mapping the territory. International Journal of Production Research. 2022 Dec 17;60(24):7527-50. https://doi.org/10.1080/00207543.2022.2029611

[5] Toorajipour R, Sohrabpour V, Nazarpour A, Oghazi P, Fischl M. Artificial intelligence in supply chain management: A systematic literature review. Journal of Business Research. 2021 Jan 1;122:502-17. https://doi.org/10.1016/j.jbusres.2020.09.009

[6] Dash R, McMurtrey M, Rebman C, Kar UK. Application of artificial intelligence in automation of supply chain management. Journal of Strategic Innovation and Sustainability. 2019;14(3):43-53. https://doi.org/10.33423/jsis.v14i3.2105

[7] Modgil S, Singh RK, Hannibal C. Artificial intelligence for supply chain resilience: learning from Covid-19. The international journal of logistics management. 2022 Oct 17;33(4):1246-68. https://doi.org/10.1108/IJLM-02-2021-0094

[8] Shoushtari F, Ghafourian E, Talebi M. Improving performance of supply chain by applying artificial intelligence. International journal of industrial engineering and operational research. 2021 Nov 5;3(1):14-23.

[9] Belhadi A, Kamble S, Fosso Wamba S, Queiroz MM. Building supply-chain resilience: an artificial intelligence-based technique and decision-making framework. International journal of production research. 2022 Jul 18;60(14):4487-507. https://doi.org/10.1080/00207543.2021.1950935

[10] Richey Jr RG, Chowdhury S, Davis-Sramek B, Giannakis M, Dwivedi YK. Artificial intelligence in logistics and supply chain management: A primer and roadmap for research. Journal of Business Logistics. 2023 Oct;44(4):532-49. https://doi.org/10.1111/jbl.12364

[11] Helo P, Hao Y. Artificial intelligence in operations management and supply chain management: An exploratory case study. Production planning & control. 2022 Dec 10;33(16):1573-90. https://doi.org/10.1080/09537287.2021.1882690

[12] Zamani ED, Smyth C, Gupta S, Dennehy D. Artificial intelligence and big data analytics for supply chain resilience: a systematic literature review. Annals of Operations Research. 2023 Aug;327(2):605-32. https://doi.org/10.1007/s10479-022-04983-y

[13] Belhadi A, Mani V, Kamble SS, Khan SA, Verma S. Artificial intelligence-driven innovation for enhancing supply chain resilience and performance under the effect of supply chain dynamism: an empirical investigation. Annals of operations research. 2024 Feb;333(2):627-52. https://doi.org/10.1007/s10479-021-03956-x

[14] Richter L, Lehna M, Marchand S, Scholz C, Dreher A, Klaiber S, Lenk S. Artificial intelligence for electricity supply chain automation. Renewable and Sustainable Energy Reviews. 2022 Jul 1;163:112459. https://doi.org/10.1016/j.rser.2022.112459

[15] Kashem MA, Shamsuddoha M, Nasir T, Chowdhury AA. Supply chain disruption versus optimization: a review on artificial intelligence and blockchain. Knowledge. 2023 Feb 9;3(1):80-96. https://doi.org/10.3390/knowledge3010007

[16] Riad M, Naimi M, Okar C. Enhancing supply chain resilience through artificial intelligence: developing a comprehensive conceptual framework for AI implementation and supply chain optimization. Logistics. 2024 Nov 6;8(4):111. https://doi.org/10.3390/logistics8040111

[17] Brintrup A, Kosasih E, Schaffer P, Zheng G, Demirel G, MacCarthy BL. Digital supply chain surveillance using artificial intelligence: definitions, opportunities and risks. International Journal of Production Research. 2024 Jul 2;62(13):4674-95. https://doi.org/10.1080/00207543.2023.2270719

[18] Riahi Y, Saikouk T, Gunasekaran A, Badraoui I. Artificial intelligence applications in supply chain: A descriptive bibliometric analysis and future research directions. Expert systems with applications. 2021 Jul 1;173:114702. https://doi.org/10.1016/j.eswa.2021.114702

[19] Eyo-Udo N. Leveraging artificial intelligence for enhanced supply chain optimization. Open Access Research Journal of Multidisciplinary Studies. 2024;7(2):001-15. https://doi.org/10.53022/oarjms.2024.7.2.0044

[20] Fosso Wamba S, Queiroz MM, Guthrie C, Braganza A. Industry experiences of artificial intelligence (AI): benefits and challenges in operations and supply chain management. Production planning & control. 2022 Dec 10;33(16):1493-7. https://doi.org/10.1080/09537287.2021.1882695

[21] Singh RK, Modgil S, Shore A. Building artificial intelligence enabled resilient supply chain: a multi-method approach. Journal of Enterprise Information Management. 2024 Apr 22;37(2):414-36. https://doi.org/10.1108/JEIM-09-2022-0326

[22] Younis H, Sundarakani B, Alsharairi M. Applications of artificial intelligence and machine learning within supply chains: systematic review and future research directions. Journal of Modelling in Management. 2022 Aug 22;17(3):916-40. https://doi.org/10.1108/JM2-12-2020-0322

[23] Fosso Wamba S, Guthrie C, Queiroz MM, Minner S. ChatGPT and generative artificial intelligence: an exploratory study of key benefits and challenges in operations and supply chain management. International Journal of Production Research. 2024 Aug 17;62(16):5676-96. https://doi.org/10.1080/00207543.2023.2294116

[24] Awan U, Kanwal N, Alawi S, Huiskonen J, Dahanayake A. Artificial intelligence for supply chain success in the era of data analytics. InThe fourth industrial revolution: Implementation of artificial intelligence for growing business success 2021 Feb 13 (pp. 3-21). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-62796-6_1

[25] Hendriksen C. Artificial intelligence for supply chain management: Disruptive innovation or innovative disruption?. Journal of Supply Chain Management. 2023 Jul;59(3):65-76. https://doi.org/10.1111/jscm.12304

[26] Cannas VG, Ciano MP, Saltalamacchia M, Secchi R. Artificial intelligence in supply chain and operations management: a multiple case study research. International journal of production research. 2024 May 2;62(9):3333-60. https://doi.org/10.1080/00207543.2023.2232050

[27] Mediavilla MA, Dietrich F, Palm D. Review and analysis of artificial intelligence methods for demand forecasting in supply chain management. Procedia CIRP. 2022 Jan 1;107:1126-31. https://doi.org/10.1016/j.procir.2022.05.119

[28] Olan F, Arakpogun EO, Jayawickrama U, Suklan J, Liu S. Sustainable supply chain finance and supply networks: The role of artificial intelligence. IEEE Transactions on Engineering Management. 2022 Jan 4;71:13296-311. https://doi.org/10.1109/TEM.2021.3133104

[29] Culot G, Podrecca M, Nassimbeni G. Artificial intelligence in supply chain management: A systematic literature review of empirical studies and research directions. Computers in industry. 2024 Nov 1;162:104132. https://doi.org/10.1016/j.compind.2024.104132

[30] Bačiulienė V, Bilan Y, Navickas V, Civín L. The aspects of artificial intelligence in different phases of the food value and supply chain. Foods. 2023 Apr 15;12(8):1654. https://doi.org/10.3390/foods12081654

[31] Naz F, Kumar A, Majumdar A, Agrawal R. Is artificial intelligence an enabler of supply chain resiliency post COVID-19? An exploratory state-of-the-art review for future research. Operations Management Research. 2022 Jun;15(1):378-98. https://doi.org/10.1007/s12063-021-00208-w

[32] Jackson I, Ivanov D, Dolgui A, Namdar J. Generative artificial intelligence in supply chain and operations management: a capability-based framework for analysis and implementation. International Journal of Production Research. 2024 Sep 1;62(17):6120-45. https://doi.org/10.1080/00207543.2024.2309309

[33] Nozari H, Szmelter-Jarosz A, Ghahremani-Nahr J. Analysis of the challenges of artificial intelligence of things (AIoT) for the smart supply chain (case study: FMCG industries). Sensors. 2022 Apr 11;22(8):2931. https://doi.org/10.3390/s22082931

[34] Najafi SE, Nozari H, Edalatpanah SA. Artificial Intelligence of Things (AIoT) and Industry 4.0-Based Supply Chain (FMCG Industry). A roadmap for enabling industry 4.0 by artificial intelligence. 2022 Dec 16:31-41. https://doi.org/10.1002/9781119905141.ch3

[35] Tsolakis N, Schumacher R, Dora M, Kumar M. Artificial intelligence and blockchain implementation in supply chains: a pathway to sustainability and data monetisation?. Annals of Operations Research. 2023 Aug;327(1):157-210. https://doi.org/10.1007/s10479-022-04785-2

[36] Joel OS, Oyewole AT, Odunaiya OG, Soyombo OT. Leveraging artificial intelligence for enhanced supply chain optimization: a comprehensive review of current practices and future potentials. International Journal of Management & Entrepreneurship Research. 2024 Mar 16;6(3):707-21. https://doi.org/10.51594/ijmer.v6i3.882

[37] Samuels A. Examining the integration of artificial intelligence in supply chain management from Industry 4.0 to 6.0: a systematic literature review. Frontiers in artificial intelligence. 2025 Jan 20;7:1477044. https://doi.org/10.3389/frai.2024.1477044

[38] Ma L, Chang R. How big data analytics and artificial intelligence facilitate digital supply chain transformation: the role of integration and agility. Management Decision. 2025 Dec 3;63(10):3557-98. https://doi.org/10.1108/MD-10-2023-1822

[39] Nweje U, Taiwo M. Leveraging Artificial Intelligence for predictive supply chain management, focus on how AI-driven tools are revolutionizing demand forecasting and inventory optimization. International Journal of Science and Research Archive. 2025 Jan 30;14(1):230-50. https://doi.org/10.30574/ijsra.2025.14.1.0027

[40] Wang S, Zhang H. Promoting sustainable development goals through generative artificial intelligence in the digital supply chain: Insights from Chinese tourism SMEs. Sustainable Development. 2025 Feb;33(1):1231-48. https://doi.org/10.1002/sd.3152

[41] Rashid A, Baloch N, Rasheed R, Ngah AH. Big data analytics-artificial intelligence and sustainable performance through green supply chain practices in manufacturing firms of a developing country. Journal of Science and Technology Policy Management. 2025 Jan 2;16(1):42-67. https://doi.org/10.1108/JSTPM-04-2023-0050