WEP Warwick Evans Publishing

Academic Journal of Management Science and Engineering

ISSN: 3080-9827 (Print), ISSN: 3080-9835 (Online)
Volume 1, Number 7, Year 2026
DOI: https://doi.org/10.62051/ajmse.v1n7.13



Uncertainty-Aware Thinking-Model Planning for Dynamic Supply Chain Network Design

-- A Risk-Sensitive Candidate-Sequence Evaluation Study

Lana Harris

Wisconsin School of Business, University of Wisconsin–Madison, WI, USA

ABSTRACT

Strategic facility activation and operational inventory allocation are coupled sequential decisions under demand volatility. This study develops an uncertainty-aware thinking-model decision evaluator (UA-TD) for multi-period supply chain network design. A bootstrapped one-step demand predictor and empirical multivariate innovations generate scenario paths; exact cost accounting and a feasible shipment recourse rule evaluate enumerated facility sequences. An expectation–conditional-value-at-risk objective extends the candidate-sequence evaluation of the Think4SCND thinking model, retaining the first action and replanning after each observed demand. Fully synthetic, explicitly specified experiments include three warehouses, four customer zones, ten periods, and 180 held-out episodes per demand regime. Relative to risk-neutral ensemble planning, UA-TD increases mean fill rate by 0.77, 0.85, and 0.52 percentage points in normal, volatile, and disruption conditions, respectively, with paired bootstrap intervals excluding zero. Mean episode cost is 0.68% higher in normal conditions and 0.29% and 0.20% lower in volatile and disrupted conditions. The evidence supports a measurable service–cost trade-off, not uniform dominance. The study contributes an inspectable, risk-sensitive planning module and a fully reproducible evaluation protocol for dynamic facility and inventory decisions.

KEYWORDS

Supply Chain Network Design; Model-based Reinforcement Learning; Uncertainty Quantification; Scenario Planning; Conditional Value-at-risk; Inventory Control.

1. INTRODUCTION

Supply chain network design links long-lived capacity decisions to short-lived customer requirements. Facility activation, sourcing, transportation, and safety stock interact over time, so a favorable configuration at one date may be an expensive configuration at the next. Location and network-design reviews document the value of modeling these decisions jointly [1], [2]. Disruption-specific work further distinguishes the cost of proactive capacity from the cost of recovering after a shock [3], while digital-twin studies argue for repeated revision of decisions as operational state changes [4].

Inventory is not merely a secondary constraint: it carries the consequences of past decisions into future periods, an insight central to multi-echelon inventory research [5]. Reinforcement learning provides one possible framework for this dependence; recent reviews [6] and numerical studies [7] demonstrate its relevance to difficult inventory policies. Yet a point-estimate future is a weak basis for opening facilities under uncertain, correlated demand. Underprediction induces shortages, whereas conservative opening schedules create avoidable fixed costs.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

Think4SCND formulates dynamic supply chain network design as a Markov decision process with mixed discrete–continuous actions and couples a Supply Chain Transformer Network with a thinking model that predicts future states, estimates rewards and evaluates candidate decision sequences before acting; trained end to end, this combination of model-free learning and model-based planning reached near-optimal solutions at a fraction of the computational cost and transferred from medium-sized to larger unseen instances [8]. The thinking model is the component that makes those decisions forward-looking, and it is the direct starting point of this study. We investigate whether augmenting its candidate evaluation with stochastic scenario paths and upper-tail cost sensitivity improves decision quality when demand uncertainty is consequential. The focus is not network representation capacity, but what information the evaluator should preserve about the future.

The contributions are threefold. First, we adapt the state-predictor/reward-evaluator/candidate-refinement loop to a feasible mixed discrete–continuous facility–inventory simulator. Second, we replace a single predicted trajectory with bootstrapped dynamics and jointly sampled residual innovations, ranking plans by an expectation–CVaR functional. Third, we provide held-out paired evaluations, a no-risk ablation, a no-lookahead ablation, and an unseen demand-shift experiment with released seeds and episode-level results.

2. RELATED WORK AND POSITIONING

Model-based reinforcement learning spans explicit dynamics learning and planning as well as policy improvement with simulated trajectories [9]. Probabilistic ensemble trajectory sampling [10] shows how uncertainty can be propagated during planning; model-based policy optimization [11] emphasizes that shorter imagined rollouts can mitigate compounding model error. These methods concern general control problems rather than facility activation with continuous delivery recourse. Ensembles are also a pragmatic mechanism for predictive uncertainty [12], although calibration needs separate attention [13].

For supply chains, model predictive control offers a distinct route to responsive operation. A recent distributed robust MPC formulation considers uncertain perishability and forecast errors in multi-stage networks [14]. Our decisions instead concern discrete warehouse availability, with realized demand allocated through bounded continuous flows. We retain an explicit two-period horizon, evaluating all activation sequences at this small scale; there is no approximation by a learned value function in the present implementation. Multi-agent reinforcement learning for resilient green public supply chains demonstrates the value of testing network decisions against disruption-sensitive operational objectives [25].

Risk-sensitive optimization provides a formal vocabulary for comparing mean performance with the loss tail. CVaR measures the expected severity of outcomes beyond a chosen quantile [15], and robust optimization emphasizes the price paid for conservatism [16]. Scenario-based optimization [17], sample-based CVaR [18], and risk-sensitive Markov decisions [19] motivate the chosen criterion. Their theoretical guarantees do not automatically transfer to finite-scenario, misspecified demand models, so we evaluate realized costs and fill rates rather than asserting a safety guarantee.

Domain-specific information can stabilize inventory-oriented reinforcement learning, as demonstrated through reward shaping for perishable stocks [20]. More broadly, predictive-to-prescriptive analytics [21] separates accuracy from downstream optimization value; inventory simulations on M5 data reveal that their ranking can differ [22]. The M5 forecasting competition provides a high-quality real-world demand reference [23], though it is not a substitute for calibrated counterfactual warehouse and transportation attributes. Risk-aware decision-focused demand forecasting emphasizes measuring forecast quality through subsequent inventory and stock-out consequences, providing a strong decision-oriented reference for supply-chain evaluation [26].

3. DYNAMIC NETWORK AND DECISION PROCESS

3.1. Network, State, and Operational Recourse

The benchmark contains W = 3 candidate warehouses, Z = 4 customer zones, a single aggregate product, and T = 10 periods. Warehouse j has production capacity Pj, storage limit Sj, fixed operating expense fj, and reopening cost oj. At the beginning of period t, the controller observes warehouse inventory It, prior activation zt−1, the previous demand vector dt−1, period index, and a publicly known volatility indicator. It chooses a binary facility vector $z_t \in \{0,1\}^3$. After demand dt is realized, a cost-aware feasibility-preserving recourse routine chooses nonnegative warehouse-to-zone shipments xjzt and pooled external sourcing ot. This information order precludes using realized demand when selecting facility openings.

Active warehouses replenish stock subject to production and storage limits. Shipments are then assigned by ascending unit transport cost, never exceeding remaining stock or customer demand. Remaining shortfall is outsourced up to a pooled limit; any residual demand is lost. This fixed dispatch rule is intentionally transparent and need not be a globally optimal flow solution. The replenishment and dispatch decisions are continuous quantities coupled to the discrete activation vector. Stock at a closed site is retained but cannot be shipped until the site reopens.

$$A_{j,t} = \min \{P_j z_{j,t}, S_j - I_{j,t}\} ; \qquad I_{j,t+1} = I_{j,t} + A_{j,t} - \Sigma_z x_{jz,t} . \tag{1}$$

Each period incurs operation and reopening expenses, procurement, shipment, outsourcing, holding, and lost-sales charges. With $Q_t = \Sigma_z d_{z,t}$ and served quantities b_t, the observed stage cost is:

$$c_t = \Sigma_j [f_j z_{j,t} + o_j (z_{j,t} - z_{j,t-1})_+ + 2 A_{j,t} + 0.22 I_{j,t+1}] + \Sigma_{j,z} \tau_{jz} x_{jz,t} + 9 o_t + 24(Q_t - b_t). \tag{2}$$

The numeric coefficients in (2) are benchmark cost units, not estimates from an industrial company. External sourcing is restricted to 17 units per period; plant capacity is 32, 29, and 25 units and warehouse storage is 45, 43, and 38 units. A disruption reduces the second facility's available production capacity to 45% in periods 3–5. The controller knows this announced capacity schedule, while the simultaneous demand uplift is omitted from training.

3.2. Objective and Feasible Horizon Search

The stochastic optimization target is small mean total cost while controlling severe demand loss. The undiscounted episode cost is $C = \Sigma_{t=0}^{T-1} c_t$; fill rate is $\Sigma_t b_t / \Sigma_t Q_t$. For comparing action sequences locally, a discount factor $\gamma = 0.97$ only weights the two predicted stage costs. The first binary action in the selected sequence is implemented; continuous dispatch is computed after observing demand; and a new sequence is selected at the next period. Thus the policy is closed-loop even though each candidate sequence is evaluated open-loop.

4. UNCERTAINTY-AWARE THINKING-MODEL EVALUATION

4.1. Technical Adaptation and Model Training

The thinking model in Think4SCND separates a state predictor, a reward estimator and a decision evaluator, which refine the action initially proposed by its transformer decision network [8]. UA-TD implements the same three functional roles in a compact form: seven bootstrapped ridge regressors predict next-period zone demand, the exact stage-cost simulator replaces the learned reward estimator, and the decision evaluator rolls out candidate activation sequences. At the scale studied here all 8^H sequences can be enumerated, so exhaustive evaluation takes the place of the network's initial proposal; for larger networks, that proposal is what keeps the candidate set tractable (Section VII-C).

This preserves the prediction–evaluation–selection mechanism of the thinking model while making the dispatch constraints and risk terms directly inspectable; the transformer decision network and its policy training are not components of this experiment. Coordinated physics-informed multi-agent reinforcement learning for supply-chain optimization illustrates how risk-aware constraints and coupled network decisions can be evaluated within a structured optimization framework [27].

Each regressor receives previous four-zone demand, next-period seasonal baseline, sine and cosine seasonal indices, scaled period index, and the visible binary volatility flag. Training comprises 250 independent ten-period trajectories (2,500 labeled transitions), split equally between normal and volatile generators. Bootstrap sampling occurs by episode so within-trajectory dependence is preserved. An independent 140-trajectory validation collection supplies four-dimensional residual vectors conditioned on regime, maintaining contemporaneous cross-zone error correlation when disturbances are sampled.

Let $m_k(\cdot)$ denote bootstrap predictor k. To simulate one scenario path, an ensemble member is sampled and a previously observed four-zone validation residual is added at each imagined step. Its predicted demand becomes the autoregressive input for the next imagined step. The fixed regional cost matrix and known stock transition propagate the remaining state. These residual paths approximate aleatoric variation, whereas differences among bootstrap predictions approximate some estimation variation; neither mechanism yields a guarantee of coverage under unseen shifts.

$$\hat{d}_{t+h}^{(s)} = [m_{k(s,h)}(\varphi(\hat{d}_{t+h-1}^{(s)}, t+h)) + \varepsilon_{s,h}]_{+} . \tag{3}$$

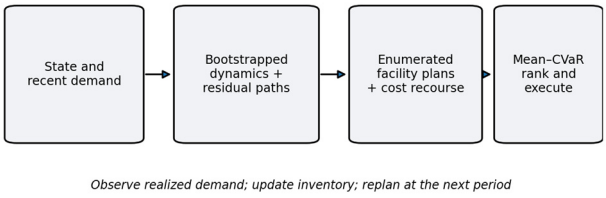


Fig 1. Implemented prediction–candidate evaluation–replanning flow. Every module is represented in the released code.

4.2. Candidate Risk Ranking

For $H = \min(2, T-t)$, all 8^H warehouse activation sequences are enumerated. Each sequence is evaluated against $S = 12$ common sampled demand paths, using the same stochastic realizations for all candidates within that decision. The realized cost of candidate a under scenario s is $V_s(a) = \Sigma_{h=0}^{H-1} \gamma^h \hat{c}_{t+h}^{(s)}$. Shared scenario draws reduce pairwise ranking noise. Sorting these 12 trajectory costs yields an empirical upper-tail CVaR; the highest $\text{ceil}((1-\alpha)S)$ outcomes enter the tail average, with $\alpha = 0.8$.

$$J_\beta(a) = (1 - \beta)\,(1/S)\,\Sigma_s V_s(a) + \beta\,\widehat{\text{CVaR}}_\alpha\,(V(a)); \quad \beta = 0.55. \tag{4}$$

$$a^* = \arg\min_{a \in \{0,1\}^{3H}} J_\beta(a); \qquad z_t = \text{first}(a^*). \tag{5}$$

Equation (4) is the technical modification. For $\beta \in [0,1]$, its scenario-functional is a convex combination of empirical expectation and empirical CVaR. Both are coherent on a fixed finite uniform scenario distribution, so their convex combination remains translation equivariant, positively homogeneous, monotone, and subadditive. This establishes a consistent risk ranking on those scenarios; it does not establish optimality or calibrated real-world loss bounds. The computational burden is bounded by 64 activation sequences and 12 simulated demand paths at $H = 2$, plus twelve shipment-arc passes per simulated period.

Algorithm 1. UA-TD Online Facility-Sequence Selection

Stage	Implemented computation
Input	Inventory I, previous mask z, last demand d, period t, regime flag
Predict	Sample S = 12 bootstrapped/residual demand paths of $H \leq 2$ steps
Propose	Enumerate each of the 8^H candidate binary facility sequences
Evaluate	Simulate capacity-feasible production, flows, external sourcing and cost
Rank	Compute sample mean and upper-tail CVaR; minimize $J\beta$ for $\beta = 0.55$
Act	Apply the first activation vector; reveal demand and dispatch
Update	Carry inventory and observed demand to period t + 1

4.3. Reproducible Inference Procedure

At each date the algorithm (1) constructs scenario paths from the previous observed demand; (2) enumerates candidate binary facility sequences; (3) propagates replenishment, shipment, outsourcing, and inventory for each candidate–scenario combination; (4) computes its mean and worst-tail cost; and (5) executes only the first facility vector of the minimum-score sequence. All operations use deterministic NumPy array functions given a scenario seed. The output is a planning policy informed by a fitted model; it is not trained using actor–critic gradients.

5. EXPERIMENTAL PROTOCOL

5.1. Synthetic Demand, Regimes, and Data Partition

A public transaction dataset does not itself supply counterfactual facility openings, storage limits, transport matrices, disruption schedules, and outsourcing prices. Consequently this study uses a controlled synthetic benchmark rather than assigning invented operational attributes to retail sales. The M5 competition documents the availability of real hierarchical demand records [23], but no M5 sales observations are used in the reported evaluation. This decision allows full auditability of intervention outcomes while limiting the strength of real-world generalization claims.

Zone baselines are (16,17,15,18). Demand follows a 0.56 autoregressive coefficient plus a 0.44 seasonal baseline, zone-specific Gaussian innovations, a shared Gaussian innovation, and occasional asymmetric bursts. Normal and volatile regimes differ in common-shock scale (4 versus 9), zone noise (2.8 versus 5), and burst probability (0.04 versus 0.10). The disruption regime also adds 12 and 11 units to zones 1 and 3 in periods 3–5 and activates the warehouse-2 capacity curtailment. All nonnegative clipping, random number calls, periods, and parameters are fixed in experiment.py.

Training trajectories use seeds 100000–100249; validation trajectories use 500000–500069 and 510000–510069. Development checks used separate early pilot trajectories. The untouched reported test sets have 180 independent episodes each, using 900000–900179 for normal, 910000–910179 for volatile, and 920000–920179 for disruption. Across policies, the same episode and demand realization are replayed; training and final testing have disjoint seeds. The validation one-step MAE is 4.35 units in normal and 8.72 in volatile conditions.

5.2. Baselines, Ablations, and Metrics

Six policies are run on every test episode: Always Open (all three facilities enabled); Reactive (one-step point-forecast plan); Nominal Lookahead (two-step point-forecast plan, i.e., a deterministic Think4SCND-style thinking-model evaluator [8]); Ensemble Expected (two-step 12-path mean-cost plan, $\beta = 0$); Risk Myopic (one-step 12-path mean–CVaR plan); and UA-TD (two-step 12-path mean–

CVaR plan). The latter two comparisons respectively isolate the risk term and the two-step candidate evaluator. Nominal Lookahead plays the evaluator's role in Think4SCND with a single predicted trajectory, so its comparison with UA-TD isolates the value of scenario and tail information inside the thinking model. These are common-environment baselines, not reproductions of other publications' full learning systems. PPO [24], for example, optimizes a learned policy and is therefore conceptually distinct from the enumerative planners compared here.

Table 1. Fixed Synthetic Benchmark Configuration

Setting	Experimental value
Topology and period	3 warehouses, 4 zones, 1 product, 10 periods
Production cap.	(32, 29, 25) units per warehouse/period
Storage cap.	(45, 43, 38) units
Operating costs	(110, 100, 90) per active site and period
Reopening costs	(40, 36, 30) per transition from off to on
Continuous costs	2 production; 0.22 holding; arc transport 1–2.5
External/loss	17-unit pooled limit; 9 sourcing; 24 unserved
Shifted capacity	Warehouse 2 at 45% capacity, periods 3–5

Outcomes include ten-period total cost, standard deviation of episode cost, empirical worst-decile average cost ($CVaR_{0.9}$), pooled-demand fill rate, average facilities activated, and decision time divided by ten. The 95% paired percentile bootstrap intervals use 3,000 resamples of the 180 within-seed differences for costs and 5,000 resamples for service, without asymptotic normality assumptions. Independent resampling is by episode. All reported evaluation outcomes come from saved CSV rows; figures are generated from the same files.

6. RESULTS AND ANALYSIS

6.1. Aggregate Cost and Service

Table 2. Held-Out Performance (180 Episodes Per Regime; Cost in Arbitrary Units; Nominal H2 = Deterministic Think4scnd-Style Evaluator)

Reg.	Policy	Mean cost	CVaR90	Fill %
Norm	Nominal H2	6085	10382	95.11
Norm	Ensemble H2	5974	10058	96.63
Norm	Risk H1	5996	10018	96.90
Norm	UA-TD H2	6014	9914	97.40
Vola	Nominal H2	8135	15084	86.55
Vola	Ensemble H2	7848	14643	90.28
Vola	Risk H1	7862	14687	90.66
Vola	UA-TD H2	7825	14622	91.13
Disr	Nominal H2	10593	19277	80.89
Disr	Ensemble H2	10266	19020	83.58
Disr	Risk H1	10272	19058	83.90
Disr	UA-TD H2	10246	19044	84.10

Table 2 provides the primary empirical evidence. In normal operation, UA-TD incurs mean cost 6014.3 versus 5973.6 for Ensemble Expected (a 0.68% increase), but improves fill rate from 96.63% to 97.40% and reduces worst-decile mean cost from 10058.0 to 9913.7. In volatile conditions, its mean cost is 7825.2 versus 7848.0 and its fill rate is 91.13% versus 90.28%. Under disruption, it records 10246.3 versus 10266.4 cost and 84.10% versus 83.58% fill. Thus risk-aware selection changes the realized service–cost balance more consistently than it reduces mean cost.

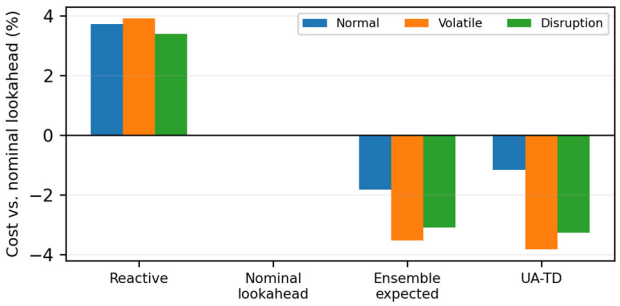


Fig 2. Realized mean cost relative to nominal two-step planning; negative percentages indicate savings.

6.2. Paired Uncertainty Intervals and Interpretation

Normal: mean service improvement over Ensemble Expected is 0.77 percentage points (paired 95% bootstrap interval 0.59 to 0.96 points).

Volatile: mean service improvement over Ensemble Expected is 0.85 percentage points (paired 95% bootstrap interval 0.62 to 1.08 points).

Disruption: mean service improvement over Ensemble Expected is 0.52 percentage points (paired 95% bootstrap interval 0.33 to 0.72 points).

Paired cost differences give a less uniform picture. The mean Ensemble Expected minus UA-TD costs are −40.7 units in normal operation (95% interval −75.2 to −4.5), +22.8 in volatile operation (−12.3 to +62.1), and +20.1 under disruption (−7.1 to +50.5). Therefore, the normal-regime cost premium is detectable under the paired bootstrap, while the small stress-regime cost differences are not sharply resolved. The strongest consistently observed signal is improved demand satisfaction rather than a claim of statistically established cost superiority.

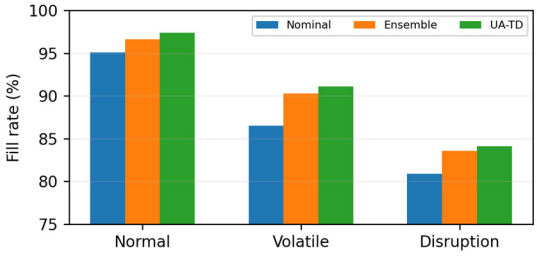


Fig 3. Observed fill rates from identical test episodes for three controlled planning variants.

Table 3. Paired UA-TD Minus Ensemble Expected Service Gains

Regime	Gain (pp)	95% CI (pp)
Normal	0.77	[0.59, 0.96]
Volatile	0.85	[0.62, 1.08]
Disruption	0.52	[0.33, 0.72]

6.3. Lookahead and Uncertainty Ablations

Removing both ensemble scenarios and risk sensitivity, i.e., returning to a deterministic Think4SCND-style thinking-model evaluator [8] (Nominal Lookahead), increases average total cost relative to UA-TD by 70.3, 310.1, and 346.5 units for normal, volatile, and disrupted test sets. The gain from uncertainty-aware evaluation is thus largest exactly where a single predicted trajectory is least representative, in the volatile and disrupted regimes. The corresponding paired 95% intervals are [15.1, 124.2], [220.1, 400.8], and [282.7, 412.8]. This comparison supports the joint value of scenario-aware evaluation over a deterministic future, rather than attributing all improvement to a single component.

Against Risk Myopic, adding the second planning period changes mean cost by +18.4 (normal), −36.9 (volatile), and −26.1 (disruption) units. Against Ensemble Expected, adding CVaR increases mean cost by 40.7 (normal) and decreases it by 22.8 and 20.1 units in the two stressed regimes. These mixed effects matter: a two-step risk planner may open extra capacity to avoid the simulated tail, but fixed and reopening costs are realized even when the adverse demand does not occur. The ablation thus identifies a measurable risk–expense interaction, not a universal advantage of longer horizons.

6.4. Feasibility, Operational Cost, and Limits

The shipment routine respects warehouse stock, customer demand, external sourcing bounds, and storage caps by construction. UA-TD enables an average 2.72, 2.93, and 2.95 warehouses per period under normal, volatile, and disruption conditions, respectively; this behavior explains its high fill rate and some of its operating-cost premium. Its recorded decision time is 1.85–2.27 ms per ten-period step on a single CPU, excluding the one-time fit. These figures measure a small enumerated network and should not be extrapolated to hundreds of candidate facilities, for which exhaustive 8^H search is infeasible.

The experiment isolates single-product, lost-sales dynamics, instantaneous replenishment, and a fixed deterministic greedy shipment allocator. It neither establishes multi-echelon optimality nor tests deployment on real fulfillment records. The exogenous demand process is learned, but warehouse inventory and reward transitions use known simulation rules; changes in those operational rules would require model recalibration or system identification. The announced capacity curtailment is known to the planner, whereas the additional disruption demand is not. Large-scale variants should sample or factor candidate configurations instead of enumerating all possibilities.

7. VALIDITY AND COMPUTATIONAL CONSIDERATIONS

7.1. Information Separation and Leakage Controls

An important requirement for sequential optimization experiments is that only information available before facility activation may influence activation. In the simulator, the action mask is selected from inventory, previous facility status, lagged demand, time, and the regime flag. The demand array used for scoring is produced by the learned predictor and residual sampler rather than supplied from the held-out episode. The actual demand vector is passed to the transition routine only after that action has been fixed. Shipment quantities can react to demand immediately, but the warehouse-opening decision cannot. The same distinction is applied to the deterministic and uncertainty-aware baselines. Since shipment recourse has the same dispatch ordering and the same outsourcing limit for every strategy, differences in observed performance arise from activation choices rather than changes in dispatch feasibility.

For each final test seed, the synthetic demand generator constructs one complete episode independently of the selected strategy. Thus comparison policies face identical demand paths, and

their outcome differences are naturally paired. Model fitting is performed once on training episodes and is not updated with any held-out future target. Scenarios used at decision time are regenerated deterministically from a seed based on the episode number and period. Each candidate within the same decision receives common forecast scenarios; otherwise an apparently better action could simply have benefited from a more favorable random sample. Evaluation metrics are computed from realized episode costs after all ten decisions, not from the planning score. This separation is necessary because a risk-sensitive planning functional and the actual realized expense measure different objects.

7.2. What the Uncertainty and Tail Estimates Represent

The ensemble contains seven episode-bootstrap ridge regressors. It does not represent seven independently discovered physical laws, and its between-model dispersion will typically be modest after fitting thousands of transitions. Much of the evaluated predictive dispersion consequently comes from joint historical residual innovations. Joint resampling matters because common demand shocks simultaneously pressure several customer zones; drawing each zone error independently would erase part of this competition for shared production and outsourcing capacity. The volatility flag selects either the normal or volatile residual pool. During the disruption test, the volatile pool is used while the structural demand uplift is deliberately absent from the training distribution. This is a useful controlled stress test, but its outcome should not be read as a numerical guarantee of out-of-distribution calibration.

With twelve forecast trajectories and confidence-tail level 0.8, each candidate tail estimate averages its three most expensive simulated trajectories. It therefore has nontrivial sampling variability, even with common random numbers. Raising the number of trajectories could stabilize the estimate at additional evaluation cost. Nor does the empirical worst-decile cost in Table 2 equal the score optimized online: the former uses the most expensive eighteen of 180 realized ten-period episodes, whereas the latter uses three expensive simulated two-period futures at each decision. A two-step risk objective should be understood as a local operational preference for avoiding severe shortages rather than an exact optimizer of ten-period CVaR.

7.3. Complexity and Scope of Application

For W candidate facilities and horizon H, exhaustive binary sequences number 2^(WH). With S forecast scenarios and a cost-aware dispatch pass visiting WZ warehouse-zone arcs per period, the evaluator takes O(2^(WH) S H WZ) simple allocation operations, in addition to prediction. In the tested case this means 64 candidate sequences, twelve scenarios, at most two steps and twelve transportation arcs. This bounded search avoids training an additional combinatorial actor but is the dominant scaling limitation: doubling the warehouse count from three to six while retaining H = 2 multiplies the candidate count by 64. Beam search, progressive widening, mixed-integer sequence selection, or regionwise decomposition would be natural alternatives for genuinely large network design problems.

Several validity boundaries follow from these design choices. Facility availability acts immediately within a period, so the benchmark does not cover construction lead times or irreversible site establishment. Procurement costs and warehouse capacities are fixed, and the setting aggregates products into a single interchangeable category; neither contractual supplier restrictions nor perishability are represented. The static shipping-price order is a useful feasible heuristic but can be dominated by global transportation optimization when capacity is tight. Furthermore, an announced disruption capacity schedule is available to all planners in the stress regime, which is more informative than an entirely unannounced outage. Generalization claims should therefore be restricted to the specified simulator and its held-out stochastic regimes until external network data and additional operational models are evaluated.

8. DISCUSSION

The operational implication is conditional rather than unconditional. For a service-sensitive operator, a scenario path evaluator offers a systematic method of spending more on capacity only when simulated downside risk warrants it. For a cost-only operator in comparatively stable demand, the risk-neutral ensemble can be preferable. The transparent score makes this choice explicit through β, and the realized fill/cost curve is interpretable by planners. Incorporating domain knowledge directly into an RL-related architecture has also proved useful elsewhere in inventory research [20]; this work similarly favors physical feasibility and inspectable cost accounting over opaque dispatch outputs.

Decision value is not reducible to forecast error alone. Prescriptive analytics emphasizes the optimization consequences of predictive information [21], while an empirical M5 inventory study identifies cases where prediction ranking and inventory cost ranking diverge [22]. Our separate reporting of demand-prediction MAE and downstream service/cost follows that distinction. The M5 accuracy competition [23] also underscores why an evaluation using real demand would be attractive, but real network metadata would still be required to make counterfactual activation comparisons meaningful.

For future research, the most important tests are capacity networks calibrated from public facility and transportation datasets, rolling forecasts on external real sales series, non-anticipative production lead times, and scalable mixed-integer or graph-guided candidate generation. Distribution shift may be handled through recalibration or robust uncertainty sets; at larger action dimensions, Think4SCND's learned decision network is the natural way to propose a tractable candidate set for the uncertainty-aware evaluator, and a joint comparison with its end-to-end training would quantify the benefit of scenario- and tail-aware evaluation at scale.

9. CONCLUSION

This paper presents and executes a risk-sensitive extension of thinking-model candidate-sequence planning for dynamic supply chain network design. The implemented method couples a bootstrapped demand state predictor, exact reward accounting, feasible shipping recourse, and CVaR-weighted two-period action evaluation. In 540 held-out synthetic episodes, it increases service levels relative to risk-neutral ensemble planning by 0.52–0.85 percentage points, with a detectable mean cost premium under normal demand and small, statistically unresolved mean cost savings in stressed regimes. The evidence favors treating uncertainty awareness as a configurable service–risk mechanism rather than a guarantee of lower cost. Reusable source code, seed partitions, per-episode results, confidence intervals, and plotting scripts are supplied for independent inspection.

REFERENCES

[1] Melo, M. T., Nickel, S., & Saldanha-da-Gama, F. (2009). Facility location and supply chain management—A review. European Journal of Operational Research, 196(2), 401–412. https://doi.org/10.1016/j.ejor.2008.05.007.

[2] Chen, J., Zhang, X., & Han, X. (2023). Explainable identification of tax-related financial reporting anomalies in U.S. public companies: Cross-fitted tax-consistency residuals and Owen-value Shapley propagation. World Journal of Economics and Business Research, 1(1), 33–45.

[3] Aldrighetti, R., Battini, D., Ivanov, D., & Zennaro, I. (2021). Costs of resilience and disruptions in supply chain network design models: A review and future research directions. International Journal of Production Economics, 235, 108103. https://doi.org/10.1016/j.ijpe.2021.108103.

[4] Ivanov, D., & Dolgui, A. (2021). A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0. Production Planning & Control, 32(9), 775–788. https://doi.org/10.1080/09537287.2020.1768450.

[5] 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.

[6] Boute, R. N., Gijsbrechts, J., van Jaarsveld, W., & Vanvuchelen, N. (2022). Deep reinforcement learning for inventory control: A roadmap. European Journal of Operational Research, 298(2), 401–412. https://doi.org/10.1016/j.ejor.2021.07.016.

[7] 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.

[8] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 12, 195974–195985. https://doi.org/10.1109/ACCESS.2024.3521439.

[9] Moerland, T. M., Broekens, J., Plaat, A., & Jonker, C. M. (2023). Model-based reinforcement learning: A survey. Foundations and Trends in Machine Learning, 16(1), 1–118. https://doi.org/10.1561/2200000086.

[10] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. Journal of Trends in Finance and Economics, 1(1), 85–100.

[11] Janner, M., Fu, J., Zhang, M., & Levine, S. (2019). When to trust your model: Model-based policy optimization. Advances in Neural Information Processing Systems, 32, 12519–12530. https://doi.org/10.48550/arXiv.1906.08253.

[12] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1612.01474.

[13] Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. Proceedings of the 35th International Conference on Machine Learning, PMLR, 80, 2796–2804. https://doi.org/10.48550/arXiv.1807.00263.

[14] Jetto, B., & Orsini, V. (2024). Resilient and robust management policy for multi-stage supply chains with perishable goods and inaccurate forecast information: A distributed model predictive control approach. Optimal Control Applications and Methods, 45(5), 2383–2414. https://doi.org/10.1002/oca.3162.

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

[16] Bertsimas, D., & Sim, M. (2004). The price of robustness. Operations Research, 52(1), 35–53. https://doi.org/10.1287/opre.1030.0065.

[17] Birge, J. R., & Louveaux, F. (2011). Introduction to stochastic programming (2nd ed.). Springer. https://doi.org/10.1007/978-1-4614-0237-4.

[18] Tamar, A., Glassner, Y., & Mannor, S. (2015). Optimizing the CVaR via sampling. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1). https://doi.org/10.1609/aaai.v29i1.9561.

[19] Teng, D. (2025). RSRBench: Reproducibility and performance benchmarking for regional seismic risk workflows. Academic Journal of Architecture and Civil Engineering, 3(1), 40–52.

[20] De Moor, B. J., Gijsbrechts, J., & Boute, R. N. (2022). Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management. European Journal of Operational Research, 301(2), 535–545. https://doi.org/10.1016/j.ejor.2021.10.045.

[21] Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. Management Science, 66(3), 1025–1044. https://doi.org/10.1287/mnsc.2018.3253.

[22] Theodorou, E., Spiliotis, E., & Assimakopoulos, V. (2025). Forecast accuracy and inventory performance: Insights on their relationship from the M5 competition data. European Journal of Operational Research, 322(2), 414–426. https://doi.org/10.1016/j.ejor.2024.12.033.

[23] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting, 38(4), 1346–1364. https://doi.org/10.1016/j.ijforecast.2021.11.013.

[24] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv:1707.06347. https://doi.org/10.48550/arXiv.1707.06347.

[25] Bian, M., et al. (2026). Assessing Green Public Supply Chains Resilience Under Carbon Neutrality Goals: A Multi-Agent Reinforcement Learning Framework. IEEE Access. https://doi.org/10.1109/ACCESS.2026.3720389.

[26] Zhang, X., Yang, Y., & Wang, Y. (2026). Risk-Aware Decision-Focused Demand Forecasting for Multi-Channel Inventory Optimization under Promotion and Stock-Out Constraints. Highlights in Business, Economics and Management, 69, 263–279.

[27] 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.