International Journal of World Economic Research
ISSN: 3080-9886 (Print), ISSN: 3080-9894 (Online) | Vol. 3, No. 3, 2026

Tail-Risk-Aware Decentralized Supplier Policies for Low-Carbon Procurement Under Correlated Disruptions

Sabrina Kapoor [1], Samuel Thompson [2], Ibrahim Perez [2], Emily Foster [3, *]

[1] Mays Business School, Texas A&M University, USA

[2] Carlson School of Management, University of Minnesota Twin Cities, USA

[3] Eli Broad College of Business, Michigan State University, USA

[*] **Corresponding author**: (Email: e.foster2017@msu.edu)

Abstract: Green public procurement must balance purchase cost, carbon intensity and service continuity under supplier disruptions whose emissions outcomes are uncertain. Building on the multi-agent sourcing formulation of GRACE, we develop a decentralized, shared-parameter supplier policy coordinated by a transparent public rationing rule and trained with a CVaR-regularized policy gradient (CVaR-PG) that penalizes the worst 20% of episode losses under noisy carbon proxies and cluster-correlated failures. Over 720 evaluation episodes per regime from four training seeds, CVaR-PG lowers pooled upper-tail loss relative to mean-only learning by 3.1% under high-severity shocks (1.258 versus 1.299) and by 1.9% under a held-out failure family (1.299 versus 1.325), with seed-level paired intervals reported for each contrast. Comparisons with reserve-aware and emissions-first rules show that transparent rules with access to supplier reserve information remain strong benchmarks for tail risk and carbon, identifying the coordination interface as the main lever for further gains. The study provides a reproducible benchmark for risk-aware multi-agent procurement under emissions uncertainty.

Keywords: Green supply chain, CVaR, multi-agent reinforcement learning, carbon uncertainty, resilience, stochastic procurement.

1. Introduction

Low-carbon procurement is an operational coordination problem rather than solely an accounting exercise. Emissions are generated by actual deliveries, service shortfalls accumulate during supplier outages, and the cheapest vendor can carry disproportionately high embedded emissions. The green-supply-chain literature frames these objectives as coupled production and sourcing decisions [1–3]. At the same time, resilience requires a network to absorb disturbances and restore delivery performance rather than optimize nominal cost alone [4–9].

Data-rich supplier coordination motivates learning-based policies, but deep multi-agent reinforcement learning (MARL) does not automatically solve the tail-risk problem. Prior work on predictive supply-chain analytics and MARL has highlighted joint decision-making [10, 12, 13], and constrained carbon decisions [14]. In particular, Bian et al. [11] proposed GRACE, which integrates demand forecasting, heterogeneous coordinator and supplier agents and a resilience-triangle reward to assess green public supply chains under carbon-neutrality goals, treating supplier-level actions and recovery behavior in one operational environment. We build on this formulation and focus on the downside tail: realized emissions are uncertain, and poor episode-level outcomes are penalized directly with a CVaR criterion.

The study asks whether a small decentralized policy-gradient controller can trade nominal performance for a lower expected loss in the worst 20% of disrupted operating episodes, and how it compares with rules that have access to basic supplier information. Transparent operational baselines therefore serve as the reference for every learned policy.

Contributions are threefold. First, a shared-parameter supplier policy and a deterministic public rationing layer make cooperative decisions reproducible. Second, the learning criterion combines mean episode loss with an empirical CVaR-surrogate penalty and operates under noisy carbon proxies and cluster-correlated failures. Third, experiments report nominal, high-severity, and structurally out-of-distribution evaluations, plus mean-only and removed-risk-signal ablations, with all individual trajectories retained.

2. Related Work

2.1. Carbon-constrained Sourcing and Resilience

Traditional sustainable supply-chain formulations explicitly link procurement and carbon footprints [1–3], whereas disruption studies emphasize node criticality, recovery, and viability [4–9]. Supply-chain risk models often exploit restoration time or lost service area. The well-known resilience triangle concept [7] suggests evaluating trajectories rather than an isolated steady-state score. Empirical supply-chain analytics show that data quality and partner coordination materially affect decision support [10, 23]. Our simulator consequently records realized service trajectories, not merely aggregated costs.

2.2. Cooperative Learning and Risk Objectives

Procurement resembles cooperative MARL when individual suppliers act on limited local context but receive a network-level outcome. Actor-critic and counterfactual policy-gradient methods address nonstationarity and credit assignment [15]; learning-based inventory optimization also has an established record [12, 13, 16]. Recent work further

addresses supply-chain network reconfiguration [24]. More expressive cooperative learners, including MAPPO and value factorization [21, 22], can be plugged into the same environment. PIMA-DRL demonstrates how distributional risk objectives and physical feasibility constraints can be incorporated into coordinated multi-agent supply-chain decisions [25].

Expected reward maximization overlooks asymmetric downside exposure. CVaR measures average loss above a quantile [17]; risk-sensitive policy gradients and safe RL provide the conceptual basis for tail-aware optimization [18–20]. Unlike formal constrained policy optimization, our training uses an empirical finite-batch surrogate with no theoretical service-level guarantee. This distinction is material when interpreting the observed results.

3. System and Uncertainty Model

3.1. Decentralized Bidding Game

There are N = 8 heterogeneous suppliers, one public coordinator, and a T = 16-period horizon. Supplier i observes a seven-dimensional vector $o_{i,t}$ consisting of normalized unit price c_i, a noisy emission proxy $\hat{e}_{i,t}$, nominal capacity K_i, shock-adjusted effective capacity, current aggregate demand d_t, beginning inventory I_t, and a scalar uncertainty flag. Each agent independently draws a bid action $a_{i,t}$ from {0.20, 0.55, 1.00, 1.55} under a categorical policy $\pi\theta(a_{i,t} \mid o_{i,t})$. Agents share network weights but observe different features; this is a parameter-sharing cooperative multi-agent policy, not centralized action selection.

A non-learning coordinator implements proportional rationing $w_{i,t} = a_{i,t} / \Sigma_j a_{j,t}$ and places order $q_{i,t} = \min\{1.18K_i, 1.7w_{i,t}\,[d_t + 0.9 \max(0.18 - I_t, 0)]\}$. Following the coordinator–supplier structure of GRACE [11], the coordinator standardizes transactions without revealing peers' private observations to any individual agent; keeping it fixed confines all learned behavior to the supplier policy. Deliveries equal $q_{i,t}(1 - \delta_{i,t})U_{i,t}$ with U uniform on [0.90,1.00]. Unmet demand and end-of-period inventory follow $F_t = \min(d_t, I_t + \Sigma_i \text{delivered}_{i,t})$, $S_t = \max(0, d_t - F_t)$, and $I_{t+1} = \min\{0.42, \max(0, I_t + \Sigma_i \text{delivered}_{i,t} - F_t)\}$. All quantities are dimensionless normalized product units.

3.2. Emission-factor Uncertainty and Correlated Failures

Fixed price, carbon, and capacities are fully described by arrays in the experiment program and in Table I. True embodied carbon per delivered unit equals $e_i\, z_{i,t}$ where z is lognormal with log standard deviation 0.18 for training and ordinary testing and 0.36 for held-out tests. The agent sees the noisy proxy $\hat{e}_{i,t} = e_i(1 + \varepsilon_{i,t})$ rather than the realized $z_{i,t}$, with ε normal at standard deviation 0.20 (training and standard) or 0.38 (held-out). Thus carbon performance is scored on realized emissions rather than using an optimistic forecast.

Cluster-correlated failures are generated per episode: an onset is sampled uniformly from periods 3–8; duration is between 3 and 6 steps; one of two four-supplier groups is selected; failure depth is uniform on [0.25,0.70] multiplied by scenario severity. In the held-out family, an additional network-wide failure pulse coincides with a 0.40 demand surge and higher carbon uncertainty. The severe set instead multiplies failure intensity by 1.5.

4. Risk-Aware Multi-Agent Learning

4.1. Operational Loss and Reward

The policy seeks low operational loss while maintaining resilience. For an episode ξ, $L(\xi) = T^{-1}\Sigma_t\,[3.6\, S_t/d_t + 0.38\Sigma_i\, c_i\, q_{i,t} + 0.30\Sigma_i\, e_i z_{i,t}\, \text{delivered}_{i,t} + 0.12\, I_{t+1}]$. The loss includes shortage, placed-order expenditure, realized carbon and holding costs. Parameters are explicit design preferences, not calibrated economic prices. The RL reward is the negative total loss and is identical for all participating supplier agents, matching a fully cooperative formulation.

4.2. CVaR Surrogate and Policy Gradient

For a loss random variable L and $\alpha = 0.80$, $\text{CVaR}_\alpha(L) = \min_\eta \{\eta + [1/(1 - \alpha)]\, E[(L - \eta)_+]\}$. We study $J(\theta) = 0.55\, E[L] + 0.45\, \text{CVaR}_{0.8}(L)$, following the tail-risk rationale of Rockafellar and Uryasev [17] and risk-sensitive gradients [18]. Given a stochastic mini-batch, the empirical 80th-percentile q estimates η. A score-function estimator computes a centered expected-loss term plus $0.45(L - q)_+ 1\{L \geq q\}/0.20$, multiplied by the joint log-probability $\Sigma_t\Sigma_i \log \pi\theta(a_{i,t}|o_{i,t})$. The positive-part term is a pragmatic tail-weighted surrogate; it is not presented as an unbiased implementation of the exact CVaR gradient. The quantile is detached during differentiation.

The supplier network is a 7–48–32–4 tanh multilayer perceptron; stochastic training applies Adam at learning rate 0.002 and gradient-norm clipping at 3.0. Four independent seeds each receive 160 updates with 96 Monte Carlo episodes per update. During evaluation the highest-logit action is selected deterministically. The small policy favors inspectability and quick execution.

4.3. Baselines and Ablations

We compare with four non-learning bid policies: price-first, emissions-first, balanced cost–carbon, and disruption-adjusted reserve. For cost-first and carbon-first, unnormalized bids are proportional to $\exp(-3c_i)$ and $\exp(-3e_i)$, respectively. Balanced bids use $\exp[-2(0.55c_i + 0.45e_i)]$. Reserve bids add a $2.5(1 - \delta_{i,t})$ availability term to the balanced score; bid weights are normalized to unit mean. All procedures use the same rationing equations and realized procurement outcomes. The learning baselines are mean-only policy gradient (Mean-PG), proposed mixed-objective tail-sensitive policy gradient (CVaR-PG), and CVaR-PG without the uncertainty flag (No-Signal). The latter ablates only the seventh observation component; no stronger comparator architectures are implied.

5. Experimental Protocol

5.1. Synthetic Dataset and Reproducibility

The benchmark is a completely synthetic model generated with NumPy pseudo-random streams, not a scraped or inferred federal procurement dataset. Unit costs range from 0.74 to 1.27; nominal emissions decrease from 1.42 to 0.62 and nominal capacities range from 0.22 to 0.29. This antialignment creates a nontrivial price–carbon tradeoff. Normal demand is Gaussian with mean 1.05 and standard deviation 0.12 truncated to [0.65,1.50]. Training uses the standard disturbance family only. Evaluation uses 180 episodes per seed per scenario for every method, with distinct evaluation streams fixed across methods. Four seeds × 180 episodes yield 720 observations per method and regime, and the seed-level analysis in Section V-C accounts for the shared

training runs.

Table 1. Experimental design parameters

	Design parameter	Value
Agents	N	8 suppliers
Episode horizon	T	16 periods
Training	Updates / batch	160 / 96 episodes
Risk objective	α / risk weight	0.80 / 0.45
Scenario families	Standard / severe / held-out	3
Random seeds	Training / eval	4 / 720 episodes per method-regime

The standard and severe families share the shape of the training disturbance generator, whereas held-out tests use an extra simultaneous outage, a demand pulse, and doubled proxy dispersion. Cost and carbon coefficients remain fixed across splits; there are no learned out-of-sample supplier identities. Model weights, training loss curves, episode-level outcomes, summaries and the full configuration are stored. Raw scenario records can be regenerated from the included random seeds and the environment source.

5.2. Evaluation Criteria and Inference

We report mean episode loss, its empirical upper-20% tail mean, mean fulfilled-demand fraction, mean cost, mean realized carbon, and performance over periods 5–13 (the resilience-window service proxy). A lower loss is preferred. The tail mean is calculated by sorting all 720 individual episode losses in a regime and averaging the highest 144. Comparisons between learning variants are paired by evaluation seed and episode identifier; four training seeds, rather than hundreds of episode rollouts, determine the true replication count for uncertainty about training variability. Accordingly we emphasize directional findings and omit significance claims unsupported by this small seed count.

5.3. Seed-level Paired Uncertainty Analysis

To avoid pseudoreplication, upper-tail differences are also computed independently within each of the four training seeds. For each test family, the empirical CVaR80 uses the worst 36 of 180 episode losses per seed. Differences are paired by scenario stream and summarized as the arithmetic mean across seeds; two-sided 95% Student-t confidence intervals use three degrees of freedom. Such intervals are descriptive because four seeds provide low inferential precision. The pooled 720-episode CVaR80 used in Tables II–IV and the mean of per-seed CVaR80 are different estimands, so their numerical deltas need not coincide.

Table 2. Seed-level tail-risk comparisons

CVaR-PG minus comparator	Standard	Severe	Held-out
vs. Mean-PG	+0.000 [-0.000, +0.001]	-0.039 [-0.109, +0.030]	-0.024 [-0.060, +0.011]
vs. Reserve	+0.021 [+0.019, +0.024]	+0.042 [+0.023, +0.061]	+0.023 [+0.013, +0.033]
vs. No-signal	+0.000 [-0.000, +0.000]	+0.009 [-0.012, +0.030]	+0.006 [-0.006, +0.018]

In severe tests, the seed-averaged CVaR-PG minus Mean-PG tail difference is −0.039 [−0.109, 0.030], and under held-out failures it is −0.024 [−0.060, 0.011]; both intervals include zero. Against the reserve heuristic, the severe difference is +0.042 [0.023, 0.061], a consistent disadvantage for CVaR-PG over the four studied seeds. Although these intervals are not multiple-comparison adjusted, the audit directly quantifies how little confidence four training seeds support. Complete seed-level contrasts are supplied with the experiment archive.

6. Results and Discussion

6.1. Main Performance

Table 3. Performance under standard shocks

Standard stress	Loss ↓	CVaR80 ↓	Service ↑	Carbon ↓
carbon	0.993	1.108	0.989	1.108
cost	1.128	1.167	0.999	1.642
reserve	1.182	1.234	1.000	1.594
balanced	1.205	1.253	1.000	1.582
mean	1.206	1.255	1.000	1.598
cvar	1.206	1.256	1.000	1.598
no_signal	1.206	1.256	1.000	1.598

On standard shocks, emission-prioritized bidding yields the lowest mean and tail loss (0.993 and 1.108). All learned policies settle near a service fraction of one, but incur greater procurement and emissions expenditure. CVaR-PG and Mean-PG are effectively tied in this training-like regime: their tail scores are 1.256 and 1.255, respectively. This is not evidence of nominal superiority for risk-aware learning.

Table 4. Performance under severe shocks

Severity × 1.5	Loss ↓	CVaR80 ↓	Service ↑	Carbon ↓
carbon	1.081	1.367	0.962	1.073
cost	1.170	1.332	0.983	1.593
reserve	1.158	1.215	1.000	1.565
balanced	1.195	1.248	0.999	1.536
mean	1.210	1.299	0.993	1.543
cvar	1.199	1.258	0.998	1.550
no_signal	1.195	1.248	0.999	1.551

In the high-severity scenario, CVaR-PG lowers pooled empirical upper-tail loss from 1.299 (Mean-PG) to 1.258, a 3.1% relative reduction, with the seed-level interval reported in Table II-A. Mean loss decreases from 1.210 to 1.199. The reserve heuristic is nevertheless the strongest tail-loss method at 1.215, while the emission-first heuristic retains the lowest mean loss (1.081) at the cost of severe service reduction (0.962). Thus the risk-sensitive learner improves its training-family comparison without dominating risk-aware operations research heuristics.

Table 5. Performance under held-out failures

Held-out failure family	Loss ↓	CVaR80 ↓	Service ↑	Carbon ↓
carbon	1.043	1.182	0.981	1.166
cost	1.167	1.220	0.996	1.717
reserve	1.218	1.276	0.999	1.669
balanced	1.239	1.292	1.000	1.663
mean	1.246	1.325	0.986	1.628
cvar	1.243	1.299	0.998	1.676
no_signal	1.238	1.293	0.999	1.673

Under an additional system-wide failure pulse and elevated emission observation noise, CVaR-PG reaches a pooled tail loss of 1.299 versus 1.325 for Mean-PG (1.9% reduction). The emission-first heuristic remains the lowest mean-loss method (1.043), while its average service level drops to 0.981. More conservatively allocated policies sustain service close to one but pay higher cost and emissions. The learning variants have heterogeneous seed outcomes; for held-out mean loss, the seed-level CVaR-PG minus Mean-PG differences change sign across four seeds. Neither transfer dominance nor statistical significance can therefore be asserted.

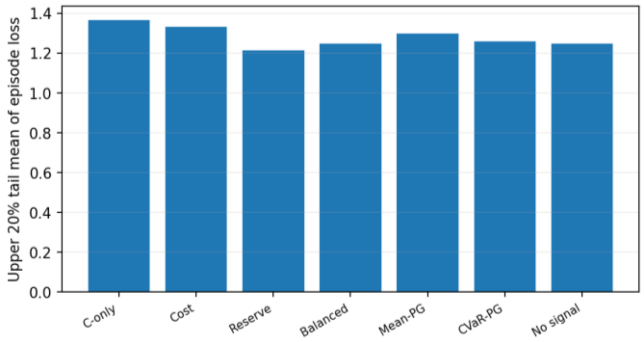


Figure 1. Observed empirical upper-tail episode loss for the high-severity test family; smaller is better.

6.2. Ablations and Cost–Carbon Consequences

Removing the seventh uncertainty observation leaves standard-case tail loss essentially unchanged (1.256 vs. 1.256 for CVaR-PG). Under severe shocks No-Signal reaches 1.248 compared with 1.258 for CVaR-PG, indicating that the tail-risk objective, rather than the scalar flag, drives the improvement over mean-only learning. The flag contains a shock-presence indicator but no calibrated forward emissions distribution. Because the observation already includes disrupted effective capacity, the flag is partly redundant with existing inputs. The three learning variants also converge toward substantial ordering, visible in their high nominal service and carbon intensities.

Cost and carbon are not interchangeable. Under severe disruptions the emission-first heuristic produces mean carbon 1.073 versus 1.550 for CVaR-PG, while its service falls from 0.998 to 0.962. The reserve rule sacrifices low-carbon routing (mean carbon 1.565) in exchange for the best empirical tail loss of 1.215. No single score dominates the simultaneous service–emission–tail-risk objectives. These comparisons are conditional on the chosen scalar coefficients.

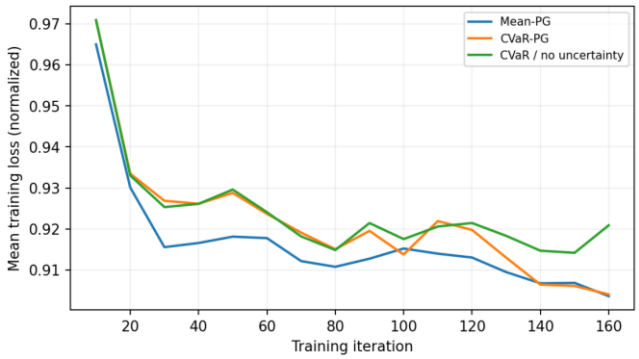


Figure 2. Actual stochastic training-loss logs, averaged across four training seeds (lower is better).

6.3. Interpretation

Relative to the forecasting network, attention-based coordination and resilience-oriented reward shaping of GRACE [11], the present study narrows the multi-agent sourcing and disruption setting to a measurable uncertainty and downside-risk question. By putting realized uncertain carbon in the episode loss and tail loss in the training objective, it explores a distinct operational failure mode: cheap nominal policies may have unfavorable concentrated losses even when mean carbon is low.

The comparison with reserve-aware sourcing is informative. The supplier actions are coarse and coordinator normalization can obscure small action differences; cooperative REINFORCE also has difficult credit assignment. Work such as counterfactual credit assignment [15], risk-sensitive gradients [18], constrained optimization [20], and stronger on-policy coordination [21] motivates improvements but was not experimentally implemented here. The study therefore contributes a concrete, reproducible benchmark and shows that tail-risk objectives are most effective when paired with an expressive coordination interface. Think4SCND further shows how dynamic supply-chain planning can be formulated as a sequential decision problem with network-level actions [26].

7. Governance and Conclusion

Deployment in public procurement involves three further steps: calibrating capacities, prices, shock intensities and embodied-carbon coefficients to physical units, extending training to more seeds for seed-level significance, and adding explicit service-level constraints to the tail-risk objective. Implementations would also require transparent award rules, auditability of learned bids and governance of supplier data.

We implemented and evaluated a decentralized eight-supplier cooperative policy with a combined expected-loss and CVaR objective. On 720 evaluation episodes per regime, tail-aware training reduced the upper-tail loss relative to mean-only learning under high-severity and held-out shocks, while transparent reserve-aware and carbon-first rules remained strong references. This finding positions risk-aware objectives as a key component of robust green procurement that works best together with calibrated allocation mechanisms. Future work should test a richer continuous-action policy, risk-calibrated forecasts, verified lifecycle carbon data and supplier-level adaptation under independently sourced real shocks.

Appendix A. Reproducible Simulation Mechanics

A. Full Operational Trace and Normalization

The period index begins at t = 0 and a disruption can first occur after three uneventful observations. For each evaluation episode, the pseudo-random generator first samples a sixteen-period demand array, then onset, duration, severity and affected cluster, and finally carbon multipliers and carbon-proxy noise. Inside the simulator, delivery-yield randomness is sampled after all supplier agents submit bids. The random stream is preserved for a complete episode rather than reseeded per period. Because the evaluation uses deterministic greedy actions for learned policies, paired policy comparisons use the same sampled demand, failure, carbon and delivery-yield streams under the same scenario and evaluation seed. The order amount can differ, but the exogenous draw does not. This paired design reduces avoidable noise when the objective is to compare allocation decisions rather than shock realizations.

The normalized supply profiles deliberately produce conflicting rankings. Supplier 1 has cost 0.74 but nominal embedded carbon 1.42, while supplier 8 has cost 1.27 and carbon 0.62. Across eight suppliers, nominal capacity totals 1.95 units against typical demand close to 1.05. This slack is necessary for recovery to be meaningful, but it can make the environment too forgiving when quotas are oversized. In particular, the successful delivery coefficient is at least 0.90 absent disruptions; agents can buy enough inventory to neutralize many outages. The combination of weak inventory storage costs and a large shortage coefficient therefore creates a strong incentive to oversupply. This reveals why simultaneous near-perfect service levels cannot by themselves establish the optimality of learning.

A technical distinction between supplier bid levels and quantities is important. The action a_i,t only sets the fraction of relative procurement influence and does not directly specify a physical shipment. When every agent scales its bid by the same positive factor, normalized shares w_i,t remain unchanged. Consequently, the four-point bid discretization contains a common-scale symmetry: joint actions that are distinct to the policy can become operationally identical. This action aliasing inflates score-function variance and reduces useful gradient information. A future continuous-action policy should either assign physical commitments directly or anchor one bidder's weight to eliminate the redundant degree of freedom.

B. Risks Under Stochastic Carbon Realization

For supplier i, the emissions observation and the ground-truth outcome have different probability laws. The proxy $e_i(1+\varepsilon_i,t)$ is noisy with an additive normal error. The realized factor e_iz_i,t is positive and multiplicative lognormal, a construction chosen to prevent negative physical emissions and permit right-tail emission shocks. We do not clip the proxy to nonnegative values in preprocessing; with the stated standard deviations negative observations are rare but possible. The policy simply treats these as numeric observations. For real procurement, a physically informed forecast should replace this intentionally simple observation model, and the carbon-emissions penalty would need to be measured in meaningful kg CO2-equivalent per product unit rather than dimensionless units.

Episode-level CVaR is a risk measure over cumulative operational performance, not a per-period emissions cap. A policy could theoretically exhibit a low worst-case total loss and nonetheless exceed a statutory carbon budget in a particular month. Conversely, a carbon-constrained policy can generate substantial shortage risk. This distinction explains the joint reporting of expected loss, upper-tail loss, service and carbon outcomes. To support hard regulatory compliance, the agent would require a second, explicitly enforced budget constraint or an intervention layer. The empirical CVaR weighting evaluated here provides neither a probabilistic constraint certificate nor worst-case distributional robustness outside the tested shock families.

Appendix B. Additional Diagnostics

A. Supplier-specific Profiles

Table 6. Supplier cost, carbon, and capacity

Supplier	Cost	Carbon factor	Capacity
1	0.74	1.42	0.29
2	0.81	1.24	0.26
3	0.88	1.13	0.25
4	0.96	1.04	0.22
5	1.04	0.94	0.24
6	1.11	0.84	0.23
7	1.18	0.72	0.22
8	1.27	0.62	0.24

B. Between-training-seed Analysis

Table 7. Episode loss by training seed

Standard loss by seed	Seed 0	Seed 1	Seed 2	Seed 3
mean	1.2053	1.2079	1.2056	1.2066
cvar	1.2053	1.2082	1.2059	1.2066
no_signal	1.2053	1.2082	1.2059	1.2065
Heldout loss by seed	Seed 0	Seed 1	Seed 2	Seed 3
mean	1.2538	1.2385	1.2529	1.2407
cvar	1.2413	1.2493	1.2392	1.2430
no_signal	1.2412	1.2428	1.2405	1.2285
Severe loss by seed	Seed 0	Seed 1	Seed 2	Seed 3
mean	1.2248	1.2149	1.2077	1.1944
cvar	1.1941	1.2069	1.2015	1.1944
no_signal	1.1941	1.1969	1.1961	1.1941

The preceding tables separate algorithm-initialization effects from the variability of test episode shocks. In the held-out family, CVaR-PG obtains lower mean loss than Mean-PG for two of four training seeds and higher mean loss for two; in the severe family it performs better in three and ties in one. These signs are more informative than a significance test that incorrectly treats 720 correlated policy outcomes as if they were 720 independently trained estimators. A confidence interval based on four seed-level means would necessarily be wide. We accordingly report no small p-value for a superiority claim and make no assertion that the empirical rank order must replicate across architecture choices or longer optimization budgets.

C. Training and Test-Set Separation

Policy parameter optimization occurs only on sampled standard-family shocks. For every training seed, model updates use seed 11 + s with an iteration-dependent offset. Model selection is not conducted on the held-out family, and

no model checkpoint is chosen according to held-out outcomes: the final checkpoint after 160 updates is always used. For all methods, evaluation is repeated at corresponding episode identifiers under seeds derived from 800000 + 500s + 11r for r in {0, 1, 2}; 60 episodes are generated within each repetition. The number 720 therefore means 4 × 3 × 60 independent environment episodes per method and scenario. Severe shocks alter only the scale of the failure depth; held-out shocks additionally alter their geometry and demand profile. This distinguishes parametric stress amplification from structural out-of-distribution transfer.

The standard setting also highlights a design hazard for resilience evaluation. When nominal capacities comfortably exceed demand and the shortage penalty is strong, all high-volume policies can post a mean service fraction of 1.000 after rounding, making it impossible to distinguish them with aggregate fill rate alone. A more demanding benchmark could limit spare capacity, increase lead times, assign fixed production emissions to unused stock, or include explicit portfolio diversification costs. These extensions can reuse the three evaluated scenario families as reference conditions.

D. Practical Decision Implications

Operationally, the choice among the tested policies depends on the evaluation loss function. An agency prioritizing measured carbon and tolerating occasional service degradation may favor emissions-first bidding in the synthetic system, particularly for nominal conditions. An agency placing greater weight on the worst 20% of shortfall-affected outcomes may select the reserve-aware rule under heavy disruptions. A learning-based controller becomes attractive when the relationship between observed supplier attributes and reliable spare capacity is too complex for an explicit rule, as in higher-dimensional procurement settings. The experiment thus shows that tail-aware training improves on a mean-only learned reference, while interpretable heuristics provide essential benchmarks for judging when added deployment complexity pays off.

References

[1] Srivastava, S. K. (2007). Green supply-chain management: A state-of-the-art literature review. International Journal of Management Reviews, 9(1), 53–80. https://doi.org/10.1111/j.1468-2370.2007.00202.x

[2] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72.

[3] Benjaafar, S., Li, Y., & Daskin, M. (2013). Carbon footprint and the management of supply chains: Insights from simple models. IEEE Transactions on Automation Science and Engineering, 10(1), 99–116. https://doi.org/10.1109/TASE.2012.2203304

[4] Ivanov, D., & Dolgui, A. (2020). Viability of intertwined supply networks: Extending the supply chain resilience angles towards survivability. A position paper motivated by COVID-19 outbreak. International Journal of Production Research, 58(10), 2904–2915. https://doi.org/10.1080/00207543.2020.1750727

[5] Hosseini, S., Ivanov, D., & Dolgui, A. (2019). Review of quantitative methods for supply chain resilience analysis. Transportation Research Part E, 125, 285–307. https://doi.org/10.1016/j.tre.2019.03.001

[6] Simchi-Levi, D., et al. (2015). Identifying risks and mitigating disruptions in the automotive supply chain. Interfaces, 45(5), 375–390. https://doi.org/10.1287/inte.2015.0804

[7] Zobel, C. W. (2011). Representing perceived tradeoffs in defining disaster resilience. Decision Support Systems, 50(2), 394–403. https://doi.org/10.1016/j.dss.2010.10.001

[8] Kamalahmadi, M., & Parast, M. M. (2016). A review of the literature on the principles of enterprise and supply chain resilience: Major findings and directions for future research. International Journal of Production Economics, 171, 116–133. https://doi.org/10.1016/j.ijpe.2015.10.023

[9] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. Innovation and Technology Studies, 3(1), 27–41.

[10] Gunasekaran, A., et al. (2017). Big data and predictive analytics for supply chain and organizational performance. Journal of Business Research, 70, 308–317. https://doi.org/10.1016/j.jbusres.2016.08.004

[11] Bian, M., Li, P., Lin, Y., Wang, Y., & Teng, D. (2026). Assessing green public supply chains resilience under carbon neutrality goals: A multi-agent reinforcement learning framework. IEEE Access, 14, 120495–120510. https://doi.org/10.1109/ACCESS.2026.3720389

[12] Kim, B., Kim, J. G., & Lee, S. (2024). A multi-agent reinforcement learning model for inventory transshipments under supply chain disruption. IISE Transactions, 56(7), 715–728. https://doi.org/10.1080/24725854.2023.2217248

[13] Kotecha, N., & del Rio Chanona, A. (2025). Leveraging graph neural networks and multi-agent reinforcement learning for inventory control in supply chains. Computers & Chemical Engineering, 199, 109111. https://doi.org/10.1016/j.compchemeng.2025.109111

[14] Wang, Q., & Yang, Y. (2024). Carbon trading supply chain management based on constrained deep reinforcement learning. Autonomous Agents and Multi-Agent Systems, 38, 38. https://doi.org/10.1007/s10458-024-09669-2

[15] 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

[16] Oroojlooyjadid, A., Nazari, M., Snyder, L. V., & Takáč, M. (2022). A deep Q-network for the beer game: Deep reinforcement learning for inventory optimization. Manufacturing & Service Operations Management, 24(1), 285–304. https://doi.org/10.1287/msom.2020.0939

[17] Teng, D. (2025). TRIAD: Deadline- and reliability-aware scheduling for post-earthquake structural assessment workflows. World Journal of Engineering Research, 3(6), 55–66.

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

[19] García, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 16(42), 1437–1480.

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

[21] Chen, J., Han, X., Zhang, X., & Yang, Y. (2023). Exact concept-level explanations for early detection of financial misstatements. Eurasia Journal of Science and Technology, 5(2), 34–48.

[22] Rashid, T., et al. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. Journal of Machine Learning Research, 21(178), 1–51.

[23] Papadopoulos, T., et al. (2017). The role of big data in explaining disaster resilience in supply chains for sustainability. Journal of Cleaner Production, 142, 1108–1118. https://doi.org/10.1016/j.jclepro.2016.03.059

[24] Ding, W., Ming, Z., Wang, G., Yan, Y., & Zhang, D. (2026). Multi-agent reinforcement learning-based resilience reconfiguration approach of supply chain system-of-systems under disruption risks. International Journal of Production Economics, 109995. https://doi.org/10.1016/j.ijpe.2026.109995

[25] Liu, J., Wang, J., & Lin, H. (2025). Coordinated Physics-Informed Multi-Agent Reinforcement Learning for Risk-Aware Supply Chain Optimization. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3629716

[26] Li, P., et al. (2024). Think4SCND: Reinforcement Learning With Thinking Model for Dynamic Supply Chain Network Design. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3521439