Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 23, No. 2, 2026

Adaptive Conservation-Consistent Multi-Agent Policy Optimization for Tail-Risk-Aware Inventory Networks

Andre Kim *, Lana Foster, Hannah Lopez, Jade Ross

Manning College of Information and Computer Sciences, University of Massachusetts Amherst, USA
Corresponding author: Andre Kim (Email: ad.kim29@umass.edu)

Abstract: Capacity-limited inventory networks must reconcile locally requested shipments with executable warehouse flows while controlling losses under supply and demand disruptions. We introduce a decentralized, parameter-shared policy trained through a differentiable, mass-balanced simulator. A nodewise command-to-realization residual, derived from discrete inventory conservation, penalizes infeasible requested flows. An online-adjusted risk weight couples mean episode cost with the upper-tail conditional value-at-risk (CVaR). Ten independent seeds compare seven policies across four synthetic disturbance regimes (400 held-out trajectories per method, seed and regime). Under combined stress, mean cost is 4906.8 versus 4979.7 for risk-neutral policy optimization, while CVaR90 is 8464.8 versus 8580.9. Against a fixed-risk comparator with a closely matched final risk weight, the adaptive formulation further reduces stress CVaR90 by 39.4 cost units. Requested-but-unshipped flow drops from 15.80 to 2.27 units/day. Realized transitions conserve inventory under every policy: the residual regulates planning consistency, not physical conservation itself. Results reflect a reproducible synthetic benchmark rather than field validation.

Keywords: Multi-agent policy optimization; Supply chain inventory; Conservation constraints; CVaR; Adaptive risk control; Differentiable simulation.

1. Introduction

Inventory control is fundamentally a network problem: replenishment requests at one outlet can displace stock needed elsewhere, and an otherwise sensible local policy may become infeasible when a common warehouse is capacity limited. The classical multi-echelon formulation of Clark and Scarf [1] shows that upstream inventory availability couples downstream ordering decisions. Multi-agent reinforcement learning has consequently been studied for the beer game [2], learning-assisted newsvendor control [3], and difficult multi-echelon inventory settings [4]. In these applications, the relevant decision is often an executable physical shipment, not merely the request issued by a learned agent.

Modern inventory-learning systems handle lead-time constraints and multiple products [5], and libraries such as OR-Gym [6] and MABIM [7] provide controlled simulation environments. Hierarchical multi-node control is also a practical means of managing many simultaneous decisions [8]. These developments motivate a narrower question that is rarely isolated experimentally: should a decentralized policy be trained against its own requested material movement, or against the movement that remains feasible after shared supply is allocated?

The distinction matters most during disruption. If retailers concurrently request more than the warehouse can deliver, the environment must scale down their orders. A policy that repeatedly over-requests can still record apparently reasonable returns when the simulator silently repairs the action. Such a policy is difficult to interpret and may create operational planning churn. Conversely, aggressive penalties on requests can sacrifice fill rate during normal operation. Tail-sensitive objectives introduce another tension: optimizing average cost is insufficient when intermittent inbound failures coincide with correlated spikes in customer demand.

We study a conservation-consistent alternative in a transparent warehouse–three-retailer benchmark. The contributions are: (i) an inventory-balance residual between proposed and capacity-projected material movement; (ii) a scenario-CVaR objective with an online risk multiplier trained jointly with decentralized retailer actions; and (iii) ten-seed tests across nominal, demand-only, supply-only and combined stresses, including a tuned base-stock rule, a fixed-risk baseline matched to the adaptive final weight, and mechanism-removal ablations. Complete scenario arrays, checkpoints and paired-seed results are retained for reproducibility.

2. Related work

2.1. Multi-agent inventory decisions

Centralized training with decentralized execution is prominent in multi-agent actor–critic methods [9] and counterfactual policy gradients [10]. Value factorization methods such as QMIX [11] address joint action selection, while recent systematic studies of cooperative PPO [12] underscore the importance of implementation choices and comparable baselines. Our method instead differentiates through a known, short-horizon simulator. Its actors use local retailer observations and a publicly broadcast warehouse-availability signal; the shared simulator is used only for training. This choice retains decentralized decisions without adding a learned joint critic.

2.2. Physics-informed learning and return uncertainty

Physics-informed neural networks incorporate known equations into learning objectives [13]. Their broader role as structural inductive biases is well established [14], although scaling and balancing physics terms remain difficult [15], and different loss components can train at unequal rates [16]. In an inventory system, physical constraints are not differential equations approximated by a neural solver: discrete

conservation of items can be written exactly. We therefore apply a conservation residual directly to the commanded inventory update, and measure the discrepancy from the executable transition, rather than imposing a redundant constraint on a simulator that already conserves material.

Distributional reinforcement learning studies the full distribution of return [17], with quantile regression supplying a practical representation [18]. For cost minimization, CVaR is particularly suitable because it averages losses in the upper tail rather than using a single high quantile [19]. Risk-constrained policy gradients and their multiplier updates have been studied by Chow et al. [20], while coherent-risk policy gradients provide a wider framework [21]. Liu, Wang, and Lin [24] unify these strands for supply chains: their coordinated physics-informed multi-agent framework embeds conservation constraints in learning, optimizes CVaR over learned return distributions, and coordinates agents through Lagrangian multiplier updates. We build on its conservation residual (Section III-B) and its risk-aware objective, but compute CVaR directly from sampled episode costs of a differentiable simulator instead of training a return critic.

2.3. Coordinated constraints

Lagrangian dual decomposition and the alternating direction method of multipliers give principled means to coordinate local objectives [22], while constrained policy optimization considers limits during policy search [23]. Here an adaptive scalar risk multiplier serves as a lightweight controller; it is neither an ADMM consensus solver nor a certificate that a potentially infeasible CVaR target is met.

3. Inventory dynamics and physical consistency

3.1. Decentralized actions and stochastic disturbances

Let $i \in \{1,2,3\}$ index retailers and let t denote a daily period. The warehouse has on-hand inventory W_t and each retailer has $I_{i,t}$. Exogenous warehouse arrivals are P_t, with usable availability $A_t = \min\{120, W_t + P_t\}$ and overflow $Z_t = \max\{0, W_t + P_t - 120\}$. Demand $D_{i,t}$ is realized after dispatch. Each actor receives local inventory, its demand scale, previous actual shipment, seasonal phase, retailer identity, and the current warehouse availability broadcast. The policy generates the requested shipment $0 \le Q_{i,t} \le C_i$, with transport capacities $C = (28,30,32)$ units per day. All quantities are measured in units, not abstract normalized rewards.

The warehouse may receive simultaneous requests exceeding available stock. Define the common rationing factor $\rho_t = \min\{1, A_t / (\Sigma_j Q_{j,t} + 10^{-6})\}$. Executed shipments $F_{i,t} = \rho_t Q_{i,t}$ are nonnegative, do not exceed requests or capacities, and jointly satisfy $\Sigma_i F_{i,t} \le A_t$. Immediate customer sales $S_{i,t} = \min\{D_{i,t}, I_{i,t} + F_{i,t}\}$ induce the next inventories:

$$I_{i,t+1} = I_{i,t} + F_{i,t} - S_{i,t}, \quad W_{t+1} = A_t - \Sigma_i F_{i,t}. \tag{1}$$

Nonnegative initial stocks are preserved. Moreover $W_{t+1} + \Sigma_i I_{i,t+1} = A_t + \Sigma_i I_{i,t} - \Sigma_i S_{i,t}$, exactly separating exogenous warehouse arrivals, shipped inventory, and consumption. The capacity projection is piecewise differentiable and requires no numerical solver.



Fig. 1 Capacity-constrained warehouse-to-retailer flow and adaptive CVaR feedback in the training objective.

3.2. Commanded-flow conservation residual

The commanded-flow penalty adopts the discrete nodewise conservation residual of [24]—observed inventory change minus inflow, plus outflow and demand—together with its differentiable capacity-limited flow rule, which limits flow by source availability and link capacity. Our extension changes the physical comparison: the actor's raw requests Q represent a provisional plan, while F denotes the material that can actually move after shared-capacity rationing. This distinction exposes an otherwise hidden planning inconsistency.

Define provisional one-step predictions $\tilde{I}_{i,t+1} = I_{i,t} + Q_{i,t} - S_{i,t}$ and $\tilde{W}_{t+1} = A_t - \Sigma_i Q_{i,t}$. Using the actual states of (1), the per-period nodewise squared residual becomes

$$R_t = (W_{t+1} - \tilde{W}_{t+1})^2 + \Sigma_i (I_{i,t+1} - \tilde{I}_{i,t+1})^2 \tag{2}$$

$$= (\Sigma_i (Q_{i,t} - F_{i,t}))^2 + \Sigma_i (Q_{i,t} - F_{i,t})^2. \tag{3}$$

Equation (2) thus carries this conservation loss from realized to commanded flows. Its derivative passes through the requested action and the dispatch projection; it is therefore positive when planned and feasible flows disagree. It is zero for feasible requests. Importantly, if one instead calculated the residual using only realized F and S in (1), the error would be identically zero under every policy. The experiment reports this realized-balance diagnostic separately and never attributes changes in exact simulator conservation to learning.

3.3. C. System cost and physical properties

The unit costs distinguish holding inventory from stockouts. Retailer-specific shortage costs are $b = (10,12,14)$; on-hand retailer holding costs 0.46 per unit per day; warehouse holding 0.21; dispatched transport 0.36; and discarded warehouse overflow 0.75. The daily cost is

$$L_t = \Sigma_i [b_i (D_{i,t} - S_{i,t}) + 0.46 I_{i,t+1} + 0.36 F_{i,t}] + 0.21 W_{t+1} + 0.75 Z_t. \tag{4}$$

Each episode cost $C(\theta,\xi)$ sums L_t over 28 days under scenario ξ. Conservation and positivity follow immediately from the feasibility projection and sale definition. For any nonnegative W_t, $I_{i,t}$, P_t and $Q_{i,t}$, we have $A_t \ge \Sigma F_{i,t}$ and $S_{i,t} \le I_{i,t} + F_{i,t}$, so the two next-state inventories are nonnegative and (1) holds exactly. This is a property of the environment's dynamics; no separate claim of policy convergence follows from it.

4. Adaptive risk-aware multi-agent optimization

4.1. Joint simulation, decentralized policy

A shared feed-forward actor network with two tanh hidden layers of 48 units maps each retailer's nine-component local feature vector to a sigmoid action, scaled by its own edge capacity. Retailer identity is one-hot encoded, allowing heterogeneous actions while sharing statistical strength. Training simultaneously unrolls all three actors through the same physically coupled supply graph. Gradients pass

through time and the rationing factor using automatic differentiation, which is a model-based form of deterministic multi-agent policy optimization. No learned transition model, categorical critic, replay buffer or off-policy Bellman update is involved.

In each epoch, a fresh batch of B = 64 independent stochastic 28-day episodes is drawn. The empirical mean episode cost is $M = B^{-1}\Sigma^{b}C^{b}$. Let k = ceil(0.1B). Define an upper-tail CVaR estimate as the arithmetic mean of the k largest episode costs. This empirical choice measures exactly the worst 10% of sampled cost outcomes; unlike a 95th-percentile point estimate it averages tail severity. Within a batch the top-k operation is differentiable almost everywhere.

$$\mathrm{CVaR}_{0.9}(C) \approx k^{-1} \Sigma_{\beta} \in \mathrm{Top}_{k}(C)\, C_{\beta}. \quad (5)$$

4.2. Adaptive risk pressure and physics penalty

The loss combines risk sensitivity with the mean physical mismatch $\bar{R} = (BT)^{-1}\Sigma_{\beta,t}R_{\beta,t}$, using fixed normalization for numerical stability:

$$J(\theta) = [M + \lambda\, \mathrm{CVaR}_{0.9}(C)] / 1000 + 3.4\bar{R} / 1200. \quad (6)$$

The multiplier λ starts at 0.25 for the adaptive variants and changes once per gradient step according to the batch CVaR relative to a preselected reference cost $c\star = 4200$:

$$\lambda \leftarrow \mathrm{clip}[\lambda + 0.08(\mathrm{CVaR}_{0.9} - c\star)/c\star, 0, 1.8]. \quad (7)$$

This nonnegative feedback increases the attention paid to high-cost episodes when measured risk is above the reference and relaxes it when risk is below. The reference is a design setpoint, not a hard upper bound. It was fixed prior to held-out evaluation; as demonstrated in the results, the stressed system remains above it. We therefore interpret λ as a risk-weight adaptation rather than a mathematical certificate of risk feasibility.

The actor is optimized by Adam [25] at learning rate 0.008 for 45 epochs using clipping of the gradient norm at 5. Because the environment is differentiable, direct backpropagation implements the basic policy-gradient principle of optimizing expected outcomes with respect to policy parameters [26]. The method has lower training complexity than a full distributional multi-agent critic, but it requires a known differentiable transition mechanism; discontinuous lead-time events or unknown dynamics would require approximation or alternate gradient estimators.

4.3. Ablations and controlled variants

Seven methods share identical environment transitions and evaluation episodes. A base-stock rule selects its retailer target from {20,26,32,38,44,50}, using 200 independent validation trajectories per seed. Six learned variants have the same 48–48 actor and 45 training epochs: risk-neutral (no risk or physics terms); fixed-risk at λ = 0.55 without physics; a second fixed-risk control at λ = 0.90 without physics; physics only; adaptive risk without physics; and the proposed adaptive risk plus physics. The λ = 0.90 comparator approximates the mean final adaptive multiplier observed in pilot training logs and tests whether gain stems merely from raising a fixed risk weight. All learned variants use model-based policy search; model-free actor–critic methods are discussed in Section VII.

5. Experimental protocol

5.1. Stochastic scenario generation

All trajectories are synthetic, not derived from industrial or public-company transaction logs. Base retailer daily demand follows independent Poisson distributions with means (17,19,21), multiplied by a 14-day sinusoid of amplitude 0.12. A shared gamma increment occurs on 10% of nominal days and affects all retailers, introducing correlated spikes. Warehouse arrivals are $N(60,12^2)$, truncated at zero; on 9% of days receipts are reduced to 25% of their original value. Three held-out shifts isolate the causes of adversity: demand-only increases shared-shock probability to 18%; supply-only raises receipt-disruption probability to 17% and cuts disrupted receipts to 12%; combined stress changes both sets of parameters. The shifts were specified explicitly and evaluated for all seven methods, with neither policy retraining nor calibration to the shifted test episodes.

Initial warehouse stock is 43 units and initial retailer stock is 23 units each. Capacity never exceeds 120 units in the warehouse; fixed retailer transport capacities are (28,30,32). Shipments occur within the time step and unmet demand is lost rather than backordered. The absence of pipeline lead times and arbitrary number of retailers is a deliberate tractability limitation.

5.2. Separation of training, validation, and test

Ten independent initializations use seeds 0–9. Training scenarios use NumPy seed 1,000,000 × s + e for seed s and epoch e; base-stock validation uses 800,000 + s. Held-out scenario seeds are 910,000 + s for nominal, 920,000 + s for combined stress, 930,000 + s for supply-only and 940,000 + s for demand-only. Each method is evaluated on the same 400 trajectories within a seed and regime. The first five actor checkpoints were preserved from the initial five-seed run under unchanged training parameters; the remaining five were trained independently, and the matched fixed-risk models were trained for all ten seeds. This design avoids test-driven retraining while extending independent policy initializations and distribution shifts.

5.3. Outcomes and statistical estimation

Metrics are mean 28-day episode cost, empirical worst-decile CVaR90 cost, fill rate (one minus aggregate lost-sales fraction), average request–shipment gap $\Delta = E[\Sigma_i(Q_i - F_i)]/28$, and the nodewise squared plan-residual in units squared. Executed material balance is verified independently using realized shipments and sales. Reported mean ± SD values aggregate ten independent training seeds, not 4000 pseudo-independent training replications. We compute paired 95% Student-t intervals over seed-level differences and, as a complementary small-sample check, enumerate all 2^{10} sign flips to report two-sided randomization p-values. Multiple contrasts are exploratory. The emphasis on seed variability is consistent with reproducibility research [27].

6. Results

6.1. Nominal and stressed performance

Table I reports held-out results directly from the archived per-seed file. Under nominal conditions, adaptive control gives mean cost 2650.7 ± 56.8, compared with 2643.2 ± 62.2 for the risk-neutral policy. Their paired mean-cost difference is 7.4 [-4.5, 19.3]; zero remains plausible. Conversely, nominal CVaR90 falls from 4988.0 to 4904.0, a paired difference of -84.0 [-103.4, -64.5]. Thus tail improvements need not imply lower nominal mean cost.

Under combined demand-and-supply stress, the proposed policy yields mean cost 4906.8 ± 73.0 and CVaR90 8464.8 ±

171.8. The risk-neutral values are 4979.7 and 8580.9, differences of 1.47% and 1.35%, respectively. The paired mean-cost difference is -73.0 [-85.9, -60.1]; the CVaR90 difference is -116.1 [-135.7, -96.5]. For both, the exact two-sided seed-level sign-flip test gives $p = 0.0020$ and $p = 0.0020$, respectively. These are simulation effects under the specified test generator.

Table I. Nominal and combined-stress cost, mean ± SD across 10 training seeds (400 test episodes per seed and regime).

Method	Nom. mean	Nom. CVaR	Stress mean	Stress CVaR
Base stock	2726±64	5131±202	5110±79	8720±175
Risk-neutral	2643±62	4988±207	4980±82	8581±192
Fixed 0.55	2645±63	4941±204	4940±77	8520±191
Fixed 0.90	2654±62	4932±200	4930±67	8504±179
No physics	2652±67	4926±205	4927±73	8498±187
Physics only	2645±59	4942±203	4940±74	8524±178
Adaptive	2651±57	4904±202	4907±73	8465±172

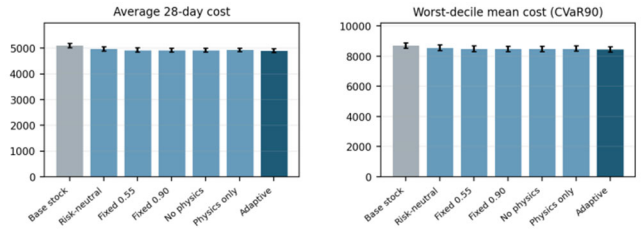


Fig. 2 Mean and worst-decile cost under combined stress; error bars are SD across 10 independent training seeds.

6.2. Physical consistency of intended shipments

The physics penalty acts on commanded movement, distinct from operational cost. Under combined stress the request–shipment gap is 2.27 units/day for the complete policy, 15.80 for risk-neutral learning, and 13.76 when the residual term is removed while preserving risk adaptation. The relative reductions are 85.7% and 83.5%. The mean nodewise squared prediction residual likewise falls from 612.3 to 25.9 units². These gains characterize requested-flow executability, not changes in realized conservation, which holds for all methods by construction.

Across every stored evaluation episode and method the largest absolute mass-balance diagnostic is below 1.2×10^{-4} units, consistent with float32 summation error. Consequently, it would be misleading to claim fewer simulated conservation-law violations for the proposed policy. Equation (3) makes the planning benefit explicit: the warehouse term penalizes total shared over-request quadratically, and nodewise terms penalize individual unexecuted requests. The very low request gap of the physics-only model additionally indicates that pure feasibility need not coincide with optimal tail-cost control.

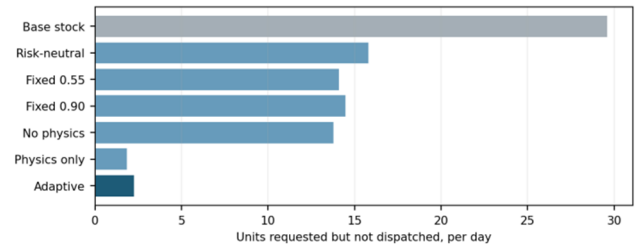


Fig. 3 Stressed requested-but-unshipped flow, lower is better; all values computed from the simulation.

6.3. Adaptive risk-weight behavior and ablation

The adaptive risk multiplier rises from 0.25 to 0.93 ± 0.05 at epoch 44, as shown in Fig. 4. To distinguish that update from simply applying a stronger static risk penalty, the matched fixed-risk control uses $\lambda = 0.90$, close to the learned final mean. Under combined stress it yields mean cost 4930.3 and CVaR90 8504.3, while the adaptive complete method yields 4906.8 and 8464.8. The paired gaps are -23.6 [-34.9, -12.2] for mean cost and -39.4 [-61.9, -17.0] for CVaR90, with exact sign-flip $p = 0.0039$ for the latter. Thus the matched-weight ablation supports an effect beyond the final multiplier magnitude in this simulation, although the adaptive schedule and physics term also differ between models.

An especially informative ablation retains the identical adaptive dual update while setting the physical penalty to zero. This no-physics model has combined-stress cost 4926.9 and CVaR90 8498.0; adding the conservation residual adapted from [24] decreases them by 20.1 and 33.1 units. Seed-paired intervals are -20.1 [-29.3, -10.9] and -33.1 [-50.9, -15.3]. Physics-only learning achieves a still smaller request gap but weaker stress cost than the complete formulation. Together, these ablations separate risk pressure from command feasibility, while remaining specific to the selected simulator and hyperparameters.

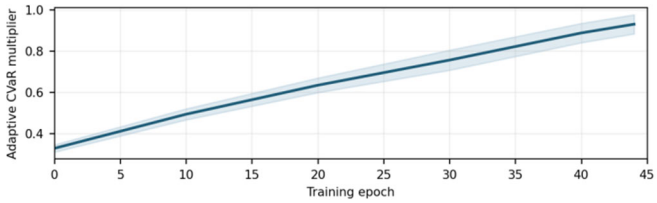


Fig. 4 Empirical adaptive risk multiplier during training: across-seed mean and ±1 SD, sampled every 10 epochs.

Table II. Combined-stress service and command-feasibility metrics, mean over 10 training seeds.

Method	Fill %	Request gap/day	Residual squared
Base stock	81.30	29.60	1993.1
Risk-neutral	81.38	15.80	780.0
Fixed 0.55	81.34	14.10	642.7
Fixed 0.90	81.36	14.47	654.1
No physics	81.34	13.76	612.3
Physics only	81.40	1.85	21.6
Adaptive	81.37	2.27	25.9

6.4. Tail quantile sensitivity

To check sensitivity to the tail fraction without retraining, we reran every saved policy on its original held-out scenario arrays and computed the worst-20%, worst-10% and worst-5% mean episode costs. Under combined stress, the adaptive model obtains CVaR80 = 7633.5, CVaR90 = 8464.8, and CVaR95 = 9204.4; the respective risk-neutral estimates are 7744.5, 8580.9, and 9324.6. The matched fixed-risk method gives 7669.6, 8504.3, and 9248.3. The same direction holds at all three tail cuts; these are recalculations of the held-out distributions, not separate policies tuned at each quantile.

These figures are empirical scenario averages, not calibrated tail-probability bounds. Each 5% estimate uses the 20 most expensive trajectories among 400 per seed; rare events beyond the modeled gamma and receipt-disruption distributions could substantially change them. Compared with reporting only average return, the three thresholds clarify

what the CVaR component is actually optimizing and distinguish sensitivity to tail severity from general changes in mean cost.

6.5. Disentangling held-out demand and supply shifts

Combined stress can obscure which stochastic mechanism produces a policy advantage. Fig. 5 disaggregates the shifts, without retraining any policy: demand-only raises the correlated-shock probability to 18% while keeping nominal receipts; supply-only raises inbound disruption frequency to 17% and lowers affected receipts to 12%, retaining nominal demand surges. The complete method has cost/CVaR90 3620.7/6585.8 under demand-only stress and 3735.7/6689.5 under supply-only stress. Both are below the matched fixed-risk controls (3634.8/6622.2 and 3757.0/6726.1, respectively). These results distinguish robustness to the two specified stress mechanisms, not arbitrary unmodeled distribution shift.

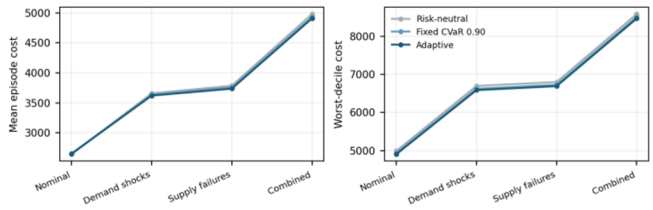


Fig. 5 Held-out performance under separately intensified demand surges and inbound supply failures (10 training seeds).

6.6. Computational cost and balance diagnostics

Each learned model uses 45 batches of 64 stochastic episodes, or 80,640 simulated decision-days per training seed; the base-stock rule is validation tuned separately. Experiment scripts launch one independent process per training seed and archive process logs and elapsed times. Host load varied between runs, so no algorithmic speedup claim is based on the recorded seconds. The largest test workload comprises seven evaluated policies across ten seeds and four regimes, each using 400 independently generated 28-day trajectories. The experiment remains CPU feasible because it optimizes actors directly through a known differentiable simulator instead of fitting a distributional critic.

The cost of each policy iteration scales with the number of retailers N, the trajectory length T, the episode batch size B and the cost of evaluating the shared actor. For fixed-width actors, the forward state simulation is O(BTN); reverse-mode differentiation requires retaining the state and action graph across T periods. The system-wide flow allocation uses only one reduction over retailer requests and one broadcast scale per period; it does not require a quadratic graph-attention operation. This computational structure supports modest network expansion, but it is not evidence that our fixed three-retailer experiment is scalable to dozens of echelons.

The exact-balance diagnostic is $E = \Sigma_t(|W_{t+1} - A_t + \Sigma_i F_{i,t}| + \Sigma_i|I_{i,t+1} - I_{i,t} - F_{i,t} + S_{i,t}|)$. Its maximum over all held-out runs is below 1.2×10^{-4} units. The same invariant holds for the policies with the largest planning gaps, confirming that simulated material conservation and the quality of the order plan are distinct concepts. This separation is essential when evaluating a physics-aware controller in an environment where the physical transition kernel already enforces feasibility.

7. Interpretation and limitations

Three observations shape interpretation. First, average cost and CVaR improvements are small in relative terms, even where paired differences are consistent across seeds. Second, a conservation-derived residual is useful as an action regularizer because it distinguishes what agents request from what can execute; every rollout conserves mass by construction. Third, adaptive CVaR weighting exhibits a measured advantage over both a weaker fixed weight and a fixed weight selected near the trained final multiplier. The feedback setpoint itself is not met, so these results do not constitute a risk-feasibility guarantee.

The experiments use a three-retailer, zero-lead-time network with proportional stock rationing, lost sales, and fully observed warehouse availability. Transport execution is immediate; there are no delayed pipelines, supplier selection, perishability ages, or decentralized negotiation protocols. These assumptions are important because differentiating through (1) would become more demanding with discrete orders, delayed transport and partially observed upstream capacity. Direct comparisons with end-to-end MADDPG, QMIX, D4PG, COMA or PPO, and with the complete physics-informed distributional framework of [24], are left to future work; carrying the commanded-flow residual into that framework's learned critics and dual coordination is a natural next step.

Poisson, gamma and Gaussian shocks are controlled scientific assumptions, not estimates from a real business. All three stressed regimes retain the same warehouse topology, immediate transportation and fully observed upstream availability; other distributions, longer disruptions or nonzero lead times may reverse the ranking. Although ten seeds and paired sign-flip tests improve reproducibility, exploratory multiple comparisons should not be mistaken for universal guarantees. Future work should test multi-warehouse topology, lead times, private supply signals, real inventory data and stronger model-free actor–critic comparators before making industrial deployment claims.

8. Conclusion

A command-to-realization conservation penalty and adaptive upper-tail risk weight were implemented in a jointly differentiable, multi-agent inventory controller. Ten-seed comparisons against matched-weight fixed risk, risk-neutral learning, a tuned base-stock rule and mechanism-removal variants show lower command mismatch and modest cost/tail benefits over four specified synthetic regimes. The primary practical distinction is between guaranteed simulator material balance and learned plans that are executable without corrective rationing. The resulting transparent benchmark and saved experimental artifacts support further tests with unknown dynamics and real supply-chain data.

References

[1] Clark, A. J., & Scarf, H. (1960). Optimal policies for a multi-echelon inventory problem. Management Science, 6(4), 475-490. https://doi.org/10.1287/mnsc.6.4.475

[2] Chaharsooghi, S. K., Heydari, J., & Zegordi, S. H. (2008). A reinforcement learning model for supply chain ordering management: An application to the beer game. Decision Support Systems, 45(4), 949-959. https://doi.org/10.1016/j.dss.2008.03.007

[3] Oroojlooyjadid, A., Snyder, L. V., & Takáč, M. (2020). Applying deep learning to the newsvendor problem. IISE Transactions, 52(4), 444-463. https://doi.org/10.1080/24725854.2019.1632502

[4] Gijsbrechts, J., Boute, R. N., Van Mieghem, J. A., & Zhang, D. J. (2022). Can deep reinforcement learning improve inventory management? Performance on lost sales, dual-sourcing, and multi-echelon problems. Manufacturing & Service Operations Management, 24(3), 1349-1368. https://doi.org/10.1287/msom.2021.1064

[5] Meisheri, H., Sultana, N. N., Baranwal, M., Baniwal, V., Nath, S., Verma, S., Ravindran, B., & Khadilkar, H. (2022). Scalable multi-product inventory control with lead time constraints using reinforcement learning. Neural Computing and Applications, 34(3), 1735-1757. https://doi.org/10.1007/s00521-021-06129-w

[6] Hubbs, C. D., Perez, H. D., Sarwar, O., Sahinidis, N. V., Grossmann, I. E., & Wassick, J. M. (2020). OR-Gym: A reinforcement learning library for operations research problems. arXiv, arXiv:2008.06319. https://arxiv.org/abs/2008.06319

[7] Yang, X., Liu, Z., Jiang, W., Zhang, C., Zhao, L., Song, L., & Bian, J. (2023). A versatile multi-agent reinforcement learning benchmark for inventory management. arXiv, arXiv:2306.07542. https://arxiv.org/abs/2306.07542

[8] Sultana, N. N., Meisheri, H., Baniwal, V., Nath, S., Ravindran, B., & Khadilkar, H. (2020). Reinforcement learning for multi-product multi-node inventory management in supply chains. arXiv, arXiv:2006.04037. https://arxiv.org/abs/2006.04037

[9] Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in Neural Information Processing Systems, 30, 6379-6390.

[10] Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2974-2982. https://doi.org/10.1609/aaai.v32i1.11794

[11] Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018). QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80 (pp. 4295-4304). PMLR.

[12] Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The surprising effectiveness of PPO in cooperative multi-agent games. Advances in Neural Information Processing Systems, 35, 24611-24624. https://doi.org/10.52202/068431-1787

[13] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707. https://doi.org/10.1016/j.jcp.2018.10.045

[14] Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3, 422-440. https://doi.org/10.1038/s42254-021-00314-5

[15] Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., & Piccialli, F. (2022). Scientific machine learning through physics-informed neural networks: Where we are and what's next. Journal of Scientific Computing, 92, Article 88. https://doi.org/10.1007/s10915-022-01939-z

[16] Wang, S., Yu, X., & Perdikaris, P. (2022). When and why PINNs fail to train: A neural tangent kernel perspective. Journal of Computational Physics, 449, Article 110768. https://doi.org/10.1016/j.jcp.2021.110768

[17] Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70 (pp. 449-458). PMLR.

[18] Dabney, W., Rowland, M., Bellemare, M. G., & Munos, R. (2018). Distributional reinforcement learning with quantile regression. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2892-2901. https://doi.org/10.1609/aaai.v32i1.11791

[19] Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. The Journal of Risk, 2(3), 21-41. https://doi.org/10.21314/JOR.2000.038

[20] Chow, Y., Ghavamzadeh, M., Janson, L., & Pavone, M. (2018). Risk-constrained reinforcement learning with percentile risk criteria. Journal of Machine Learning Research, 18(167), 1-51.

[21] Tamar, A., Chow, Y., Ghavamzadeh, M., & Mannor, S. (2015). Policy gradient for coherent risk measures. Advances in Neural Information Processing Systems, 28, 1468-1476.

[22] Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 3(1), 1-122. https://doi.org/10.1561/2200000016

[23] Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. In Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70 (pp. 22-31). PMLR.

[24] Liu, J., Wang, J., & Lin, H. (2025). Coordinated physics-informed multi-agent reinforcement learning for risk-aware supply chain optimization. IEEE Access, 13, 190980-190993. https://doi.org/10.1109/ACCESS.2025.3629716

[25] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR).

[26] Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. Advances in Neural Information Processing Systems, 12, 1057-1063.

[27] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 3207-3214. https://doi.org/10.1609/aaai.v32i1.11694