# Integrating AI with Bayesian Tracking in Supply Chain Management

**John Yata Raymond Lubari**[1], **Li Yongjun** [2], **Shuguang Zhang** [1], **Alladoumbaye Ngueilbaye**[3,4,5,6]
[1]Department of Statistics and Finance, School of Management, University of Science and Technology of China
[2]Department of Management Science, School of Management, University of Science and Technology of China
[3]School of Artificial Intelligence, Shenzhen University, Shenzhen 518060, China
[4]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University,Shenzhen 518060, China
[5]Big Data Institute, College of Computer Science and Software Engineering, Shenzhen University,Shenzhen 518060, China
[6]School of Industry and Urban Construction, Qingdao Hengxing University of Science and Technology, Qingdao, Shandong, China
johnyata@mail.ustc.edu.cn, lionli@ustc.edu.cn, sgzhang@ustc.edu.cn, angueilbaye@szu.edu.cn

**Abstract**

We study online replenishment under non-stationary demand, where decisions must be made sequentially from noisy demand-related signals while the underlying process can drift. We propose Artificial Intelligence Bayesian Tracking Supply Chain Management (ABT-SCM), which separates state estimation from ordering by combining Bayesian tracking of a latent demand driver (state-space model) with an optimistic exploration-exploitation rule over a discrete order grid. Each period, ABT-SCM updates a belief via Kalman filtering and chooses an order quantity by maximizing an upper-confidence objective computed from a learned reward surrogate. We compare ABT-SCM with Thompson Sampling, Sliding-Window Upper Confidence Bound, Discounted UCB, Sliding-Window Thompson Sampling, EXP3-IX, and a Random policy using cumulative reward and cumulative cost-regret relative to a discrete oracle. Over 30 replications, ABT-SCM delivers statistically significant gains on synthetic non-stationary data and remains robust to stronger drift, higher observation noise, heavy-tailed shocks, asymmetric holding or backorder costs, and alternative reward mappings. In multi-node supply-chain networks with chain, star, and Erdős–Rényi topologies, it consistently improves reward and reduces regret, suggesting robustness under stylized network scaling. On a real dataset evaluated via block bootstrap, ABT-SCM achieves the lowest mean cumulative cost-regret, with statistically significant regret improvement against EXP3-IX and directionally favorable but not consistently significant differences against the other baselines.

**Code** — https://github.com/JohnYata/ABT-SCM-code-for-supply-Chain_R-code

**Datasets** — https://www.kaggle.com/datasets/harshsingh2209/supply-chain-analysis

## Introduction

Recent work highlights the growing role of artificial Intelligence (AI) and machine learning (ML) in supply-chain digital transformation and decision support, including predictive analytics, smart and connected supply chains, and disruption-aware planning (Rana and Daultani 2023; Nozari, Szmelter-Jarosz, and Ghahremani-Nahr 2022; Camur, Ravi, and Saleh 2024). At the same time, supply-chain decisions are often made under partial observability: the true operational state, such as the underlying demand regime, lead-time regime, or in-transit pipeline condition, is not directly observed and must instead be inferred from noisy enterprise and sensor signals (Djennas, Benbouziane, and Djennas 2012). This naturally yields a partially observed sequential decision problem in which the posterior belief over the latent state is more informative than raw observations or point forecasts alone (Oroojlooy 2019). In this manuscript, we solve the deployable supply-chain decision support that faces three challenges. (i) The challenges of sensing and enterprise records provide only indirect and noisy proxies of the true state. (ii) Non-stationary environment, for instance, demand, supply, and disruption processes can drift over time because of seasonality, promotions, supplier shocks, and other regime changes. (iii) The most challenge is real operations impose hard KPI requirements, such as stock-out probability limits and fill-rate or OTIF targets, that must be respected while learning online (van der Laan et al. 2022). Prior supply-chain methods often optimize actions using point estimates or treat forecast errors as exogenous, which can be fragile under latent-state uncertainty and temporal drift. Meanwhile, non-stationary online learning methods model drifting reward environments (Deng et al. 2022), but they typically assume more direct feedback and do not explicitly study how state-tracking quality affects downstream operational decisions. Likewise, chance-constrained optimization addresses service-level control (van der Laan et al. 2022), but it is rarely integrated with Bayesian state tracking and online exploration–exploitation in a single closed-loop architecture. In this work, we propose ABT-SCM, a belief-based architecture that couples Bayesian state tracking with AI decision-making for non-stationary supply-chain control. Our contributions are as follows:

- **Belief-based Bayesian-tracking architecture for supply-chain control.** We formalize an end-to-end closed loop in which a Bayesian tracker updates a posterior belief over the latent system state, a belief-feature interface summarizes that posterior for decision-making, and an AI policy selects actions over a practical discrete

order grid.

- **Empirical evidence under non-stationarity and partial observability.** We evaluate ABT-SCM on synthetic non-stationary environments, robustness checks, multi-node network settings, and a real dataset, using cumulative reward and cumulative cost-regret relative to a discrete oracle.
- **Belief-feature interface with practical implementation detail.** In the present implementation, the posterior belief is updated via Kalman filtering and summarized by the posterior mean, yielding a lightweight and interpretable interface between state tracking and action selection.
- **KPI-aware action-selection framework.** We formalize how service-level requirements such as stock-out risk can be incorporated through belief-aware feasibility constraints, while clarifying that the current empirical study focuses on the unconstrained benchmark setting.

## Related Work

Supply-chain management involves sequential decisions in procurement, inventory, transportation, and service operations under uncertainty in industrial settings. The true operational state is often only partially observed through delayed, noisy, or incomplete signals, which makes belief-state reasoning important for non-stationary supply-chain control (Oroojlooy 2019; Lubari et al. 2026). AI and ML have been widely used for forecasting, inventory control, warehouse operations, predictive analytics, and smart supply-chain systems (Rana and Daultani 2023; Zapke 2019; Hanson-New and Daniel 2019; Stevanović et al. 2025). However, many AI-in-SCM pipelines still rely on point forecasts and do not explicitly propagate posterior uncertainty into downstream ordering decisions under non-stationarity.

Probabilistic and sequential-decision models provide a natural foundation for uncertainty-aware SCM. Bayesian networks and related probabilistic models have been used to support decision-making under uncertainty (Samanta, Chakraborty, and Jana 2024), while POMDP and MDP formulations treat the belief state as a sufficient statistic for control in partially observed systems (Cai et al. 2026). Bayesian filtering methods, including Kalman and particle filtering, enable online tracking of latent operational drivers from noisy signals (Seeger et al. 2017). Yet much of this literature emphasizes estimation accuracy rather than the downstream effect of tracking quality on replenishment performance under drift.

Our work is also connected to non-stationary online learning and bandit optimization. Standard multi-armed bandit methods address exploration–exploitation trade-offs in sequential decisions (Slivkins 2019; Lattimore and Szepesvári 2020; Chen, Golrezaei, and Bouneffouf 2023), while non-stationary variants use mechanisms such as sliding windows, discounting, temporal-dependence models, and change-point detection to adapt to drifting rewards (Cheung, Simchi-Levi, and Zhu 2022; Cavenaghi et al. 2021; Alami 2023; Deng et al. 2022; Zhang et al. 2023). Bandit methods have been successfully applied in high-frequency online domains such as recommendation and advertising (Aramayo, Schiappacasse, and Goic 2023; Silva et al. 2022; Madhawa and Murata 2019). However, these methods usually assume relatively direct reward feedback and do not explicitly integrate Bayesian tracking of latent supply-chain states.

Operational SCM also requires service-level and risk control. Chance-constrained and data-driven newsvendor models address targets such as stock-out probability and service levels (van der Laan et al. 2022), while learning-based approaches have been used for disruption-aware planning and multi-echelon supply-chain analysis (D'Souza 2021; Sultana et al. 2020; Camur, Ravi, and Saleh 2024; Schoepf, Foster, and Brintrup 2024). These gaps motivate the need for decision systems that connect sensing, state estimation, online adaptation, and operational KPIs. In contrast to prior work, ABT–SCM uses the posterior belief as the interface between noisy observations and online ordering decisions. This allows the framework to combine Bayesian state tracking, non-stationary exploration–exploitation, and KPI-aware action selection within a single closed-loop architecture (Sutton and Barto 2018; Agarwal, Aggarwal, and Azizzadenesheli 2022; Candelieri, Ponti, and Archetti 2023).

## Method

In this manuscript, we consider a discrete-time supply-chain system operating over periods $t = 1, 2, \ldots, T$. Let $x_t \in \mathbb{R}^{d_x}$ denote the latent system state such as echelon inventories, in-transit pipeline, or lead-time modes, $a_t \in \mathcal{A}$ the control action, and $y_t \in \mathbb{R}^{d_y}$ the observation obtained from enterprise and sensing systems such as RFID, GPS, and WMS. The supply-chain system evolves as a partially observed controlled Markov process:

$$x_{t+1} = f_t(x_t, a_t) + w_t, \quad w_t \sim P_t, \tag{1}$$

$$y_t = h_t(x_t) + v_t, \quad v_t \sim Q_t, \tag{2}$$

where $f_t$ and $h_t$ may vary over time to reflect non-stationarity, and $P_t$ and $Q_t$ denote the process and observation-noise laws. The decision maker observes the history available before selecting $a_t$:

$$H_t = \{y_{1:t}, a_{1:t-1}\}. \tag{3}$$

In the experimental setting used in this paper, the latent state $x_t$ is the mean demand level, realized demand $D_t$ is generated conditionally on $x_t$, and $y_t$ is a noisy observation of $x_t$. This makes the relationship between latent state, demand, and observation explicit before presenting the experiments.

### Bayesian Tracking (Belief Update)

The Bayesian tracking module maintains a posterior belief distribution over the latent state given the available history:

$$\pi_t(dx) \triangleq \mathcal{P}(x_t \in dx \mid H_t). \tag{4}$$

The filtering posterior is updated through the standard Bayesian recursion

$$\pi_t(dx) = \frac{\ell_t(y_t \mid x) \int K_t(x \mid x', a_{t-1}) \pi_{t-1}(dx')}{\int \ell_t(y_t \mid u) \int K_t(u \mid x', a_{t-1}) \pi_{t-1}(dx')\, du}, \tag{5}$$

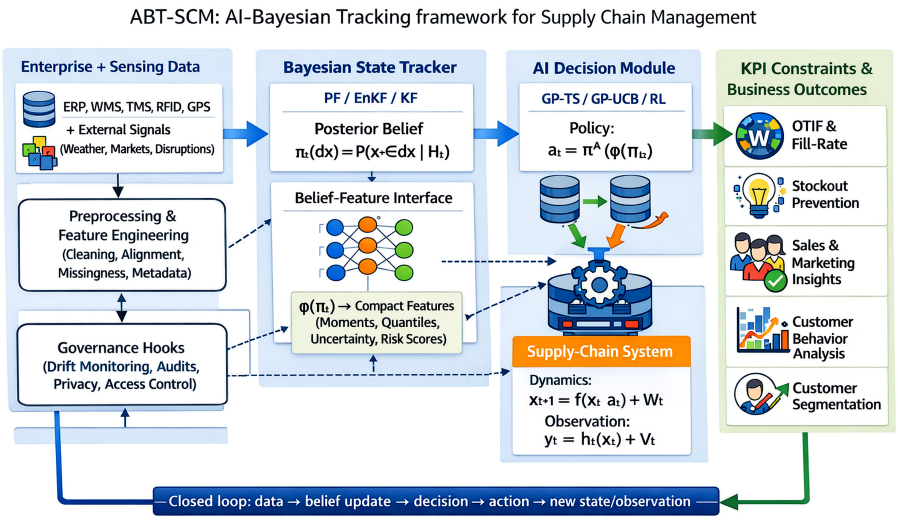


Figure 1: ABT-SCM Conceptual framework integration of Bayesian tracking and AI decision-making in supply chain management

where $K_t$ is the state-transition kernel and $\ell_t$ is the observation likelihood. In the experimental implementation, we instantiate this belief update using a Kalman filter, which provides the posterior mean $m_t$ and posterior variance $P_t$ of the latent demand state at each period.

## Belief-Feature Interface

Because the full belief $\pi_t$ may be high-dimensional, ABT-SCM maps the belief state to a compact feature representation

$$z_t = \phi(\pi_t). \tag{6}$$

In the present implementation, we use the posterior mean as the belief feature, so that

$$z_t = m_t. \tag{7}$$

This yields a lightweight and interpretable interface between tracking and decision-making. Richer summaries, such as posterior variance, quantiles, or risk scores, are natural extensions but are not the focus of the current empirical study.

## AI Decision Module

Conditioned on the belief feature $z_t$, the AI decision module selects an action using an exploration–exploitation policy, such as GP-Thompson sampling, GP-UCB, or reinforcement learning (Riquelme et al. 2018).

$$a_t = \pi^{AI}(z_t) = \pi^{AI}(\phi(\pi_t)). \tag{8}$$

Intuitively, the tracking layer answers the question: "what state are we in?" under uncertainty, while the decision layer answers "what should we do next?" by balancing learning and control. Therefore, in our experimental implementation, the historical tuples $(m_s, a_s, r_s)$ are used to fit a quadratic reward surrogate

$$\hat{r}_t(a \mid m_t) = \beta_0 + \beta_1 m_t + \beta_2 a + \beta_3 a^2, \tag{9}$$

and each candidate action is scored using an upper-confidence rule:

$$\mathrm{UCB}_t(a) = \hat{r}_t(a \mid m_t) + \sqrt{\beta_t}\,\hat{\sigma}_t(a \mid m_t) \tag{10}$$

where $\hat{\sigma}_t(a \mid m_t)$ is the prediction standard error and $\beta_t$ controls optimism, following the confidence-bound principle for exploration–exploitation (Srinivas et al. 2010).

## KPI-Aware Constraint Enforcement

Real deployments must satisfy service-level and risk constraints, such as OTIF, fill rate, or stock-out probability. ABT-SCM therefore includes a constraint-handling layer that translates the current belief state into probabilistic KPI proxies and restricts the action set accordingly. In structured replenishment settings, this can lead to tractable constrained optimization problems (van der Laan et al. 2022; Sui et al. 2015). To formalize service-level control, we define a stock-out at time $t$ as the event $\{D_t > a_t\}$ and impose the chance constraint

$$\mathcal{P}(D_t > a_t \mid \pi_t) \le \alpha \iff \mathcal{P}(D_t \le a_t \mid \pi_t) \ge 1-\alpha \tag{11}$$

where $\alpha \in (0,1)$ is a user-specified risk-tolerance parameter. This induces the belief-dependent feasible action set

$$\mathcal{A}_t^{\mathrm{feas}} = \{a \in \mathcal{A} : \mathcal{P}(D_t \le a \mid \pi_t) \ge 1-\alpha\}, \tag{12}$$

and the surrogate-UCB decision rule can then be applied over $\mathcal{A}_t^{\mathrm{feas}}$. In the present benchmark study, this KPI-aware formulation is included as part of the general ABT-SCM architecture, while the empirical comparisons focus primarily on the non-stationary unconstrained setting in order to isolate the value of belief-based tracking and adaptive decision-making. A dedicated stress test in which constraints become actively binding is left for future work. ABT-SCM operates

as a closed loop. After executing action $a_t$, the system transitions to $x_{t+1}$, generates a new observation $y_{t+1}$, updates the posterior belief to $\pi_{t+1}$, and repeats the cycle:

$$\begin{aligned}(y_t, a_{t-1}) \longrightarrow \pi_t \longrightarrow z_t = \phi(\pi_t) \longrightarrow a_t \\ \longrightarrow (x_{t+1}, y_{t+1}) \longrightarrow \pi_{t+1}.\end{aligned} \tag{13}$$

This integration allows the framework to address partial observability through Bayesian filtering, adapt to non-stationary through continual belief updates and online learning, and incorporate KPI requirements through belief-aware action restriction.

## Non-stationary and Variation Budget

To capture drift arising from seasonality, disruption, or evolving demand conditions, we allow the latent dynamics and cost landscape to change over time. In the present paper, the variation-budget discussion is used as a conceptual device for organizing non-stationary rather than as the basis of a complete finite-time theorem. Accordingly, we use it to motivate how temporal drift in the environment can affect both belief tracking and downstream decision performance.

$$V_T := \sum_{t=1}^{T-1} \sup_{x,a} \left\| \nabla_{(x,a)} c_{t+1}(x,a) - \nabla_{(x,a)} c_t(x,a) \right\| + \sum_{t=1}^{T-1} W_1(P_{t+1}, P_t) \tag{14}$$

where $W_1$ denotes the 1-Wasserstein distance and summarizes temporal changes in the latent transition environment, such as shifts in demand or lead-time regimes.

## Baseline Algorithms in the Non-stationary

All algorithms are evaluated in the same non-stationary single-item inventory environment, with the order quantity restricted to $\mathcal{A} = \{0, 2, \ldots, 40\}$. At each period, the learner observes realized demand $D_t$ and receives the bounded reward

$$r_t(a_t) = \frac{1}{1 + c(D_t, a_t)},$$

so that lower inventory cost corresponds to higher reward. We report cumulative cost-regret relative to the same discrete oracle

$$a_t^\star = \arg\min_{a \in \mathcal{A}} c(D_t, a), \qquad \Delta_t = c(D_t, a_t) - c(D_t, a_t^\star).$$

We compare ABT–SCM with Thompson Sampling (TS), Sliding-Window UCB (SW–UCB), Discounted UCB (D–UCB), Sliding-Window Thompson Sampling (SW–TS), EXP3–IX, and a uniformly random policy. TS follows the standard Beta–Bernoulli posterior update, SW–UCB and SW–TS use a window length $W = 50$, D–UCB applies geometric discounting, and EXP3–IX uses implicit exploration for adversarial feedback. All methods use the same action grid, reward mapping, statistical tests, and regret definition.

## Synthetic Environment and Experimental Protocol

To improve reproducibility, we now make the main experimental settings explicit. In the default synthetic environment, all methods are evaluated over the discrete action grid $\mathcal{A} = \{0, 2, \ldots, 40\}$ with horizon $T = 200$ and $n_{\text{rep}} = 30$ independent replications. The latent process is simulated with persistence parameter $\phi = 0.9$, initial mean $\mu_0 = 20$, state noise standard deviation $\sigma_{\text{state}} = 2$, demand noise standard deviation $\sigma_{\text{demand}} = 3$, observation noise standard deviation $\sigma_{\text{obs}} = 2$, drift magnitude 5, and drift period 50. Inventory cost is computed using the standard newsvendor form with holding and backlog penalties set to $h = b = 1$, and reward is defined using the bounded transformation $r_t(a_t) = \frac{1}{1+c_t(a_t)}$. All methods are evaluated in the same environment and are compared using cumulative reward and cumulative cost-regret relative to the same clairvoyant benchmark. We model demand through a latent state-space process and apply Bayesian filtering for online belief updates. The simulator generates (i) a latent demand driver, (ii) a noisy observation stream available to the learner, and (iii) realized demand used to compute costs and rewards. Let $x_t \in \mathcal{R}$ denote the latent demand state at time $t = 1, \ldots, T$. For the latent dynamics and non-stationary drift, let $x_t \in \mathbb{R}$ denote the latent demand state at time $t = 1, \ldots, T$. The latent process follows an AR(1) model with occasional drift shocks: Here, $\phi \in (0, 1)$ controls persistence in the latent demand process, while non-stationary is introduced through the drift term $d_t$.

$$x_1 \sim \mathcal{N}(\mu_0, \sigma^2_{\text{state}}), \tag{15}$$

$$x_t = \phi x_{t-1} + d_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2_{\text{state}}), \tag{16}$$

where $\phi \in (0, 1)$ controls persistence. Then, non-stationarity is injected through the drift term

$$d_t = \begin{cases} \zeta_t, & \text{if } t \bmod \texttt{drift\_period} = 0, \\ 0, & \text{otherwise,} \end{cases} \qquad \zeta_t \sim \mathcal{N}(0, \texttt{drift\_sd}^2). \tag{17}$$

Demand generation and observations, conditional on the latent state, demand is sampled and truncated to enforce non-negativity:

$$D_t = \max\{\eta_t, 0\}, \qquad \eta_t \sim \mathcal{N}(x_t, \sigma^2_{\text{demand}}). \tag{18}$$

The learner does not observe $x_t$ directly. Instead, it receives a noisy observation

$$y_t = x_t + \nu_t, \qquad \nu_t \sim \mathcal{N}(0, \sigma^2_{obs}). \tag{19}$$

For the action space, the decision-maker selects an order quantity from the discrete grid $\mathcal{A} = \{0, 2, 4, \ldots, 40\}$. We evaluate each decision using the standard newsvendor-style holding and backlog cost

$$c(D_t, a_t) = h(a_t - D_t)^+ + b(D_t - a_t)^+, \tag{20}$$

Therefore, the holding-cost coefficient is $(u)^+ = \max\{u, 0\}$, $h > 0$, and $b > 0$ is the backlog (shortage-cost coefficient). Costs are mapped to bounded rewards through

$$r_t = \frac{1}{1 + c(D_t, a_t)} \in (0, 1]. \tag{21}$$

To study partial observability more explicitly, we also impose missingness on the observation stream after a burn-in period. Let

$$m_t \sim \text{Bernoulli}(1-\rho), \qquad y_t \leftarrow \begin{cases} y_t, & m_t = 1, \\ \text{NA}, & m_t = 0. \end{cases} \tag{22}$$

When $y_t$ is missing, the tracking layer performs a prediction-only update without measurement correction, mimicking sensor or data-pipeline dropouts. For performance evaluation, we compute cumulative reward and cumulative regret in cost space:

$$a_t^\star = \arg\min_{a \in \mathcal{A}} c(D_t, a) \tag{23}$$

where $a_t^\star$ denotes the discrete oracle action, defined as the best grid action given the realized demand $D_t$. The instantaneous cost-regret is

$$\Delta_t = c(D_t, a_t) - c(D_t, a_t^\star), \tag{24}$$

and we report

$$R_T = \sum_{t=1}^{T} r_t, \qquad \text{Regret}_T = \sum_{t=1}^{T} \Delta_t. \tag{25}$$

**Replications and statistical comparison.** For pairwise comparisons between ABT–SCM and each baseline, we apply paired $t$-tests with non-parametric paired Wilcoxon tests. We compare ABT–SCM to standard bandit baselines adapted to discrete action grids, including Thompson Sampling (TS), Sliding-Window UCB (SW-UCB), Discounted UCB (D-UCB), Sliding-Window Thompson Sampling (SW-TS), EXP3-IX, and Random selection, all evaluated on the same simulated trajectories and cost/reward definitions above. **Multiple network sizes and topologies.** We report performance across size/topology to assess scalability and structural robustness.

We consider network sizes $N \in \{5, 10, 20\}$ and topologies such as chain, star, and Erdős–Rényi graphs, which connect the evaluation to multi-agent and networked bandit settings (Agarwal, Aggarwal, and Azizzadenesheli 2022; Madhawa and Murata 2019).

**Robustness to alternative modeling choices.** We further evaluate (i) heavy-tailed demand noise (Student-$t$ shocks), (ii) asymmetric holding/backorder penalties ($h \neq b$), and (iii) alternative bounded reward mappings (e.g., $r_t = \exp(-\lambda c_t)$).

## Synthetic Non-stationary Environment

Table 1 summarizes performance in the simulated non-stationary environment. ABT-SCM achieves the highest cumulative reward and the lowest cumulative cost-regret among the compared methods. The paired comparisons show statistically significant improvements over TS, SW–UCB, D–UCB, SW–TS, and EXP3–IX for both reward and regret.

## Robustness and Sensitivity

Figure 2 summarizes robustness checks under heavier-tailed demand shocks, asymmetric holding and backlog costs, and alternative bounded reward mappings. Across these settings, ABT-SCM maintains low cost-regret, supporting the interpretation that improved belief tracking translates into more reliable replenishment decisions under drift and sensing uncertainty.

## Network Scalability in Synthetic Multi-node Supply Chains

Across the multi-node experiments with $N \in \{5, 10, 20\}$ and chain, star, and Erdős–Rényi topologies, ABT–SCM consistently maintained strong cumulative reward and low cost-regret relative to the baselines. These results suggest that the belief-based tracking interface remains stable under stylized network scaling, although realistic coupling constraints such as shared suppliers, transportation capacity, and lead-time interactions remain important directions for future work.

## Real-world Dataset: Empirical Applicability via Block Bootstrap

We validate the empirical setting using the public Kaggle Supply Chain Analysis dataset, which contains haircare, skincare, and cosmetics supply-chain records. Observed demand-related variables are treated as noisy signals available to the learner, and costs and rewards are computed using the same newsvendor mapping as in the simulation study. Table 2 reports the real-world block-bootstrap results. ABT–SCM obtains the lowest mean cumulative cost-regret. Its regret improvement is statistically significant against EXP3–IX, while the regret differences against TS, SW–UCB, D–UCB, and SW–TS are favorable in direction but not statistically significant under the current evaluation. The reward differences depend on the bounded mapping $r_t = 1/(1+c_t)$, which compresses high-cost regions; therefore, cost-regret remains the most operationally interpretable outcome, as shown in Figures 4–5.

## Managerial and Practical Implications

ABT–SCM's lower cost-regret has a direct operational interpretation: it reduces avoidable holding and shortage penalties when demand drifts. The discrete order grid also matches common replenishment practice, where orders are often placed in case packs, pallets, or fixed increments. For analytics teams, the framework connects sensing quality, belief accuracy, and downstream ordering performance, making it useful for evaluating investments in ERP, WMS, IoT, and forecasting infrastructure.

The main limitations are as follows. First, actions are restricted to a discrete grid; continuous or capacity-constrained ordering requires further extension. Second, the empirical study uses one real dataset through block bootstrap, so broader multi-industry validation is needed. Third, the cost model is newsvendor-style and does not yet include lead times, perishability, or multi-period inventory dynamics. Finally, although the framework supports KPI-aware

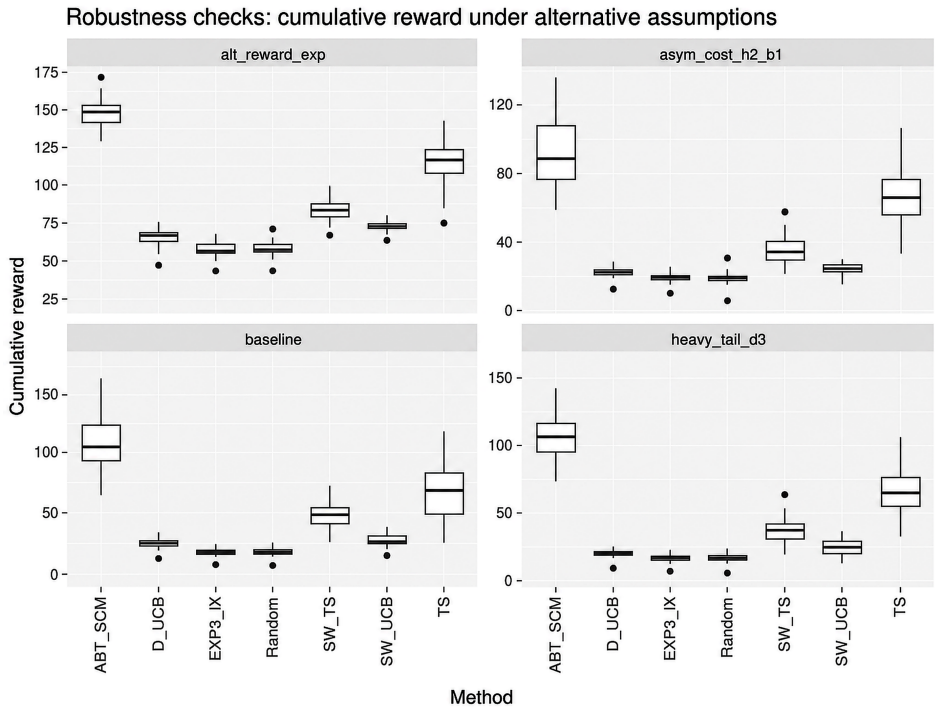


Figure 2: Robustness checks of algorithm performance under baseline, heavy-tailed demand shocks, asymmetric costs, and alternative reward assumptions in the synthetic dataset.

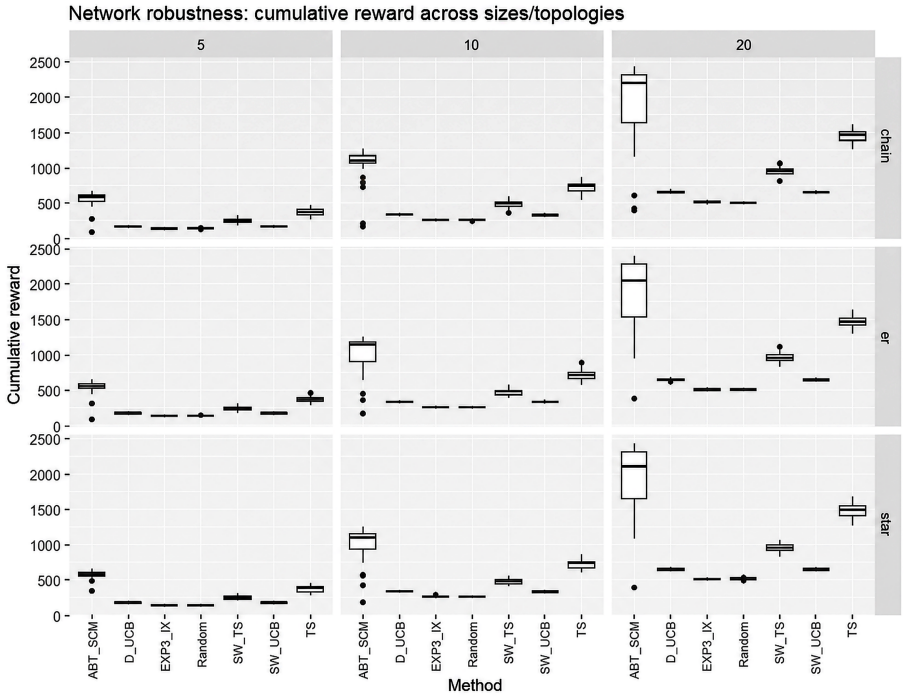


Figure 3: Network robustness in the multi-node synthesis supply-chain setting. Boxplots report cumulative reward aggregated across nodes for network sizes $N \in \{5, 10, 20\}$ and topologies including chain, star, and Erdős–Rényi networks over 30 replications.

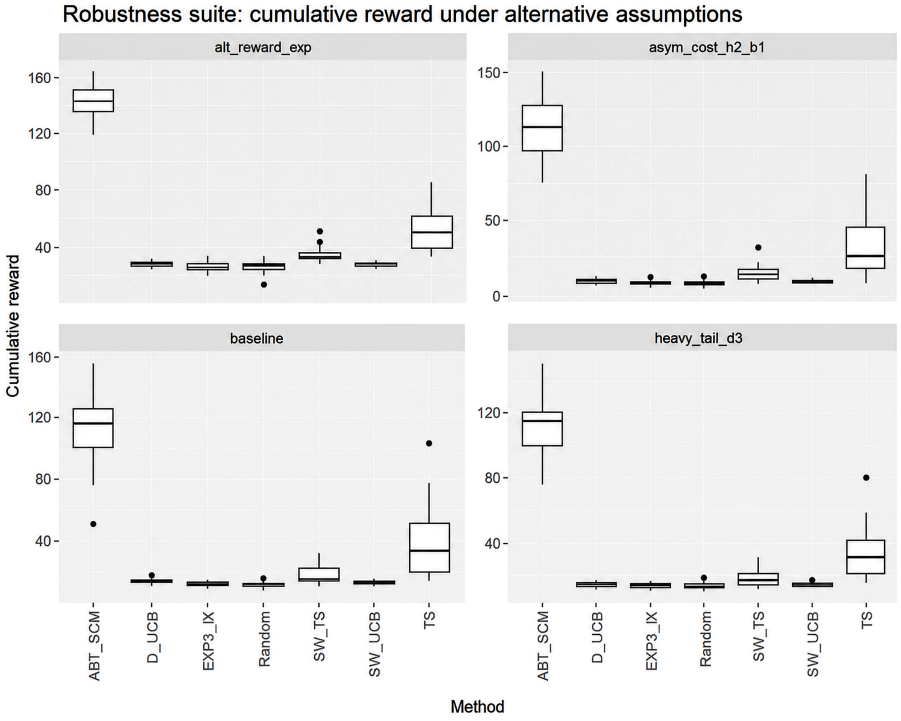


Figure 4: Robustness checks of algorithm performance under baseline, heavy-tailed demand shocks, asymmetric costs, and alternative reward assumptions in the real-world dataset.

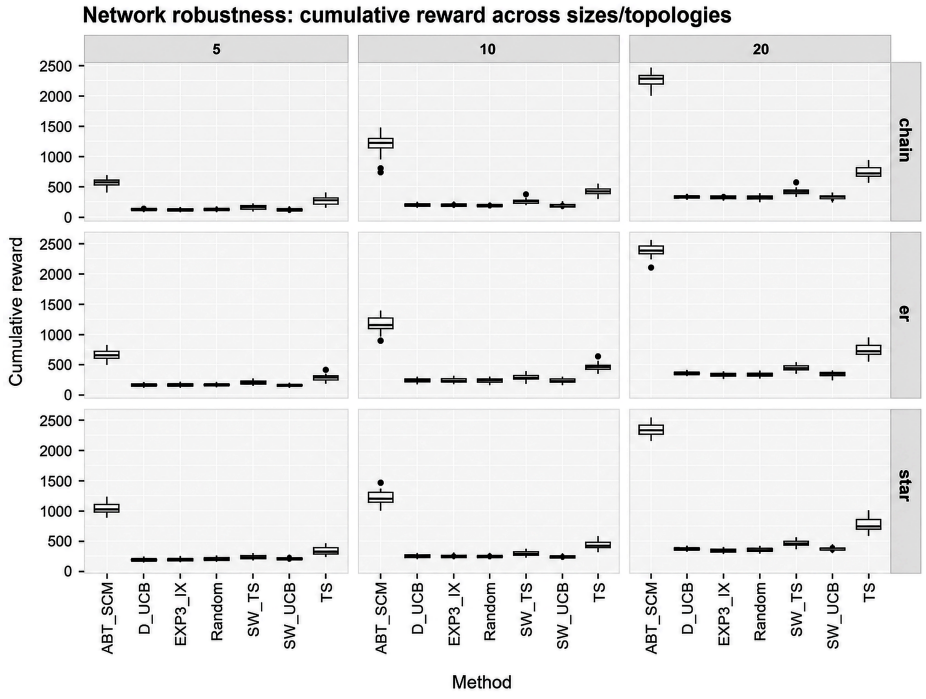


Figure 5: Network robustness in multi-node supply-chain Real-world Dataset. Boxplots report cumulative reward aggregated across nodes for network sizes $N \in \{5, 10, 20\}$ and topologies (chain, star, Erdős–Rényi), over $n_{\text{rep}} = 30$ replications

| Method | M Reward | SD Reward | M Regret | SD Regret | $\Delta$ Reward | $p$ Reward | $\Delta$ Regret | $p$ Regret |
|---|---|---|---|---|---|---|---|---|
| ABT–SCM | 108.0 | 16.10 | 841.0 | 177.0 | – | – | – | – |
| TS | 64.8 | 18.10 | 2033.0 | 318.0 | 48.26 | 0.0000 | -1079.82 | 0.0000 |
| SW–UCB | 31.6 | 2.88 | 3131.0 | 117.0 | 82.68 | 0.0000 | -2185.68 | 0.0000 |
| D–UCB | 32.7 | 3.39 | 3163.0 | 116.0 | 81.31 | 0.0000 | -2217.16 | 0.0000 |
| SW–TS | 44.4 | 11.10 | 2755.0 | 192.0 | 68.17 | 0.0000 | -1831.78 | 0.0000 |
| EXP3–IX | 26.6 | 2.65 | 3455.0 | 156.0 | 87.89 | 0.0000 | -2512.35 | 0.0000 |
| Random | 25.8 | 2.53 | 3451.0 | 196.0 | – | – | – | – |

Table 1: Synthetic dataset: summary statistics and paired comparisons against ABT–SCM over $n = 30$ replications. For regret, negative mean differences indicate lower regret for ABT–SCM.

| Method | M Reward | SD Reward | M Regret | SD Regret | $\Delta$ Reward | $p$ Reward | $\Delta$ Regret | $p$ Regret |
|---|---|---|---|---|---|---|---|---|
| ABT–SCM | 6.87 | 0.853 | 3039.0 | 217.0 | – | – | – | – |
| TS | 7.50 | 1.370 | 3113.0 | 237.0 | -0.6237 | 0.0316 | -74.1333 | 0.2240 |
| SW–UCB | 7.75 | 1.220 | 3113.0 | 196.0 | -0.8785 | 0.0067 | -73.8667 | 0.2229 |
| D–UCB | 7.88 | 1.150 | 3118.0 | 211.0 | -1.0070 | 0.0004 | -79.1333 | 0.2189 |
| SW–TS | 7.85 | 1.410 | 3137.0 | 256.0 | -0.9728 | 0.0004 | -97.7333 | 0.1265 |
| EXP3–IX | 7.58 | 1.320 | 3171.0 | 173.0 | -0.7083 | 0.0215 | -132.1333 | 0.0040 |
| Random | 7.72 | 1.330 | 3068.0 | 232.0 | – | – | – | – |

Table 2: Real-world dataset: summary statistics and paired comparisons against ABT–SCM over $n = 30$ block-bootstrap replications. Mean differences are computed as ABT–SCM minus baseline; for regret, negative values indicate lower regret for ABT–SCM.

feasibility constraints, the current experiments do not include a dedicated stress test where these constraints become actively binding.

## Conclusion

We proposed ABT–SCM as a belief-based framework for online replenishment under partial observability and non-stationary demand. The approach separates Bayesian tracking of latent demand drivers from exploration–exploitation decision-making over a practical discrete order grid. Across synthetic experiments, robustness checks, and stylized multi-node settings, ABT–SCM reduces cumulative cost-regret while remaining competitive in cumulative reward. On the real dataset, it achieves the lowest mean cumulative cost-regret, with statistically significant regret improvement against EXP3–IX and directionally favorable but not consistently significant differences against the other baselines.

Future work should extend ABT–SCM to continuous and constrained action spaces, richer latent-state models, lead-time and multi-echelon inventory dynamics, and deployment-oriented studies aligned with KPIs such as fill rate, stockout frequency, and working capital.

## References


Agarwal, M.; Aggarwal, V.; and Azizzadenesheli, K. 2022. Multi-agent Multi-armed Bandits with Limited Communication. *Journal of Machine Learning Research*, 23(212): 1–24.

Alami, R. 2023. Bayesian Change-point Detection for Bandit Feedback in Non-stationary Environments. In *Proceedings of the Asian Conference on Machine Learning*, 17–31. PMLR.

Aramayo, N.; Schiappacasse, M.; and Goic, M. 2023. A Multiarmed Bandit Approach for House Ads Recommendations. *Marketing Science*, 42(2): 271–292.

Cai, Y.; Lin, W.; Jing, C.; Liu, Z.; and Zheng, Z. 2026. FM-MDP: failure monitoring approach for DNN-based Markov decision process. *Empirical Software Engineering*, 31(2): 36.

Camur, M. C.; Ravi, S. K.; and Saleh, S. 2024. Enhancing Supply Chain Resilience: A Machine Learning Approach for Predicting Product Availability Dates Under Disruption. *Expert Systems with Applications*, 247: 123226.

Candelieri, A.; Ponti, A.; and Archetti, F. 2023. Uncertainty Quantification and Exploration–Exploitation Trade-off in Humans. *Journal of Ambient Intelligence and Humanized Computing*, 1–34.

Cavenaghi, E.; Sottocornola, G.; Stella, F.; and Zanker, M. 2021. Non-stationary Multi-Armed Bandit: Empirical Evaluation of a New Concept Drift-Aware Algorithm. *Entropy*, 23(3): 380.

Chen, Q.; Golrezaei, N.; and Bouneffouf, D. 2023. Non-stationary Bandits with Auto-regressive Temporal Dependency. In *Advances in Neural Information Processing Systems*, volume 36, 7895–7929. MIT Press.

Cheung, W. C.; Simchi-Levi, D.; and Zhu, R. 2022. Hedging the Drift: Learning to Optimize under Non-stationarity. *Management Science*, 68(3): 1696–1713.

Deng, Y.; Zhou, X.; Kim, B.; Tewari, A.; Gupta, A.; and Shroff, N. 2022. Weighted Gaussian Process Bandits for Non-stationary Environments. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, 6909–6932.

Djennas, M.; Benbouziane, M.; and Djennas, M. 2012. Agent-Based Modeling in Supply Chain Management: a Genetic Algorithm and Fuzzy Logic Approach. Preprint.

D'Souza, S. 2021. Implementing Reinforcement Learning Algorithms in Retail Supply Chains with OpenAI Gym Toolkit. arXiv preprint. ArXiv:2104.14398.

Hanson-New, C.; and Daniel, J. 2019. The Application of Big Data and AI in the Upstream Supply Chain. In *LRN Conference Proceedings*.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Lubari, J. Y. R.; Yongjun, L.; Zhang, S.; and Ngueilbaye, A. 2026. Dynamic ad selection via adaptive thompson sampling with Gaussian processes for non-stationary user behavior. *Knowledge-Based Systems*, 116119.

Madhawa, K.; and Murata, T. 2019. A multi-armed bandit approach for exploring partially observed networks. *Applied Network Science*, 4(1): 26.

Nozari, H.; Szmelter-Jarosz, A.; and Ghahremani-Nahr, J. 2022. Analysis of the Challenges of Artificial Intelligence of Things (AIoT) for the Smart Supply Chain (Case Study: FMCG Industries). *Sensors*, 22(8): 2931.

Oroojlooy, A. 2019. *Applications of Machine Learning in Supply Chains*. Ph.D. thesis, Lehigh University.

Rana, J.; and Daultani, Y. 2023. Mapping the Role and Impact of Artificial Intelligence and Machine Learning Applications in Supply Chain Digital Transformation: A Bibliometric Analysis. *Operations Management Research*, 16(4): 1641–1666.

Riquelme, C.; et al. 2018. Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling. In *International Conference on Learning Representations*, 1–15.

Samanta, S.; Chakraborty, D.; and Jana, D. K. 2024. Uncertain 4D-transportation problem with maximum profit and minimum carbon emission. *The Journal of Analysis*, 32(1): 471–508.

Schoepf, S.; Foster, J.; and Brintrup, A. 2024. Identifying Contributors to Supply Chain Outcomes in a Multiechelon Setting: A Decentralised Approach. *IEEE Transactions on Industrial Informatics*.

Seeger, M.; Rangapuram, S. S.; Wang, Y.; Salinas, D.; Gasthaus, J.; Januschowski, T.; and Flunkert, V. 2017. Approximate Bayesian Inference in Linear State Space Models for Intermittent Demand Forecasting at Scale. arXiv preprint. ArXiv:1709.07638.

Silva, N.; Werneck, H.; Silva, T.; Pereira, A. C.; and Rocha, L. 2022. Multi-armed Bandits in Recommendation Systems: A Survey of the State-of-the-art and Future Directions. *Expert Systems with Applications*, 197: 116669.

Slivkins, A. 2019. *Introduction to Multi-Armed Bandits*. Foundations and Trends in Machine Learning. Now Publishers.

Srinivas, N.; Krause, A.; Kakade, S. M.; and Seeger, M. 2010. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 1015–1022.

Stevanović, D.; Perić, V.; Roljević Nikolić, S.; Stefanović, V. M.; Oro, V.; Tabaković, M.; and Kolarić, L. 2025. Predictive Modelling of Maize Yield Under Different Crop Density Using a Machine Learning Approach. *Agriculture*, 15(20): 2138.

Sui, Y.; Gotovos, A.; Burdick, J. W.; and Krause, A. 2015. Safe Exploration for Optimization with Gaussian Processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, 997–1005.

Sultana, N. N.; Meisheri, H.; Baniwal, V.; Nath, S.; Ravindran, B.; and Khadilkar, H. 2020. Reinforcement Learning for Multi-Product Multi-Node Inventory Management in Supply Chains. arXiv preprint. ArXiv:2006.04037.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition.

van der Laan, N.; Teunter, R. H.; Romeijnders, W.; and Kilic, O. A. 2022. The Data-driven Newsvendor Problem: Achieving On-target Service-levels Using Distributionally Robust Chance-constrained Optimization. *International Journal of Production Economics*, 249: 108509.

Zapke, M. 2019. *Artificial Intelligence in Supply Chains*. Master's thesis, Universidade NOVA de Lisboa, Portugal.

Zhang, M. M.; Dumitrascu, B.; Williamson, S. A.; and Engelhardt, B. E. 2023. Sequential Gaussian Processes for Online Learning of Non-stationary Functions. *IEEE Transactions on Signal Processing*, 71: 1539–1550.