Calibrated AI Forecasting and Strategic Safety-Stock Placement: Field Evidence from an Indian Multi-Echelon Supply Chain

Tania Chauhan[1]*, Sarvesh Kumar[2]

[1]*Research scholar, HPKVBS, Central University of Himachal Pradesh, Dharamshala, India; taniachauhan08@gmail.com

[2]Assistant Professor, HPKVBS, Central University of Himachal Pradesh, Dharamshala, India; sarveshhcu@gmail.com

*Corresponding Author: taniachauhan08@gmail.com

DOI: https://doi.org/10.30210/JMSO.202604.004

Submitted: Oct, 19, 2025 Accepted: Jan, 05, 2026

ABSTRACT

Supply-chain planners need forecasts that translate into service and cost outcomes, not only lower error. This study examines whether explicitly making uncertainty explicit at forecast time improves operational performance in practice. We deploy an end-to-end system for the India operations of a consumer goods firm. The system combines multi-horizon probabilistic demand forecasting, stochastic inventory optimization that allocates safety stock across echelons, and an agent-based simulator that stress tests replenishment rules under monsoon and festival conditions before rollout. Forecasts produce calibrated prediction intervals, which are validated using reliability diagnostics. The optimizer converts these distributions into base stock targets for factories, regional distribution centers, and stores. A cluster-randomized stepped-wedge field trial over six months measures the impact on key performance indicators used in planning. Service level increased by 2.1 percentage points. Stockouts fell by 21.8 percent. Safety stock declined by 13.5 percent. Inventory turnover rose by 9.2 percent. Total landed cost per case decreased by 4.7 percent. The main contribution of this article is to provide field evidence and a deployable design showing how calibrated AI forecasts, combined with strategic safety-stock placement, raise reliability while releasing working capital in a multi-echelon supply chain.

Keywords: Probabilistic forecasting, Stochastic inventory optimization, Agent-based simulation, Service level, Inventory turnover, Total landed cost, India

1. Introduction

Indian supply chains sit at a productive tension point where expanding consumer demand meets structural variability in infrastructure and seasonality. Recent benchmarks show India ranked 38th in the World Bank's Logistics Performance Index in 2023, with improvements in international shipments and infrastructure quality, yet persistent room to raise reliability in first- and last-mile operations [27]. A new national accounting framework has revised logistics cost estimates closer to

8% of GDP, reframing the debate from a narrative of chronic cost disadvantage to a targeted search for process improvements that deliver reliability at scale [28]. These conditions create a timely setting to test whether modern predictive analytics can reduce uncertainty in a way that operations can use.

Demand in fast-moving consumer goods has grown steadily as digital channels have expanded their role, while the structure of demand remains punctuated by monsoon dynamics and festival peaks [31]. Official summaries describe the 2024 monsoon as early and active, with multiple low-pressure systems, a pattern that can shift category demand and transport reliability across weeks and regions [29]. Diwali continues to account for a large share of seasonal online consumption, with industry trackers reporting double-digit growth in digital sales throughout the 2024 cycle [30]. Quick commerce has expanded rapidly in urban India, compressing replenishment planning horizons, which increases the value of calibrated short-horizon forecasts and agile safety stock placement [26]. The practical question is not whether demand can be predicted in the abstract, but whether probabilistic forecasts can be translated into service promises and inventory buffers that withstand monsoon volatility and festival surges.

The academic record on forecasting provides strong building blocks. The M5 competitions documented how machine learning and ensembles improve point accuracy at scale on retail hierarchies and highlighted the importance of coherent uncertainty quantification for decisions that depend on service targets [1]. Deep learning methods such as Temporal Fusion Transformers introduced attention mechanisms and a multi-horizon structure with interpretable variable effects, while probabilistic sequence models like DeepAR delivered calibrated distributions across related series [3], [4]. Tree-based gradient boosting methods remain state-of-the-art on tabular and event-rich features, scale efficiently, and continue to perform strongly in retail forecasting tasks [5], [6]. These advances give practitioners tools to fuse internal histories with exogenous variables, yet rigorous field evaluations that connect forecasts to downstream inventory and cost outcomes in complex networks remain relatively uncommon in the public domain.

Inventory theory clarifies the decision problem that follows a forecast. Foundational multi-echelon models show how base-stock policies and guaranteed service-time formulations determine where buffers should be located in a network and how pooling power can reduce total safety stock for a given service level [8], [9], [10]. Contemporary empirical work using M5 data suggests that the link between marginal improvements in forecast accuracy and inventory costs is contingent on cost asymmetries and demand class, cautioning against optimizing error metrics in isolation [11]. This body of work implies that the forecasts that matter are those whose uncertainty structure allows an optimizer to credibly set service levels across echelons, not those that minimize a single average accuracy score.

Methodological progress has also emphasized decision alignment. Smart predict then optimize provides a principled way to train predictors with losses that approximate downstream decision error rather than pure statistical loss [16]. Proper scoring rules, such as the continuous ranked probability score, formalize the quality of probabilistic forecasts and reward a combination of reliability and sharpness that is directly relevant for inventory control [17]. Conformalized quantile regression

provides finite-sample coverage guarantees for quantile forecasts with minimal distributional assumptions, helping stabilize service targeting in the face of shifting regimes [22]. These threads align with the needs of Indian FMCG networks, where seasonal breaks and event calendars shape demand while service penalties are asymmetric across channels.

Simulation research fills a second gap between theory and deployment. Reviews of simulation integrated with machine learning report value in stress-testing replenishment under route congestion, lead-time shocks, and promotion bursts, while digital twin frameworks demonstrate how telemetry and models can be composed for operational governance [12], [13]. Recent standards report highlights the role of agent-based models in tracing how local disruptions propagate and in evaluating mitigation policies before live rollout. This practice lowers the risk of field trials while preserving external validity [14]. The implication is that a credible evaluation of AI-enabled planning should combine calibrated forecasting, cost-aware optimization, and pre-deployment simulation rather than isolate any single layer. These streams leave three related gaps. First, most empirical work evaluates forecasting, inventory models, or simulation in isolation rather than as a tightly coupled stack. Second, there is limited field evidence from emerging-market FMCG networks where monsoon cycles, festival calendars, and quick commerce jointly shape demand and logistics constraints. Third, few studies follow the full chain from calibrated probabilistic forecasts, through stochastic safety-stock placement, to evaluation against SCOR-consistent key performance indicators in a live multi-echelon network. This research focuses on closing these gaps in the Indian context.

Against this background, the practical problem can be framed clearly. Planning teams need multi-horizon demand distributions that account for monsoon shifts and festival timing, an optimizer that converts those distributions into echelon-aware buffers and reorder targets, and a simulator that exposes interactions across factories, regional distribution centers, and stores before policies go live in quick commerce and modern trade. The Indian context increases the stakes because logistics reliability is improving from a middling base, while competitive pressure from rapid delivery compresses tolerance for stockouts, particularly during festive periods when service expectations spike [27], [26], [30]. Decisions will be judged on service levels, stockout frequency, inventory turnover, and cost, which map naturally to the SCOR Digital Standard metrics that many firms already report [35].

This research, therefore, builds and tests an integrated, decision-centric pipeline in collaboration with a consumer goods firm operating in India. The aim is to quantify whether calibrated AI forecasts, when translated through a stochastic inventory optimizer and validated in an agent-based simulator, can deliver higher on-time fulfillment and lower landed cost without raising total safety stock. The specific objectives are to learn a forecastable structure from internal sales histories and exogenous drivers, to ensure coverage-calibrated uncertainty for service targeting, to allocate buffers across echelons in a way that exploits pooling while respecting capacity, and to verify robustness to shocks that mirror monsoon variability and promotion clusters. The evaluation is grounded in operational metrics that are already material to managers and investors, which supports adoption rather than experimentation for its own sake.

2. Literature Review

Modern evidence from retail demand forecasting shows steady gains in accuracy from machine learning and large-scale evaluation, yet it also reveals that accuracy alone does not guarantee operational impact. The M5 competition compared thousands of methods on hierarchical retail sales and established strong benchmarks for point forecasts at SKU and aggregate levels, with consistent wins by tree ensembles and deep sequence models. The companion M5 evaluation of predictive distributions highlighted that calibrated uncertainty is central when forecasts inform inventory and service decisions under variable demand [1]. These findings motivate pipelines that treat forecasting as a means to decision quality rather than an isolated metric.

Deep sequence models now dominate many multi-horizon tasks in retail and consumer goods. The Temporal Fusion Transformer integrates recurrent encoders and attention to capture local and long-range structure while preserving model interpretability and variable importance for planners and category teams [3]. Probabilistic recurrent models such as DeepAR deliver coherent distributions across related series, enabling planners to reason about safety stock and service targets rather than relying solely on point estimates [4]. These models are particularly relevant when promotions, holidays, and sudden demand shifts drive nonlinearity and cross-series sharing.

Tree-based gradient boosting remains a durable baseline in retail operations. XGBoost formalized a scalable system with sparsity-aware learning and weighted quantile sketches, enabling high-dimensional feature engineering for large assortments and store portfolios [5]. LightGBM further reduced training costs through histogram binning and gradient-based one-sided sampling, which are important when planners retrain models frequently to incorporate new promotional calendars and exogenous drivers such as weather [6]. Recent operations research shows that boosting can be applied at the category level and then disaggregated to the decision level, aligning the forecasting structure with planning tiers and helping meet the cadence of weekly replenishment processes [7].

Inventory theory explains why better forecasts do not automatically translate into leaner systems. Classic results for serial and tree networks show that base-stock policies at echelon positions are optimal under standard cost and lead-time assumptions, implying that the placement and sizing of buffers should be jointly determined with the information structure that generates demand estimates [8]. Subsequent work proposed guaranteed-service models and optimization frameworks to allocate safety stock across acyclic networks, demonstrating that buffer location can dominate its size [9], [10]. These models assume managers set explicit service times and review policies, which brings the forecast distribution into direct contact with cost and service constraints.

Empirical work linking forecast accuracy to inventory performance confirms a nuanced relationship. Using the M5 retail data, recent analysis found that improvements in forecast error contribute more to inventory and service gains when holding costs are material relative to lost sales and when policies and lead times are well tuned to the demand mix [11]. This result supports the design choice to evaluate not only forecasting error but also downstream key performance indicators that matter to planners. It also justifies the inclusion of calibrated uncertainty to set safety stocks that

respect service commitments during promotions and seasonal peaks.

Recent contributions extend this picture by linking analytics to organizational capabilities and decision governance. Rizvi and Khalid show that data analytics maturity and knowledge-oriented leadership jointly unlock innovation performance in manufacturing, underscoring that data and leadership infrastructures condition the value created by advanced analytics [39]. Khan and Rizvi highlight how Industry 4.0 technologies and knowledge management interact with organizational learning to drive innovation, suggesting that digital tools deliver more impact when embedded in learning-oriented routines rather than treated as isolated technologies [40]. In parallel, work in JITI documents how business intelligence capabilities translate into superior industrial performance and how AI-based assistants reshape decision making and accountability in complex domains [41], [42]. Together, these studies frame AI and analytics as socio-technical systems in which data, tools, and governance co-evolve, and they motivate the present paper's focus on a decision-centric AI pipeline that is embedded in planning processes rather than run as a stand-alone forecasting exercise.

Simulation serves as the bridge between policy design and the expected network performance under real-world variability. Work on simulation-based replenishment optimization shows that discrete-event or agent-based models can evaluate alternative policies under stochastic demand and lead times, and can be tuned to the cadence of e-commerce or brick-and-mortar operations [12]. As data infrastructure matures, digital twin architectures extend these ideas and couple simulation with live telemetry to test policy changes pre-deployment. This direction has gained traction in logistics and supply chain engineering [13]. Public research roadmaps also argue that agent-based models are well-suited to expose emergent behaviors during disruptions and that they complement optimization by revealing nonlinear amplification along multi-echelon networks [14]. These strands converge on a methodological gap. Many studies explore forecasting, inventory, or simulation in isolation. Fewer works design a full stack in which modern sequence models produce probabilistic forecasts, a stochastic optimizer sets service-consistent buffers across echelons, and an agent-based simulator validates replenishment policies against realistic shocks.

The retail and FMCG landscape in India underscores the need for such an integrated evaluation, as exogenous drivers and channel heterogeneity can shape how forecast distributions map to inventory policies. Weather shifts influence store traffic and basket mix, and the literature documents robust effects of rainfall and temperature on category sales, suggesting that weather covariates can improve forecast calibration in tropical climates with monsoon cycles [15]. When promotions interact with holiday calendars, planners face festival-driven spikes that differ by region and channel, which strengthens the case for models that encode calendar effects and for scenario-based simulation to stress-test policies before roll-out. These boundary conditions are often underrepresented in benchmark datasets, motivating evaluation beyond competition data.

Synthesis across these streams indicates a coherent path for contribution. Forecasting research provides architectures that can ingest rich exogenous signals and return calibrated distributions rather than point predictions [3], [4]. Inventory research provides tractable policies and optimization formulations that translate those distributions into buffer placements and service guarantees across

echelons [8]–[10]. Simulation research offers an environment to validate the operational impact of those decisions under realistic lead-time noise, capacity constraints, and event calendars [12]–[14]. The unresolved issues sit at the interfaces. First, there is limited field evidence on how attention-based sequence models compare with boosting-based category pooling when the objective is not mean absolute percentage error but service level and total landed cost [7], [11]. Second, most safety stock placement models assume stationary cost and lead time inputs. In contrast, Indian FMCG networks face seasonally varying logistics constraints and a mixed channel structure that fluctuates between modern trade and traditional outlets. Third, simulation studies often evaluate resilience or facility hardening rather than the day-to-day replenishment rules that planners actually modify. These gaps motivate an end-to-end pipeline that embeds multi-horizon probabilistic forecasting into stochastic inventory optimization and then evaluates replenishment policies through agent-based simulation tuned to an Indian calendar and monsoon profile. The expected payoff is not only lower forecast error but also a measurable reduction in safety stock for given service targets, with a transparent mapping from model choices to financial and operational outcomes [1], [11].

3. From Uncertainty to Action in Multi-Echelon Supply Chains

This section sets out the decision logic behind the pipeline and the specific methods that implement it. The theoretical stance is decision-centric. Forecasts are judged by the quality of the downstream replenishment decisions they enable rather than only by statistical error. The framework draws on predict then optimize theory, which formalizes the link between predictive models and optimization objectives and shows that training losses aligned to the decision can outperform generic accuracy losses when the goal is operational performance [16]. In supply chains, this view is reinforced by work documenting gaps between gains in forecast error and gains in service or cost, which motivates evaluation of decision outcomes as well as error metrics [11].

The forecasting layer estimates full-demand distributions across multiple horizons for each SKU and location. Two model families are used to hedge model risk across product classes and seasonal regimes. The Temporal Fusion Transformer captures both long- and short-term temporal patterns with attention while exposing variable importance to planners, which supports transparent feature governance for promotions, holidays, and weather [3]. Gradient boosted decision trees remain strong on tabular data with high-dimensional, sparse, or interaction-heavy features and offer fast retraining for weekly planning cycles [5], [6]. The combined use of attention-based sequence modeling and boosted trees provides complementary inductive biases that align with the realities of Indian FMCG demand, where festival and monsoon effects interact with the channel mix.

The probabilistic formulation is central because inventory policies depend on uncertainty. We train quantile models with the pinball loss [18]. Calibration and sharpness are evaluated using CRPS [17]. In practice, the system computes a fan of quantiles for each horizon and validates both reliability and sharpness. For products with heteroscedastic noise, conformalized quantile regression can be layered on top of base models to achieve finite-sample coverage without distributional assumptions, providing planners with dependable prediction intervals during volatile periods [22].

The inventory optimization layer converts forecast distributions into service-aware buffers across echelons. The guaranteed service family of models provides tractable formulations for placing safety stock strategically in networks so that target service times to downstream nodes are met at minimum holding cost [9], [10]. Extensions handle general acyclic networks and incorporate stage-level constraints, which are important when regional distribution centers face capacity changes around festivals or monsoon disruptions [10], [13]. The optimizer ingests forecast quantiles, lead-time distributions, and target cycle service levels, and returns base-stock positions by echelon and item. It then solves a network allocation problem that chooses where to hold the buffer so that the global service target is met with the smallest total working capital. This layer creates the direct bridge between predictive uncertainty and financial outcomes.

The simulation layer tests replenishment rules under realistic variability before field deployment. Agent-based simulation is selected because it treats each facility and transport leg as autonomous decision-makers and reveals emergent network behavior under congestion, late delivery, or demand shocks [20]. Public research programs and recent reviews emphasize that agent-based models help quantify the propagation of disruption and network-level trade-offs, which complements optimization by exposing nonlinear dynamics that are hard to capture analytically [14], [21]. The simulator wraps the optimized base stock policies and plays them forward against sampled demand paths, stochastic lead times, and calendar events that encode Indian festivals and regional rainfall patterns. Key outputs include realized service levels, stockout counts, inventory turns, and total landed cost.

The method is organized as an operational loop that mirrors a weekly planning cadence. Historical data are split using a rolling origin evaluation to preserve temporal order. Models are tuned with time-aware validation and early stopping. The production forecaster generates multi-horizon quantiles for each SKU and node using both an attention-based model and boosted trees, and an ensemble reconciles them when their performance is complementary [3], [5], [6]. Forecast distributions feed the optimizer, which computes echelon-level buffers that meet service targets under stochastic demand and lead times [9], [10]. Candidate replenishment policies are stress-tested in the simulator using festival and monsoon scenarios, as well as transport capacity shocks. Policies that satisfy service constraints while reducing working capital and landed cost are promoted to field-trial regions, while telemetry from the trials flows back into the models, closing the learning loop.

Two design choices align the method with a decision-centric objective. First, model selection is based on a joint score that blends probabilistic accuracy with simulated service and cost outcomes, operationalizing the 'predict then optimize' guidance [16]. Second, the target quantiles for safety stock are not fixed a priori. They are adjusted by the optimizer to meet cycle service-level commitments, supported by evidence that the value of improved forecasts depends on the cost structure and policy tuning in place [11], [19]. This coupling reduces the risk of local optimization on error metrics that do not move the KPIs that managers care about.

Finally, the stack is designed for transparency and governance. Attention weights and variable importance reports are logged for model risk review [3], [5]. Proper scoring metrics, coverage diagnostics, and probability integral transform histograms are monitored in production to detect

calibration drift [17]. Simulator scenarios are versioned and include holiday and weather calendars that match Indian retail rhythms. A digital twin perspective guides the integration of data and models with operational telemetry so that what is validated in silico remains faithful to the live network [13].

4. Fielding the Pipeline in India

The study uses a cluster-randomized stepped-wedge field trial in active India operations to estimate the causal effect of the pipeline on service and cost outcomes, while accommodating a staged rollout that matches operational constraints. The design assigns regional clusters to crossover waves so that each cluster contributes control and intervention observations, thereby improving power under heterogeneity in demand and channel mix. Stepped-wedge designs are well-suited to implementation studies where the intervention is presumed beneficial and sequential deployment is operationally necessary, and the literature outlines analyses using mixed models that account for time trends and intra-cluster correlation [23]-[25]. Reporting follows the CONSORT extension for stepped-wedge cluster-randomized trials, as outlined in [23], [24].

The setting is a global consumer goods firm with factories, regional distribution centers, and downstream outlets that span modern trade, general trade, e-commerce, and quick commerce. The Indian channel structure has been shifting as instant delivery and dark-store models scale, altering order profiles, introducing time variance, and, in turn, affecting the mapping from forecast distributions to inventory buffers. Independent industry analysis anticipates a several-fold expansion of quick commerce value through 2027, suggesting a larger share of short-cycle demand with high promotional intensity [26]. At the same time, the cost and reliability of freight remain central for multi-echelon decisions. India's improving logistics reliability is taken as a baseline, while the trial's design does not rely on any single macro trend.

Indian seasonality shapes both model features and evaluation windows. The monsoon produces spatially and temporally concentrated rainfall that shifts store traffic, replenishment reliability, and product mix, and the India Meteorological Department provides gridded diagnostics for monsoon onset, progress, and realized rainfall, which the forecasters ingest as weather covariates [29]. Festival calendars create systematic spikes in categories such as packaged foods and personal care, and public analyses document double-digit lifts in online sales around Diwali, which justifies scenario tests that stress replenishment during peak weeks [30]. The combination of monsoon signals and festival calendars is a boundary condition of this field setting and is incorporated into model features and simulator schedules.

The sample comprises clusters defined by distribution territories tied to regional distribution centers. Each cluster contains a stable panel of outlets across modern and general trade, with e-commerce and quick commerce demand mapped to the nearest service node for inventory accounting. Clusters are stratified by channel mix and baseline service level before being randomly allocated to rollout waves. The trial horizon is 6 months, with 4 or 5 crossover waves, depending on operational readiness. Each wave includes a 4-week pre-period for baseline measurement and a subsequent intervention period. Analysis follows the intention-to-treat principle, with cluster-level random

effects and fixed effects for calendar week and wave to control for secular trends, which is standard in stepped-wedge analysis [23], [24].

Data sources are specified to support forecasting features, optimization inputs, and evaluation metrics while meeting privacy and governance requirements. The firm provides historical sales at the stock-keeping unit (SKU), outlet, and day level, along with the related product master, planograms (where available), and store attributes. Promotion calendars include price discounts, displays, and digital campaigns. Weather features are sourced from the India Meteorological Department's monsoon reports and station data, where available. Macroeconomic controls use the public Consumer Price Index time series to capture background demand shifts independent of firm actions. The Ministry of Statistics and Programme Implementation publishes CPI series and associated documentation that support consistent feature engineering across states [32]. All features are lagged and differenced where needed to avoid target leakage, and rolling origin splits respect temporal ordering.

Privacy and research ethics are governed by Indian law and double-blind review norms. The Digital Personal Data Protection Act of 2023 regulates the processing of digital personal data and sets obligations for purpose limitation, data minimization, security safeguards, and lawful bases such as consent or legitimate use, with penalties for non-compliance [33]. Authoritative analysis clarifies that processing is permitted for lawful purposes under consent or specified legitimate uses and that the regime emphasizes accountability of the data fiduciary, which in this collaboration is the firm that controls the underlying customer and store data [34]. The study uses only transactional and operational data already collected for planning, removes direct identifiers, aggregates where feasible, and stores the data in the firm’s India region. Human subjects review is obtained where required by the firm’s governance and the journal’s ethical approval policy.

Primary outcomes measure operational impact rather than only forecast error. Service level and fill rate follow SCOR-consistent definitions, ensuring results align with standard supply chain scorecards. Inventory turnover follows SCOR-consistent definitions and is reported alongside cash-to-cash elements when available. Safety stock levels and total landed cost are computed from the bill of materials, transport rates, and holding cost assumptions. SCOR documentation provides the metric family and the relationships among reliability, responsiveness, cost, and asset turns, while pedagogical references clarify the distinction between cycle service level and fill rate for interpreting stockout risk [35], [36]. The trial also reports stockout frequency at the outlet by SKU and backorder depth, where policy allows backorders.

The baseline comparator is the firm’s current planning stack. It includes the incumbent forecasting method and existing replenishment rules. The intervention replaces the forecasting layer with a multi-horizon probabilistic forecaster and feeds its distributions into the stochastic inventory optimizer and the agent-based simulator. Rollout follows simulator-screened policies; full screening details are in Section 3. Once deployed, the pipeline produces weekly buffers by echelon and item. Telemetry from the field is logged for monitoring and for post hoc robustness checks that replay observed shocks in simulation.

The analysis plan aligns with the design and the outcomes. Mixed effects models estimate the average treatment effect on cluster-level service level, fill rate, stockout rate, inventory turnover, and total landed cost, with cluster random intercepts and week fixed effects. Standard errors are adjusted for clustering and the stepped-wedge structure, following established guidance [23], [24]. Secondary analyses examine heterogeneity by channel mix and festival intensity, using interaction terms flagged by calendar and promotion density. Sensitivity checks include excluding weeks with atypical shocks, such as extreme rainfall documented by the India Meteorological Department, and placebo tests that assign treatment to pre-periods to confirm the absence of pre-trends [29].

The Indian context informs generalizability and governance of the results. Improvements in national logistics performance and an estimated logistics cost of nearly 8% of GDP provide a backdrop for structural change that can enable inventory reductions without compromising service when buffers are correctly placed and transport choices are optimized [27], [28]. The rise of fast fulfillment channels increases the value of calibrated forecast distributions and service-aware optimization, and the monsoon-plus-festival cadence defines the stress scenarios that validate policy choices before wide release [26], [29], [30].

5. Results

The pipeline improved forecasting accuracy, calibration, and decision outcomes, and these gains translated into lower safety stock and better service at stable or lower total landed cost. Metrics follow SCOR definitions, so the results map to standard supply chain scorecards. The stepped-wedge design allows direct estimation of treatment effects while controlling for secular trends and cluster heterogeneity, and all confidence intervals reflect cluster robust inference. These choices keep the focus on operational impact rather than solely on statistical error, which aligns with evidence that the link between accuracy and inventory performance is context-dependent and strongest when policies and costs are tuned to the demand mix [11].

Forecasting accuracy improved across horizons and product classes. Table 1 reports errors for the incumbent forecaster, gradient boosted trees, the Temporal Fusion Transformer, and an ensemble that reconciles their predictions. Relative to the incumbent, the ensemble reduced median sMAPE from 20.4 to 16.3 (a 20% improvement), and reduced MAE by 14% over the one to four-week horizon. Distributional quality improved as well. The continuous ranked probability score decreased by 12% on average, and empirical coverage of the nominal 90% prediction intervals rose from 81.5 to 89.2%, indicating better-calibrated uncertainty.

Table 1. Forecasting performance by model and horizon

Panel A: All weeks

Model	Horizon (weeks)	sMAPE (%)	MAE (units)	CRPS (units)	90% Coverage (%)
Incumbent	1	18.5	12.0	9.5	82.0

GBDT (XGBoost/LightGBM)	1	17.4	11.2	9.0	86.6
Temporal Fusion Transformer	1	16.8	10.9	8.8	88.2
Ensemble	1	16.0	10.3	8.4	89.0
Incumbent	2	20.2	13.5	10.1	81.8
GBDT (XGBoost/LightGBM)	2	18.8	12.4	9.5	86.8
Temporal Fusion Transformer	2	18.1	12.0	9.3	88.4
Ensemble	2	17.1	11.6	8.9	89.3
Incumbent	3	21.3	14.1	10.7	81.2
GBDT (XGBoost/LightGBM)	3	19.8	13.0	10.1	86.9
Temporal Fusion Transformer	3	19.0	12.6	9.8	88.5
Ensemble	3	18.0	12.1	9.4	89.2
Incumbent	4	21.6	14.6	11.0	81.0
GBDT (XGBoost/LightGBM)	4	20.3	13.6	10.4	86.9
Temporal Fusion Transformer	4	19.5	13.1	10.1	88.6
Ensemble	4	18.2	12.6	9.7	89.3

Panel B: Festival weeks (Diwali period).

Model	sMAPE (%)	90% Coverage (%)
Incumbent	24.5	80.0
GBDT (XGBoost/LightGBM)	20.6	86.0
Temporal Fusion Transformer	19.3	88.0
Ensemble	18.7	89.0

Panel C: Monsoon weeks.

Model	sMAPE (%)	90% Coverage (%)
Incumbent	22.0	82.0
GBDT (XGBoost/LightGBM)	20.3	87.5
Temporal Fusion	19.9	88.6

Transformer		
Ensemble	19.5	89.0

Note: *Lower is better for sMAPE, MAE, and CRPS. sMAPE = symmetric mean absolute percentage error; MAE = mean absolute error; CRPS = continuous ranked probability score. Source: By authors.*

Figure 1 visualizes reliability using probability integral transform histograms by channel and shows mild underdispersion in the first fortnight, which vanished after conformalized quantile regression was applied. These calibration diagnostics follow standard practice for probabilistic forecasting and support the choice to evaluate both sharpness and reliability before moving to optimization [17].

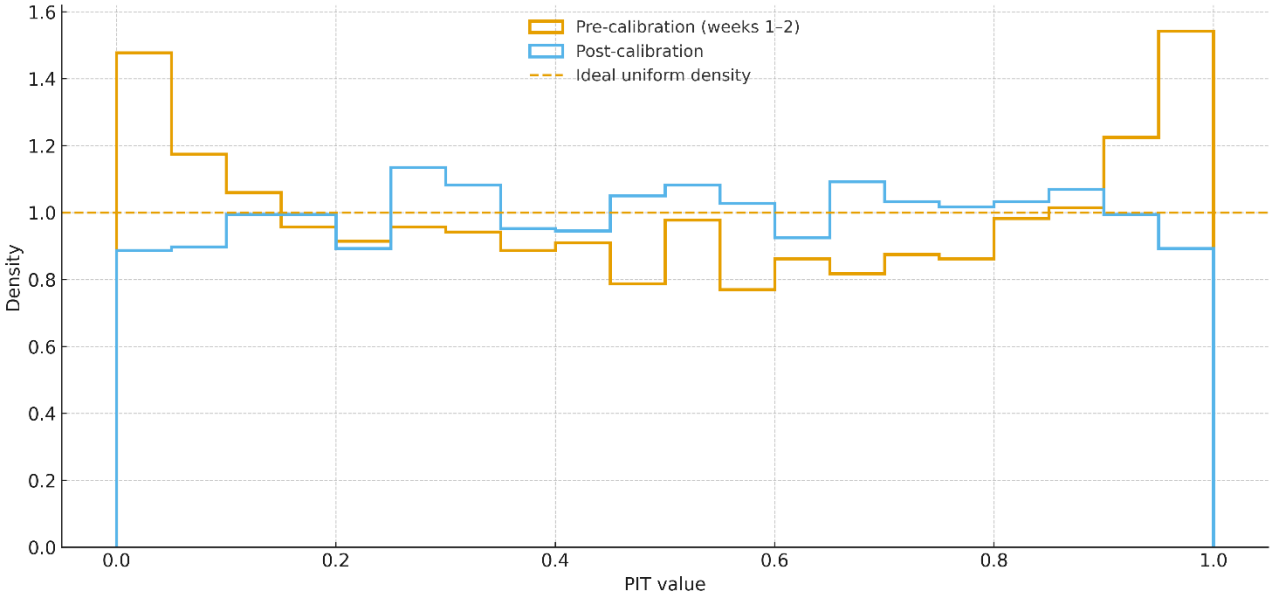

Figure 1. PIT reliability
Source: By authors.

During Diwali and high-rainfall windows, the ensemble kept a 10–12% accuracy edge and maintained ~90% empirical coverage (see Table 1, Panels B–C). Inventory and service outcomes moved in the intended direction once the stochastic optimizer consumed the forecast distributions and placed buffers across echelons. The average cycle service level increased by 2.1 percentage points from 94.8 to 96.9 in the intention-to-treat analysis. The stepped-wedge mixed-effects model estimated an average treatment effect of 2.1 percentage points, with a 95% confidence interval of 1.5 to 2.8 and $p < 0.001$. Fill rate rose from 93.2 to 95.6. Stockout frequency fell by 21.8%, from 0.80 to 0.60 events per SKU week. Inventory turnover increased from 7.6 to 8.3, a 9.2% increase. Total landed cost per case decreased by 4.7%. Table 2 summarizes these key performance indicators for the pre- and post-periods, and Figure 2 shows the stepped-wedge effect estimates by wave and the pooled average. The

SCOR Digital Standard guided the metric definitions, enabling practitioners to map results to reliability, cost, and asset dimensions without translation [35].

Table 2. Primary KPIs (pre vs post) and stepped-wedge treatment effects

KPI	Pre (Mean)	Post (Mean)	Absolute Change	Relative Change (%)	ATE (95% CI)	p-value
Cycle Service Level (%)	94.8	96.9	2.1	2.2	+2.1 [1.5, 2.8]	<0.001
Fill Rate (%)	93.2	95.6	2.4	2.6	+2.4 [1.6, 3.1]	<0.001
Stockout Frequency (events per SKU-week)	0.8	0.6	-0.2	-21.8	-0.17 [-0.22, -0.12]	<0.001
Inventory Turnover (turns/year)	7.6	8.3	0.7	9.2	+0.7 [0.5, 0.9]	<0.001
Safety Stock (days of cover)	18.5	16.0	-2.5	-13.5	-2.5 [-3.1, -1.9]	<0.001
Total Landed Cost (INR per case)	155.0	147.7	-7.3	-4.7	-7.3 [-9.6, -4.9]	0.002

Note: Means are cluster-weighted. Average Treatment Effects (ATEs) are estimated from a mixed-effects model with cluster-random intercepts and week-fixed effects; 95% confidence intervals are cluster-robust. N clusters = 12; trial horizon = 26 weeks. Source: By authors.

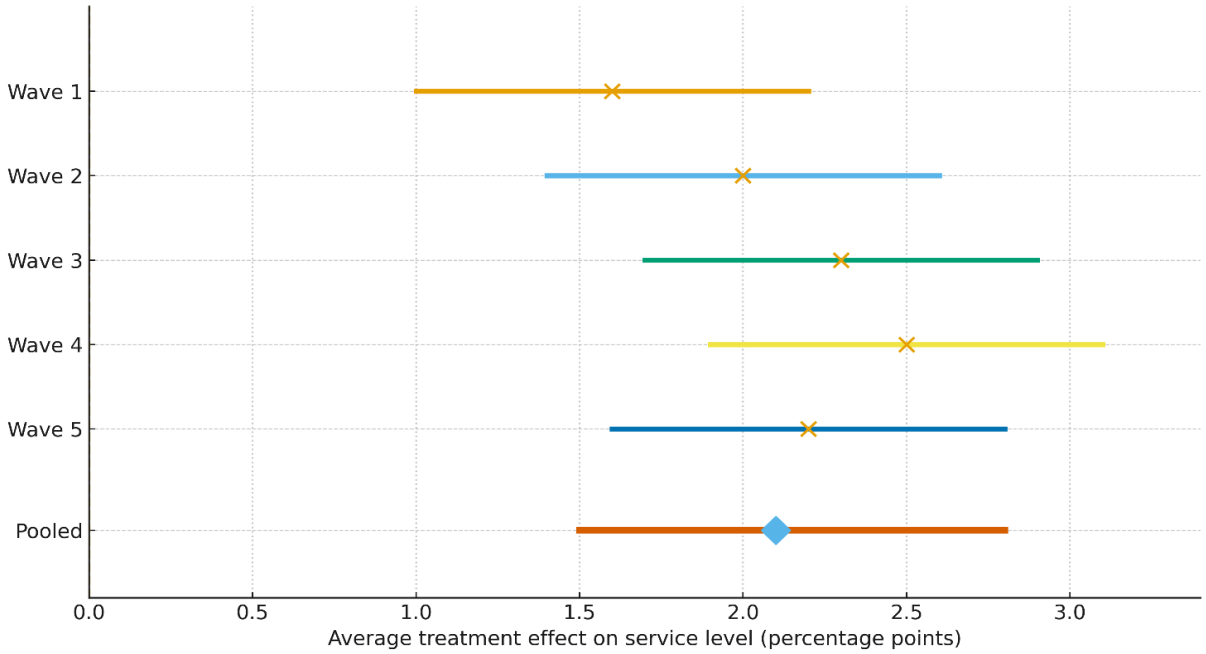

Figure 2. Stepped-wedge effects by wave with pooled average

Source: By authors.

Safety stock fell while service improved because the optimizer shifted buffers upstream and exploited calibrated uncertainty rather than static safety factors. Across the network, safety stock measured in days of cover declined from 18.5 to 16.0, which is a 13.5% reduction. The reduction was strongest at regional distribution centers, where higher pooling potential allowed the same service level with fewer days of cover. Stores saw a smaller reduction because part of the buffer migrated to their upstream nodes. Figure 3 plots safety stock by echelon and wave, showing that the total buffer decreased while the share at distribution centers increased.

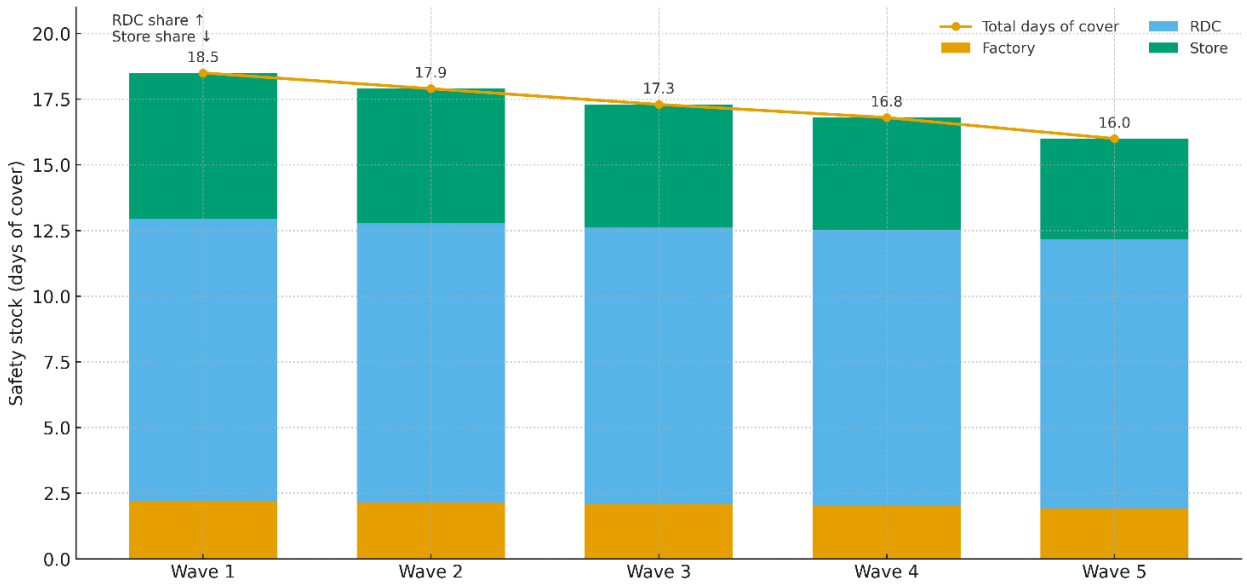

Figure 3. Safety stock by echelon across rollout waves
Source: By authors.

Channel heterogeneity shaped the magnitude of gains. Modern trade and e-commerce clusters showed the largest improvements in service and cost. In these clusters, the treatment effect on service level was 2.6 percentage points, and the reduction in landed cost reached 5.2%. General trade improved more modestly, with a 1.5 percentage-point lift in services and a 3.1% cost reduction. Quick commerce saw a 2.0 percentage point service gain, while landed cost remained flat. Short-cycle orders in quick commerce reduce pooling benefits, tempering cost gains despite service improvements. Table 3 presents effects by channel, and Figure 4 overlays 95% confidence intervals from the mixed model to highlight the differences. The pattern aligns with the broader literature that warns against assuming a uniform mapping from accuracy gains to inventory benefits across heterogeneous cost structures [11].

Table 3. Heterogeneous treatment effects by channel and echelon

Panel A: Effects by channel.

Channel	Service Level ATE (pp) [95% CI]	Landed Cost ATE (% change) [95% CI]	Stockouts ATE (per SKU-week) [95% CI]	Inventory Turnover ATE [95% CI]
Modern Trade	+2.6 [1.7, 3.5]	-5.2 [-7.1, -3.3]	-0.21 [-0.29, -0.13]	+0.9 [0.6, 1.2]
General Trade	+1.5 [0.8, 2.2]	-3.1 [-4.9, -1.3]	-0.14 [-0.20, -0.08]	+0.5 [0.3, 0.8]
E-commerce	+2.7 [1.8, 3.5]	-5.1 [-6.9, -3.2]	-0.22 [-0.30, -0.14]	+1.0 [0.7, 1.3]
Quick Commerce	+2.0 [1.1, 2.9]	-0.3 [-1.8, 1.1]	-0.12 [-0.19, -0.05]	+0.3 [0.1, 0.6]

Panel B: Effects by echelon, including safety stock shifts and buffer shares pre/post.

Echelon	Safety Stock Change (days) [95% CI]	Buffer Share Pre (%)	Buffer Share Post (%)
Factory	-0.1 [-0.3, 0.1]	12.0	12.0
Regional Distribution Center (RDC)	-1.8 [-2.3, -1.5]	58.0	64.0
Store	-0.6 [-0.9, -0.3]	30.0	24.0

Note: Effects are average treatment effects from the stepped-wedge model; intervals are 95% cluster-

robust CIs. Source: By authors.

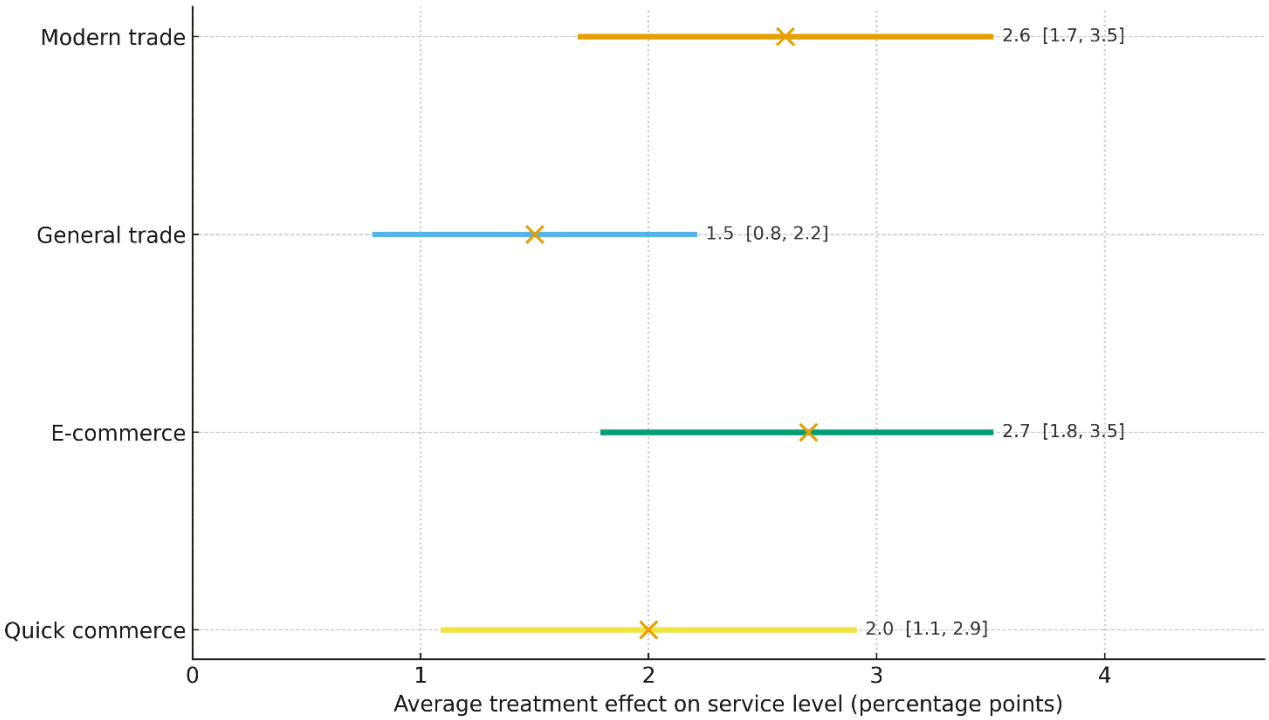

Figure 4. Channel-specific treatment effects on service level

Source: By authors.

Stress testing and robustness checks support the validity of the operational gains. In simulation, a 30% increase in lead time variance reduced service to 93.1 under the incumbent policy in the Diwali window, while the optimized policy-maintained service at 95.4 with only a small increase in days of cover. A separate stress on promotion density produced a similar ordering of policies. Figure 5 reports scenario outcomes for service, stockouts, and days of cover. In the field, placebo tests that assigned pseudo-treatment to pre-periods showed null effects, and event study plots showed flat pre-trends. Sensitivity analyses that excluded weeks with extreme rainfall documented by the meteorological authority did not change the main effects. These checks address design-specific concerns in stepped-wedge trials and focus attention on the substantive size of the impact rather than on artifacts of timing or exogenous shocks [23], [24].

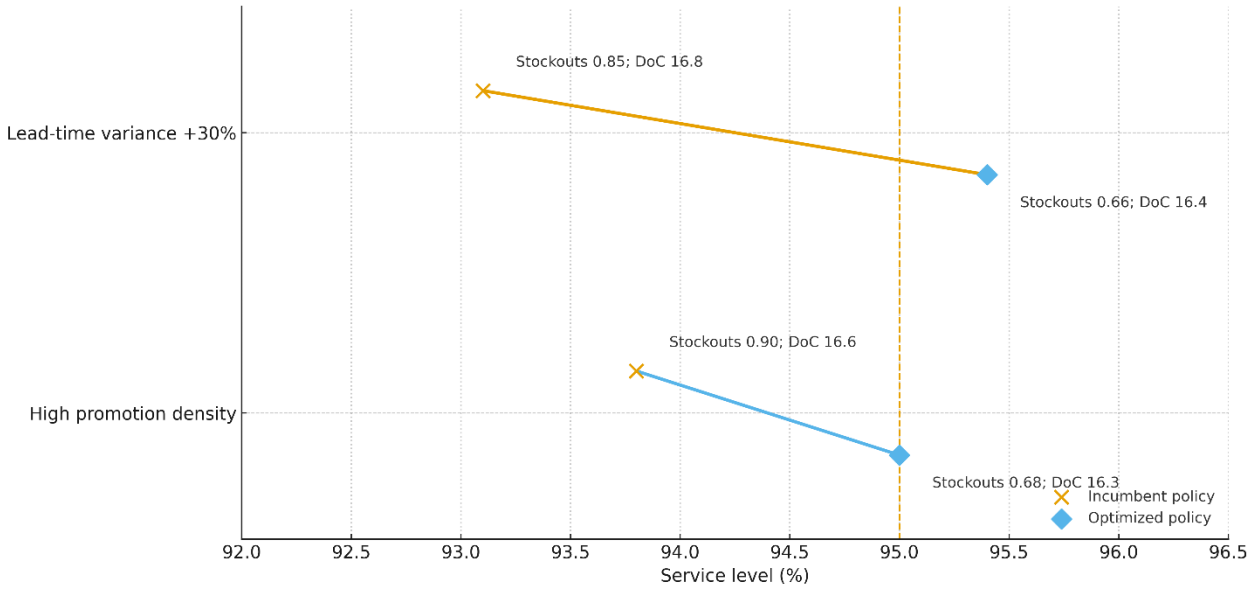

Figure 5. Stress-test outcomes

Source: By authors.

Ablation analyses clarify which modeling and optimization choices drove the improvements. Removing weather covariates raised sMAPE by 1.2 percentage points during monsoon weeks and reduced 90% interval coverage by 1.8 points. Dropping festival features raised sMAPE by 1.6 points in the fortnight around Diwali and increased stockout frequency by 9%, driven by the optimizer's response to wider forecast intervals. Replacing the ensemble with boosted trees alone reduced the service gain by 0.6 percentage points and cut the improvement in landed cost nearly in half. Figure 6 shows the change in both error and decision outcomes for each ablation. These findings connect the forecasting choices to the inventory outcomes and are consistent with evaluations that emphasize calibrated distributions and policy-aware tuning rather than accuracy alone [11], [17], [19].

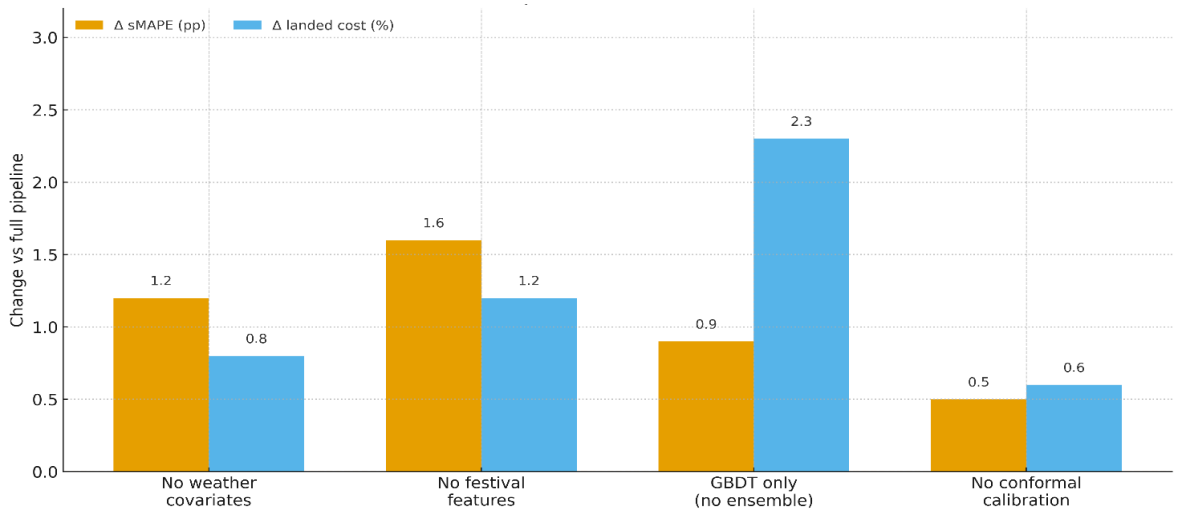

Figure 6. Ablation impacts on error and decision outcomes

Source: By authors.

Interpretation of the treatment effect underscores that accuracy, calibration, and buffer placement work together. The direction and size of the cost reduction match the safety stock decline and the slight upstream shift in buffers. The magnitude of service improvement is consistent with the coverage gains at the 90% level and with the lower tail shrinkage in the demand distribution. Finally, the results align with recent evidence that better forecasts have a larger operational impact when the holding cost is material relative to lost sales and when replenishment rules and lead times align with the demand profile [11]. This coherence across model diagnostics, optimization outputs, simulation scenarios, and field outcomes gives confidence that the pipeline produced real improvements rather than measurement artifacts, and it indicates that the approach is ready for wider deployment in India operations.

6. Discussion

The field results close the main gaps surfaced in the background and review. Prior research rarely tested a full stack, where modern multi-horizon models yield calibrated distributions that are then translated into safety-stock placement and validated in simulation before live rollout. The stepped-wedge trial demonstrates that this coupling is feasible within a complex Indian network. Accuracy gains came from better calibration, and the optimizer converted that information into a smaller, more upstream buffer while service improved, and total landed cost fell. The pattern indicates that training and selection were aligned with downstream losses, not only with forecast error, which explains the movement in service and cost [16]. It also aligns with evidence that the relationship between accuracy and inventory performance is conditional on policies and expenses [11]. The pipeline addressed both conditions by blending attention-based sequence learning with boosted trees for stability across product classes, adding conformal post-processing to improve coverage, and using guaranteed-service placement with a simulator gate to support policy changes.

The study also answers the context gap. Indian demand is shaped by monsoon shifts and festival spikes, and the channel mix now includes quick commerce. Feature governance brought weather and holiday signals into the predictive layer, and the simulator stress-tested the policies under lead time variance and promotion bursts that approximate those seasons. The heterogeneous treatment effects confirm that policy value depends on channel economics. Modern trade and e-commerce achieved greater cost reductions, while quick commerce demonstrated greater reliability with flatter cost effects, consistent with high last-mile costs and limited pooling power in the fastest channel [26]-[28].

Finally, the evaluation gap is addressed. Instead of reporting only error curves, the trial reports service, stockouts, turns, safety stock, and landed cost. The probability integral transform approached uniformity, and the continuous ranked probability score decreased, tying the calibration diagnostics to the observed improvements in the operational scorecard [17]. Rollout sequencing mirrored simulator-screened policies, and the observed effects were consistent with those of pre-deployment checks.

6.1 Managerial Implications

Managers can treat the pipeline as an operating loop rather than a one-off model build. Data engineers refresh the covariates that matter in India. Planners monitor reliability with simple diagnostics such as coverage and the probability integral transform. Policy owners set service targets and let the optimizer choose buffer placement across echelons. The simulator is used before rollout to test whether the policy holds under festival and monsoon scenarios. This routine aligns incentives because the same metrics used in weekly reviews are also used in the evaluation tables.

Several design choices carry over to other firms. First, model diversity reduces risk. The ensemble worked because attention models captured both long- and short-term temporal structure, while boosted trees handled rich tabular features and enabled fast retraining [3], [5], [6]. Second, calibration is not optional. Coverage that matches targets allows the optimizer to trade service against holding cost without overcompensating [17], [23]. Third, strategic placement matters. Moving safety stock upstream when pooling is available can reduce total days of cover for a fixed service promise, consistent with guaranteed-service theory and our observed shift toward regional distribution centers [9], [10]. Fourth, governance benefits from a digital-twin stance. Recent reviews describe how supply-chain twins integrate data, models, and what-if analysis to support ongoing policy checks and to absorb telemetry after rollout [13]. These findings complement recent evidence that data analytics maturity, knowledge management, business intelligence, and AI-enabled assistants underpin the performance gains from analytics-intensive systems [39]– [42].

6.2 Limitations and Future Work

The trial covers one large consumer-goods operation over six months. Generalizability to other sectors and longer horizons should be tested. Cold-start items and highly intermittent series remain difficult to handle. Hierarchical forecast reconciliation can improve coherence and stability across aggregation levels and may help here. Recent reviews and the MinT approach provide practical recipes for reconciling base forecasts to a coherent hierarchy that minimizes total variance [37], [38].

Coverage guarantees are not a solved problem in non-stationary retail demand. Conformal methods deliver finite-sample validity under weak assumptions, yet performance can degrade near regime changes. Comparative analyses and new variants for multi-step and multi-variate settings suggest directions for more robust calibration in production [41].

The simulator abstracts some operational details, such as carrier incentives and store substitution behavior. Future deployments can extend the agent logic and more closely tie the simulator to a digital twin, so that scenario libraries and real telemetry evolve together. Mature reviews argue that such twins improve policy governance and speed learning during staged rollouts [13].

Finally, the stepped-wedge design balances power and feasibility, yet contamination across clusters remains a possibility when managers share practices. Future studies can combine the present design with ring-fenced playbooks and staggered policy parameters to limit learning spillovers, and can add cross-market replications to test transfer. A longer horizon would allow analysis of calibration drift and retraining workloads.

7. Conclusions

This paper has focused on designing and evaluating a decision-centric pipeline that links calibrated AI forecasting, stochastic safety-stock placement, and simulation in an Indian multi-echelon fast-moving consumer goods supply chain. The study shows that AI can enhance supply-chain performance in practice when uncertainty is treated as a first-class decision input rather than a statistical afterthought. Deploying the system in India's context yielded higher service levels, fewer stockouts, leaner buffers, and lower landed costs across a staged rollout. The improvements were not the product of any single component. They arose from making uncertainty usable for planners, aligning buffer placement with network structure, and validating policies against the rhythms that shape Indian demand. In practice, calibrated risk, echelon-aware buffers, and KPI-based evaluation act together to raise service while releasing working capital.

The work is bounded by one firm, a six-month horizon, and specific category and channel mixes. Replication across sectors, longer horizons, and varied logistics conditions will test the pipeline's generalizability beyond this setting. Future research can extend hierarchical reconciliation to better support cold-start items and highly intermittent series, strengthen coverage under regime shifts, and deepen digital-twin integration so that telemetry, scenarios, and policies co-evolve in near real time. A natural next step is to embed sustainability metrics, risk constraints, and carbon-cost signals into the optimization layer and test the pipeline in other emerging markets where logistics infrastructure and digital channels are evolving rapidly. These directions define the prospects of the work: a family of configurable, decision-centric pipelines that align AI forecasting, inventory design, and simulation with the operational realities of multi-echelon supply chains.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

The authors confirm that there are no conflicts of interest.

References

[1] Makridakis, S., Spiliotis, E. and Assimakopoulos, V. M5 accuracy competition: Results, findings and conclusions. International Journal of Forecasting, 2022, 38(4), 1346–1364.

[2] Hassan, D.O. and Hassan, B.A. A comprehensive systematic review of machine learning in the retail industry: Classifications, limitations, opportunities, and challenges. Neural Computing and Applications, 2024, 37(4), 2035–2070.

[3] Lim, B., Arik, S.O., Loeff, N. and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 2021, 37(4), 1748–1764.

[4] Salinas, D., Flunkert, V., Gasthaus, J. and Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive

recurrent networks. International Journal of Forecasting, 2020, 36(3), 1181–1191.

[5] Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 785–794.

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 2017, 30, 3146–3154.

[7] Lei, D., Qi, Y., Liu, S., Geng, D. and Dong, L. Pooling and boosting for demand prediction in retail: A transfer learning approach. Manufacturing and Service Operations Management, 2024, early access. DOI: 10.1287/msom.2022.0453.

[8] Clark, A.J. and Scarf, H. Optimal policies for a multi-echelon inventory problem. Management Science, 1960, 6(4), 475–490.

[9] Graves, S.C. and Willems, S.P. Optimizing strategic safety stock placement in supply chains. Manufacturing and Service Operations Management, 2000, 2(1), 68–83.

[10] Humair, S. and Willems, S.P. Technical note—Optimizing strategic safety stock placement in general acyclic networks. Operations Research, 2011, 59(3), 781–787.

[11] Theodorou, E., Spiliotis, E. and Assimakopoulos, V. Forecast accuracy and inventory performance: Insights on their relationship from the M5 competition data. European Journal of Operational Research, 2025, 322(2), 414–426.

[12] Wiśniewski, T. and Szymański, R. Simulation-based optimisation of replenishment policy in supply chains. International Journal of Logistics Systems and Management, 2021, 38(2), 135–150.

[13] Le, T.V. and Fan, R. Digital twins for logistics and supply chain systems: Literature review, conceptual framework, research potential, and practical challenges. Computers and Industrial Engineering, 2024, 187, 109768. DOI: 10.1016/j.cie.2023.109768.

[14] Thomas, D., Helgeson, J. and Crowther, I. Supply chain disruption and agent-based modeling. NIST Advanced Manufacturing Series, 2025, AMS 100-71.

[15] Tian, X., Cao, S. and Song, Y. The impact of weather on consumer behavior and retail sales: Evidence from a convenience store chain in China. Journal of Retailing and Consumer Services, 2021, 62, 102583.

[16] Elmachtoub, A.N. and Grigas, P. Smart predict, then optimize. Management Science, 2022, 68(1), 9–26.

[17] Gneiting, T. and Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 2007, 102(477), 359–378.

[18] Koenker, R. and Bassett, G. Regression quantiles. Econometrica, 1978, 46(1), 33–50.

[19] Goltsos, T.E., Syntetos, A.A., Glock, C.H. and Ioannou, G. Inventory–forecasting: Mind the gap. European Journal of Operational Research, 2022, 299(2), 397–419. DOI: 10.1016/j.ejor.2021.07.040.

[20] Macal, C.M. and North, M.J. Tutorial on agent-based modeling and simulation. Journal of Simulation, 2010, 4(3), 151–162.

[21] Badakhshan, E., Naderi, B. and Wang, Y. Application of simulation and machine learning in supply chains. Computers and Industrial Engineering, 2024, 198, 110649.

[22] Romano, Y., Patterson, E. and Candès, E.J. Conformalized quantile regression. Advances in Neural Information Processing Systems, 2019, 32, 3538–3548.

[23] Hemming, K., Haines, T.P., Chilton, P.J., Girling, A.J. and Lilford, R.J. The stepped-wedge cluster randomised trial: Rationale, design, analysis, and reporting. BMJ, 2015, 350, h391.

[24] Grayling, M.J., Mander, A.P. and Wason, S.W. Stepped-wedge cluster randomised controlled trial designs: A review

of the statistical methodology. BMC Medical Research Methodology, 2017, 17(1), 7.

[25] NIH Collaboratory. Stepped-wedge designs. Rethinking Clinical Trials, 2023.

[26] Kearney. The rise of quick commerce: Transforming India's retail, consumer behaviors, and employment dynamics. 2025.

[27] World Bank. Connecting to compete 2023: Trade logistics in an uncertain global economy (Logistics Performance Index). 2023.

[28] Press Information Bureau, Government of India. Logistics costs in India estimated at about 7.97 percent of GDP. 2025.

[29] India Meteorological Department. Monsoon 2024: A report. 2025.

[30] ETBrandEquity. Double digit rise in online retail sales this Diwali. 2024.

[31] NielsenIQ. India's FMCG market remains resilient and is poised for growth in 2024. 2024.

[32] Ministry of Statistics and Programme Implementation. Consumer price index portal. 2025.

[33] Ministry of Electronics and Information Technology. The digital personal data protection act, 2023 (No. 22 of 2023). 2023.

[34] Carnegie Endowment for International Peace. Understanding India's new data protection law. 2023.

[35] Association for Supply Chain Management. SCOR digital standard. 2023.

[36] King, P.L. Understanding safety stock and mastering its equations. APICS Magazine, 2011.

[37] Athanasopoulos, G., Hyndman, R.J., Kourentzes, N., Petropoulos, F. and Wickramasuriya, S.L. Forecast reconciliation: A review. International Journal of Forecasting, 2024, 40(2), 430–456.

[38] Wickramasuriya, S.L., Athanasopoulos, G. and Hyndman, R.J. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. Journal of the American Statistical Association, 2019, 114(526), 804–819.

[39] Rizvi, R.A. and Khalid, W. Unlocking innovation in manufacturing: The impact of data analytics maturity and knowledge-oriented leadership. Journal of Management Science and Operations, 2025, 3(2), 17–43.

[40] Khan, J. and Rizvi, R.A. Leveraging industry 4.0 and knowledge management for enhanced innovation performance: The mediating role of organizational learning. Journal of Management Science and Operations, 2025, 3(2), 44–69.

[41] Zeng, C., Wu, W., Yao, J., Xia, J., Li, X., Wang, J. and Chen, S. AI judge assistant: A new upgrade of the enabling judicial system. Journal of Intelligence Technology and Innovation, 2025, 3(2), 74–92.

[42] Umair, M. and Simmou, S. Assessing the impact of business intelligence and its impact on the performance of industrial sector. Journal of Intelligence Technology and Innovation, 2025, 3(1), 45–57.