Supply Chain Analytics 15 (2026) 100235

Contents lists available at ScienceDirect

Supply Chain Analytics

journal homepage: www.sciencedirect.com/journal/supply-chain-analytics

A collaborative analytics framework for artificial intelligence-based demand forecasting in supply chains

Eric Weisz [a,b,*], David M. Herold [c,d], Lorenzo Bruno Prataviera [e], Maria Fernandez Pinar [f], Sebastian Kummer [a], Markus Bures [a]

[a] Institute of Transport and Logistics Management, Vienna University of Economics and Business, Vienna, Austria
[b] Department of Business, IMC Krems University of Applied Sciences, Krems, Austria
[c] Centre for Future Enterprise, School of Management, Queensland University of Technology, Brisbane, Australia
[d] Department of Logistics, SGH Warsaw School of Economics, Warsaw, Poland
[e] Department of Management, University of Verona, Verona, Italy
[f] Circly GmbH, Vienna, Austria

ARTICLE INFO

Keywords:
Demand forecasting
Artificial intelligence
Collaborative analytics
Machine learning
Data sharing
Supply chain collaboration

ABSTRACT

Despite substantial advances in artificial intelligence (AI), forecasting systems in practice often remain anchored in traditional, siloed approaches, leaving managers with unreliable and inconsistent signals for decision-making. Such forecasts tend to exacerbate the bullwhip effect, inflate safety stocks and generate significant inefficiencies in working capital and logistics costs. In fast-moving consumer goods (FMCG) supply chains, forecasts that fail to reach a minimum level of reliability are of limited value for operational planning, whereas sufficiently accurate forecasts provide a dependable foundation for managerial decisions. This practice-oriented study examines whether joint AI-based forecasting, combining downstream retailer data with advanced machine-learning techniques, enables firms to consistently achieve forecast reliability levels that are suitable for planning. Drawing on a unique dataset from a Central European manufacturer-retailer partnership, we compare three forecasting scenarios: a *Manufacturer baseline forecast* based on traditional statistical methods (ARIMA) applied to shipment data; a *Manufacturer AI forecast* using machine-learning models (XGBoost) trained exclusively on shipment data; and a *Joint AI forecast* in which machine-learning models (XGBoost) are trained on retailer warehouse outbound data. The analysis shows that Joint AI reduces total absolute error by 44.7% relative to Manufacturer AI on the test-data basis (cluster-bootstrap 95% CI: 27.8–57.3%). Under the stricter common-demand benchmark, the reduction relative to Manufacturer AI remains statistically robust at 21.2% (95% CI: 4.4–35.7%). The advantage over the traditional Manufacturer baseline is positive in point estimate but not statistically established on the common-demand benchmark. Taken together, the results indicate that, within the scope of the dyad and the two algorithmic approaches examined, collaborative data access shifts forecast errors most where their operational and financial consequences are largest, concentrating the gains of AI-based forecasting in the high-volume core of the product portfolio.

1. Introduction

Demand forecasting lies at the core of supply chain planning as production scheduling, inventory positioning, capacity allocation and transport planning all depend on the ability to anticipate future demand with sufficient accuracy and consistency [1]. Despite decades of methodological advances, however, forecasting remains a persistent source of inefficiency in many supply chains [2,3]. Forecast errors continue to propagate upstream, amplifying variability, inflating inventories and undermining service performance, in particular in fast-moving consumer goods (FMCG) environments characterized by high product variety, frequent promotions and volatile demand patterns [4,5].

Recent advances in artificial intelligence (AI) have renewed optimism that long-standing forecasting challenges can finally be overcome [6-8]. Machine learning techniques promise superior pattern recognition, the ability to model nonlinear demand drivers and improved

* Corresponding author at: Institute of Transport and Logistics Management, Vienna University of Economics and Business, Vienna, Austria.
E-mail addresses: eric.weisz@imc.ac.at (E. Weisz), d.herold@qut.edu.au (D.M. Herold), lorenzobruno.prataviera@univr.it (L.B. Prataviera), Maria@circly.at (M.F. Pinar), sebastian.kummer@wu.ac.at (S. Kummer), markus.bures@s.wu.ac.at (M. Bures).

https://doi.org/10.1016/j.sca.2026.100235
Received 31 January 2026; Received in revised form 28 July 2026; Accepted 16 August 2026
Available online 17 August 2026

responsiveness to contextual factors such as promotions or seasonality [9-11]. AI-based forecasting tools are increasingly promoted as a technological solution to the chronic limitations of traditional statistical approaches, however, despite these advances, many organizations continue to struggle with unreliable forecasts, even after adopting sophisticated analytical models [12].

We argue that this apparent problem space points to a deeper and rather underexplored issue, that is, forecasting accuracy is constrained not only by algorithms, but mainly by the organizational boundaries that shape access to demand information [13]. In practice, most forecasts are still developed within firm-level silos, relying on upstream shipment data that is temporally lagged and distorted by inventory policies, batching behavior, and ordering heuristics [14,15]. As a consequence, even advanced AI models are often trained on signals that only imperfectly reflect actual consumer demand.

Supply chain collaboration research has long argued that information sharing between manufacturers and retailers can mitigate demand distortion and reduce the bullwhip effect [16]. When firms collaborate for forecasting - or what is called 'joint forecasting' - they share demand signals that reduce distortion, improve planning and cut 'waste' across the chain by bringing forecasts closer to downstream demand signals [17]. However, empirical evidence on how collaboration interacts with modern AI-based forecasting remains limited, in particular in FMCG supply chains. In such environments, forecast errors are not evenly distributed across products as high-volume items dominate sales, logistics flows and inventory exposure, meaning that forecasting errors for these products have disproportionate operational and financial consequences. As a result, we know surprisingly little about whether and how collaborative forecasting can reshape performance in FMCG supply chains.

Against this backdrop, this study argues that forecasting should be reconceptualized as an interorganizational capability rather than a firm-level analytical function. We propose that AI-based forecasting delivers its greatest value not when deployed in isolation, but when embedded in collaborative arrangements that provide access to downstream demand signals. From this perspective, AI does not substitute for collaboration, but rather amplifies the benefits of shared data by transforming richer demand signals into actionable forecasts.

To examine this argument empirically, we conduct a comparative analysis of three forecasting approaches in an FMCG supply chain (see Fig. 1): (i) a traditional manufacturer baseline forecast based on statistical methods (ARIMA)[1]– also called *solo* forecast, (ii) a manufacturer-only AI forecast trained on upstream shipment data – also called *solo AI* forecast, and (iii) a joint AI forecast, trained on the basis of warehouse outbound data from downstream retailers, both executed with the same XGBoost[2] data model, which processed weather, promotions and calendar events. Using a unique dataset covering 87 SKUs over 36 months, we evaluate forecasting performance over a three-month test horizon. The research design isolates the effect of collaboration by holding model architecture, forecast horizon, and evaluation metrics constant across scenarios.

Our analysis focuses on three interrelated performance dimensions. First, we assess overall forecasting accuracy using both absolute and relative error metrics. Second, we examine volume-weighted performance, highlighting how forecasting improvements are distributed across high- and low-volume products. Third, we evaluate forecast stability, capturing the consistency of forecasting performance across time and products, an often-neglected, but operationally critical dimension.

This study makes three contributions: First, prior forecasting research has evaluated AI-based methods largely within firm-internal data settings, leaving it unclear whether reported AI performance gains stem from algorithmic sophistication or from the quality of the underlying demand signal. By holding model architecture (XGBoost) constant across a manufacturer-only and a joint (retailer-informed) scenario, we isolate the effect of data provenance from the effect of algorithm choice - a design that, to our knowledge, has not been applied in the FMCG joint-forecasting context. Second, whereas the collaborative-forecasting literature treats information sharing largely as a contextual enabler of performance, we show empirically that AI's forecasting gains only materialize once it is embedded in a data-sharing arrangement, repositioning AI as an amplifier of collaboration rather than a substitute for it. Third, while forecasting studies typically rely on average error metrics that treat all SKUs as equally consequential, we introduce a volume-weighted and stability-oriented evaluation lens, showing that collaboration's benefits concentrate precisely where operational and financial exposure is highest — a pattern invisible to conventional MAE/MAPE-based comparisons.

The remainder of the paper is structured as follows. Section 2 reviews the literature on forecasting, AI, and supply chain collaboration and develops the theoretical grounding for the study. Section 3 outlines the research design, data, forecasting scenarios and evaluation metrics. Section 4 presents the empirical results. Section 5 discusses the theoretical and managerial implications of the findings. Section 6 concludes with limitations and directions for future research.

2. AI Forecasting and Supply Chain Collaboration

Demand forecasting has long been recognized as a foundational capability in supply chain management [3,18]. Forecasts shape production planning, inventory positioning, capacity allocation, and transportation decisions, thereby influencing cost efficiency, service levels and overall supply chain performance [2]. Despite decades of methodological advances, however, forecasting accuracy remains a persistent challenge, particularly in environments characterized by volatile demand, frequent promotions, and high product variety [5]. This section reviews the literature on forecasting, artificial intelligence and supply chain collaboration and develops four interrelated theoretical propositions that motivate the present study. The four propositions developed in this section guide the directional interpretation of the empirical results. We examine whether the observed performance differences are consistent with each proposition's predicted direction and complement this pattern-based assessment with inferential robustness checks where SKU-level error data allow.

2.1. Forecasting as a firm-centric function versus interorganizational demand reality

The forecasting literature has traditionally conceptualized demand prediction as a firm-internal analytical function or what we call *solo* approaches. Classical work focuses on improving forecast accuracy through statistical model selection, parameter tuning and the refinement of internal data processes [19]. Within this stream of literature, forecasting performance is typically treated as the outcome of methodological sophistication applied to historical demand or shipment data available within the firm [1].

In contrast, supply chain research has consistently demonstrated that demand uncertainty is inherently interorganizational. The bullwhip effect literature shows that demand variability is amplified as information moves upstream across organizational boundaries, driven by order batching, inventory policies, lead times and information delays [20]. Importantly, these distortions do not originate within individual firms, but emerge from interactions between manufacturers, distributors and retailers. As a result, the demand signals observed by upstream actors often diverge substantially from actual consumer purchases [21].

This creates a fundamental tension in the literature and in practice. While demand distortion is widely acknowledged as an

[1] ARIMA: 'AutoRegressive Integrated Moving Average', a classical statistical time-series forecasting method.

[2] XGBoost: 'eXtreme Gradient Boosting', a machine-learning method based on gradient-boosted decision trees.

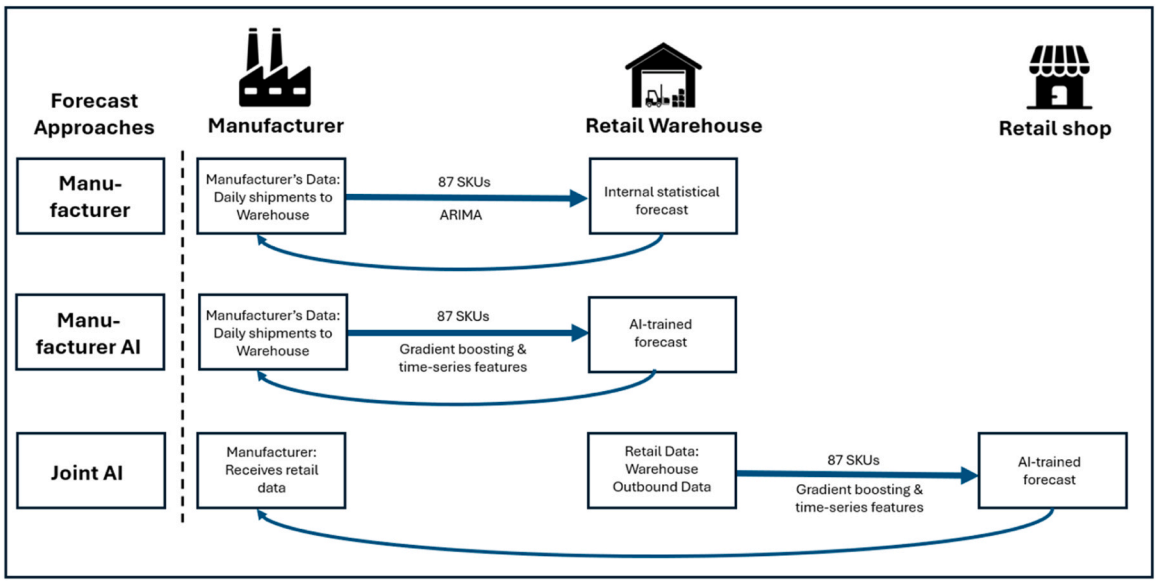

Fig. 1. Forecasting approaches considered in this study.

interorganizational phenomenon, forecasting models are typically developed and evaluated within organizational silos [22,23]. Collaborative practices such as CPFR (Collaborative Planning, Forecasting, and Replenishment), vendor-managed inventory and shared sell-through data are discussed extensively in the collaboration literature, yet their implications for forecasting performance are often treated as contextual enablers rather than as core determinants of forecast quality [24,25]. Forecasting thus remains conceptually anchored at the firm level, even as demand uncertainty spans organizational interfaces. In other words, forecasting approaches developed in isolation may be structurally limited in their ability to capture true demand dynamics, i.e. if demand distortion arises between organizations, then forecasting arrangements that remain confined within organizational boundaries are unlikely to fully resolve forecasting inaccuracies.

Proposition 1. *AI-enabled joint forecasting approaches outperform non-collaborative (solo) forecasting approaches in terms of overall forecast accuracy.*

2.2. Algorithmic sophistication versus organizational data boundaries

Recent advances in artificial intelligence have intensified interest in AI-based forecasting as a means of overcoming long-standing forecasting challenges. Machine learning techniques, including gradient boosting and neural networks, offer enhanced capabilities for capturing nonlinear relationships, interaction effects and complex demand patterns [26]. Numerous studies report that AI-based models outperform traditional statistical approaches, reinforcing the perception that forecasting limitations are primarily technical in nature [7].

However, this literature often implicitly assumes that data availability and data quality are given. In practice, access to demand data is shaped by organizational boundaries, governance arrangements and power relations along the supply chain. Many AI-based forecasting systems are trained on upstream shipment or order data, which are temporally lagged and distorted by inventory decisions and ordering heuristics [25]. While AI models may extract richer patterns from such data, they remain constrained by the informational quality of the underlying demand signal.

Despite this, empirical studies rarely distinguish between AI forecasting in collaborative versus non-collaborative settings. Algorithmic performance improvements are often attributed to model sophistication without systematically accounting for differences in data provenance. As a result, it remains unclear whether AI improves forecasting because of its analytical capabilities per se, or because it enables the exploitation of richer demand information when downstream data are available. This gives rise to a second tension: AI enhances analytical capability, but does not resolve the organizational data boundaries that shape forecasting inputs. Understanding the role of AI in forecasting therefore requires examining how advanced analytics interact with interorganizational collaboration, rather than treating AI as a standalone solution.

Proposition 2. *AI-based forecasting models trained on upstream shipment data do not deliver comparable performance improvements to AI models trained on collaboratively shared downstream demand data.*

2.3. Average forecast accuracy versus operational and economic relevance

A third tension concerns how forecasting performance is conceptualized and evaluated. The dominant approach in the forecasting literature relies on average error metrics such as MAE, MAPE or RMSE. These metrics implicitly assume that all forecast errors are equally consequential, regardless of where they occur within the product portfolio or over time.

From an operational perspective, however, forecast errors are economically asymmetric. High-volume products dominate sales revenues, logistics flows and inventory exposure, meaning that forecasting errors for these products carry disproportionately large operational and financial consequences. Errors in low-volume or niche products, by contrast, may have limited impact even when percentage deviations are large. Despite this, relatively little research explicitly examines how forecasting improvements are distributed across products or evaluates performance using volume-weighted measures.

In addition, the literature has paid limited attention to forecast stability as a performance dimension. Forecasts that are occasionally accurate but highly volatile can undermine planning processes, trigger frequent replanning and erode managerial trust. Yet most studies focus on point accuracy at a given horizon rather than on the consistency of forecasting performance across time and products. This focus on average accuracy obscures questions of forecast usability and economic relevance, particularly in high-velocity environments such as FMCG supply chains. A more nuanced understanding of forecasting performance requires accounting for both the distribution of errors across product portfolios and the stability of forecasts over time.

Proposition 3. *The performance advantages of joint AI forecasting are disproportionately concentrated in high-volume products, where forecasting errors carry greater operational and economic consequences.*

Proposition 4. *Joint AI forecasting approaches exhibit greater forecast stability over time and across products than non-collaborative forecasting approaches.*

3. Methodology

With this study, we aim to give the reader a more nuanced understanding to what extent joint AI-forecasting can improve prediction accuracy compared to non-collaborative approaches in an FMCG supply chain. To do so, we use an extensive and rarely accessible dataset, obtained through a collaboration between a large grocery retailer, operating thousands of stores across Central Europe and a snack manufacturer supplying a portfolio of savory products in Austria. The collaboration for the joint forecasting comprises an information exchange where the retailer shares warehouse inventory and outbound data (akin to Vendor Managed Inventory (VMI), but without relinquishing order control), thereby enabling the manufacturer to generate forecast models based on these downstream signals [21]. It is important to note that the research team included two industry practitioners working for an AI-forecasting provider in Vienna specializing in FMCG supply chains, who developed and implemented the forecasting models tested in this study. Their involvement, combined with academic partners, reflects what Stank et al. [27] describe as the formula for foundational work: collaboration between researchers and practitioners to tackle problems that matter in practice as well as in theory.

3.1. Data collection

The two partners, below referred to as "manufacturer" and "retailer", agreed to share selected historical operational data for 36 months for 87 stock-keeping units (SKUs). Two data streams for the forecasting were collected:

Manufacturer shipment data: daily shipment from the manufacturer to the retailer's warehouses: the data represent what the manufacturer sees as demand.
Retailer warehouse data: daily shipments from the warehouses to retail stores - the outbound volumes are closer to consumer demand and thus serve as the downstream signal for joint forecasting.

Extensive data preparation was necessary to align the datasets. The partners unified product identifiers, harmonized date formats and ensured that the same 87 SKUs were present in each stream.

3.2. Forecasting methodologies

It is crucial to ensure that the AI-forecast models are trained properly in order to forecast the future with a certain precision. To train the algorithm of AI-forecast model, Circly's AI algorithm utilizes gradient boosting (XGBoost), that is, a machine learning technique that incrementally builds decision-tree models to minimize forecast errors and capture complex nonlinear patterns from a combination of historical sales patterns with contextual factors such as weather, promotions and other relevant factors. This study focused solely on the factors of weather, calendar information and promotions. This integration allows the model to detect demand patterns like seasonality and generate more accurate forecasts of future demand. It is important to note that AI-forecasting models in the context of time-series forecasting can usually not be pre-trained, thus it can be argued that the better and further the quality of downstream information exchange between partners along the supply chain, the better the algorithm can be trained and subsequently leading to the higher prediction accuracy.

(i) Manufacturer Forecast: the manufacturer's internal statistical baseline and promotional uplift forecast (ARIMA) based on daily shipments to the retailers' warehouses. The forecast scenario provides a monthly forecast for each SKU based on historical shipments based on a typical moving-average or exponential smoothing approach,

(ii) Manufacturer AI Forecast: an AI-based forecasting model trained (incl. gradient boosting & time-series features) exclusively on the manufacturer's daily shipment data to the retailer warehouses. Forecasts were produced using a rolling horizon, i.e. on a monthly basis.

(iii) Joint AI Forecast (based on retail outbound data): uses the same AI model architecture as in (ii), but trained on retailer warehouse outbound data to the retail shops, representing more accurate consumer data. In this joint scenario, we assume that the manufacturer has access to the retailers' downstream outbound volumes through a collaborative agreement. Forecasts were likewise generated on a rolling monthly basis.

Forecasts were generated at the manufacturer aggregation level (i.e., total predicted shipments by SKU and month) for 87 stock-keeping units (SKUs) over a three-month horizon (January-March 2025), reflecting the manufacturer's planning perspective.

Let i = 1, …, N index the stock-keeping units (SKUs), with N = 87, and let t index months within the study horizon. We write $y_{i,t}$ for the realized demand of SKU i in month t, measured as the actual orders placed by the retailer's warehouses with the manufacturer (the common evaluation benchmark), $y_{i,t}^{ship}$ for the manufacturer's daily shipments to the retailer's warehouses aggregated to month t, and $y_{i,t}^{out}$ for the retailer's daily warehouse outbound volumes to its stores aggregated to month t. Forecasts are one-step-ahead at monthly granularity; $\hat{y}_{i,t+1}^{(s)}$ denotes the forecast of scenario $s \in \{1, 2, 3\}$ for SKU i and month $t + 1$, and $e_{i,t} = \hat{y}_{i,t} - y_{i,t}$ the corresponding forecast error.

The three forecasting scenarios differ in exactly two design dimensions: the hypothesis class of the forecasting function and the provenance of the training data. Scenario 1 (Manufacturer) applies a classical linear time-series model to the manufacturer's own shipment history; Scenario 2 (Manufacturer AI) applies a gradient-boosted tree ensemble to the same shipment history augmented with exogenous features; Scenario 3 (Joint AI) applies the identical tree ensemble and identical feature construction to the retailer's warehouse outbound series obtained through the collaborative data-sharing arrangement. Formally,

$$\hat{y}_{i,t+1}^{(1)} = f_i^{ARIMA}\left(y_{i,1:t}^{ship}\right) \tag{1}$$

$$\hat{y}_{i,t+1}^{(2)} = f^{XGB}\left(x_{i,t}^{ship}\right), \quad x_{i,t}^{ship} = \varphi\left(y_{i,1:t}^{ship}, \; z_{i,t}\right) \tag{2}$$

$$\hat{y}_{i,t+1}^{(3)} = f^{XGB}\left(x_{i,t}^{out}\right), \quad x_{i,t}^{out} = \varphi\left(y_{i,1:t}^{out}, \; z_{i,t}\right) \tag{3}$$

where φ(·) denotes the shared feature-construction map defined in Eq. (13) and the exogenous covariates are weather, promotions and calendar variables. Because Scenarios 2 and 3 share the hypothesis class, the feature map φ, the training procedure, and the rolling-origin protocol, and differ solely in the input series, their comparison isolates the effect of data provenance: it constitutes a data-provenance ablation in which the collaborative input signal is the single manipulated factor.

All comparative evaluation is conducted at the SKU–month level. This choice is not a data limitation but reflects the decision-relevant granularity of the setting: the manufacturer's operational planning runs on a monthly cycle within a commercial demand-planning system, so monthly forecasts are the quantities that enter production, inventory, and capacity decisions. The collaborative arrangement, by contrast, makes forecasting at daily granularity feasible for the first time, because the retailer's warehouse outbound data provide consumption-near signals at daily resolution — a level of detail that shipment data, temporally distorted by order batching and inventory policies, cannot supply.

Daily forecasting capability is therefore itself an outcome of the collaboration rather than a precondition of it; exploiting this finer granularity operationally is a natural avenue for future work, while the present evaluation deliberately remains at the granularity at which the manufacturer's planning decisions are actually made.

3.3. Model selection and disclosure constraints

The two model classes represent the two dominant paradigms of practice-facing demand forecasting and occupy deliberately different points in the bias–variance and interpretability–flexibility trade-offs. ARIMA restricts the hypothesis space to linear dependence on the series' own past and past innovations under stationarity after differencing; this yields low estimation variance on short series, transparent parameters and low maintenance cost, but introduces structural bias whenever demand responds nonlinearly to promotions, weather or calendar effects. Gradient-boosted tree ensembles remove the linearity and stationarity restrictions and approximate interactions between lagged demand and exogenous covariates; the regularization in Eq. (9) controls the associated variance, but the model class requires more data, monthly retraining and sacrifices parameter-level interpretability. Deep sequence architectures, such as LSTM or Transformer-based models, were excluded because their sample complexity exceeds the available historical window and because introducing an additional algorithmic degree of freedom would confound the data-provenance effect the design isolates. The pairing of one representative of each paradigm, with the machine-learning representative held fixed across the two data conditions, is therefore the minimal design that separates the contribution of algorithmic flexibility from the contribution of collaborative data access.

It should be noted that the XGBoost models used in this study were developed and operated on our industry partner's proprietary forecasting platform. As the underlying model architecture and feature-weighting logic constitute commercially sensitive intellectual property, feature-level diagnostics (e.g., gain-based importance scores or SHAP values) could not be extracted or disclosed as part of this collaboration. The forecasting models were retrained on a rolling monthly basis using a fixed feature set (weather, promotions, and calendar events) held constant across the Manufacturer AI and Joint AI scenarios; no feature selection procedure was applied beyond this fixed specification, and hyperparameter configuration was managed internally by the forecasting platform rather than tuned or exposed for this study. As with the feature-interpretability constraint above, this reflects the boundary of what a proprietary, production-grade forecasting system can disclose in an applied research collaboration.

Eqs. (4)–(13) report the mathematical specification of the model classes, objective function, regularization logic, learning procedure and feature construction used for analytical transparency, while proprietary implementation-level details, including exact hyperparameter values, platform-specific feature transformations, feature-weighting rules and feature-importance diagnostics, remain outside the disclosure scope of this collaboration for confidentiality and commercial reasons.

From a computational-complexity standpoint, ARIMA estimation scales approximately linearly with series length per SKU and is fitted independently for each of the 87 SKUs, making it lightweight to run and refresh on a rolling basis even without specialized infrastructure. XGBoost's training cost scales with the number of boosting rounds, tree depth, and training-sample size, and — because a separate model is retrained monthly for each SKU under a rolling horizon — its aggregate computational demand is substantially higher than ARIMA's, even though each individual model remains fast to train on datasets of the size used here (33 months of daily data per SKU). Inference (generating a forecast from an already-trained model) is fast for both methods; the higher cost of XGBoost lies in training and periodic retraining, not in producing forecasts. We do not report exact runtimes, as these depend on the proprietary platform's infrastructure and implementation choices, which fall outside the disclosure scope of this collaboration; however, the qualitative complexity ordering (ARIMA lighter, XGBoost heavier, particularly at the retraining stage) holds independent of implementation details.

The Manufacturer scenario employs an autoregressive integrated moving-average model ARIMA(p, *d*, q), estimated independently per SKU. Writing B for the backshift operator, $B^k y_t = y_{t-k}$, the model is

$$\varphi_p(B)\ (1 - B)^d\, y_t = \theta_q(B)\, \varepsilon_t \tag{4}$$

$$\varphi_p(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p \tag{5}$$

$$\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{6}$$

with $\varepsilon_t \sim WN(0, \sigma^2)$ white noise. Here p is the autoregressive order (dependence of current demand on its own p most recent values), d the degree of differencing required to render the series stationary, and q the moving-average order (dependence on the q most recent shocks). The model assumes that, after d-fold differencing, the series is covariance-stationary and its dynamics are adequately captured by a linear combination of past values and past innovations; the roots of φ and θ lie outside the unit circle (stationarity and invertibility). Orders are selected per SKU by information-criterion-based search on the training window. These assumptions — linearity and stationarity — are exactly what restrict ARIMA's ability to exploit nonlinear promotional uplifts or exogenous drivers such as weather, motivating the machine-learning alternative.

Scenarios 2 and 3 employ gradient-boosted regression trees in the XGBoost formulation [28]. The prediction for feature vector x_i is an additive ensemble of K regression trees,

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{7}$$

where $\mathcal{F}$ is the space of regression trees, each mapping an input to one of T leaf weights $w \in \mathbb{R}^T$. Training minimizes a regularized objective that trades data fit against ensemble complexity,

$$L(\varphi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{8}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{9}$$

with l a differentiable convex loss (squared error in the regression setting), γ penalizing the number of leaves T, and λ an L2 penalty on leaf weights. The ensemble is built stagewise. At boosting round k, with $g_i = \partial_{\hat{y}^{(k-1)}} l\left(y_i, \hat{y}^{(k-1)}\right)$ and h_i the first and second derivatives of the loss with respect to the previous-round prediction, the objective is approximated to second order,

$$L^{(k)} \approx \sum_{i=1}^{n} \left(g_i f_k(x_i) + \frac{1}{2} h_i f_k^2(x_i) \right) + \Omega(f_k) \tag{10}$$

For a fixed tree structure with instance set I_j in leaf j, the optimal leaf weight and the resulting split gain follow in closed form,

$$w_j^* = -\frac{G_j}{H_j + \lambda}, G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i \tag{11}$$

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \tag{12}$$

Trees are grown greedily by maximizing this gain over candidate splits; γ acts as the minimum gain required to add a leaf, and λ shrinks leaf weights. In the forecasting application, one model per SKU is retrained monthly under the rolling-origin protocol described in Section

3.4, mapping the feature vector defined in Eq. (13) to next-month demand. Hyperparameter configuration is managed internally by the forecasting platform and is therefore reported at the level of the model class rather than as tuned values; this preserves the disclosure boundary of the industry collaboration without affecting the analytical formulation above.

3.4. Forecast Performance and evaluation metrics

Forecast performance is divided into two dimensions a) prediction accuracy, representing the degree to which forecasted values correspond to the actual observed outcomes, and b) stability, reflecting the consistency of this accuracy over time.

Forecast accuracy was assessed through two complementary evaluation targets. First, we used rolling-origin backtesting [29] over the three-month out-of-sample window. For each forecast origin preceding January, February and March 2025, models were estimated using only information available at that origin and produced one-step-ahead monthly forecasts. Second, all scenarios were evaluated against actual demand orders, i.e., the realized orders placed by the retailer's warehouses with the manufacturer, to provide a common benchmark across forecasting approaches.

Both machine-learning scenarios use the identical feature map φ, applied to the respective input series. For SKU i at forecast origin t, the feature vector concatenates four groups,

$$x_{i,t} = \left[y_{i,t-1}, \quad \ldots, \quad y_{i,t-L}; c_t; p_{i,t}; w_t \right] \tag{13}$$

comprising (i) autoregressive lags of the target series up to order L, (ii) calendar features c_t encoding month, holiday and season indicators, (iii) promotion indicators $p_{i,t}$ at SKU level, and (iv) weather covariates w_t. The feature set is fixed a priori and held constant across the Manufacturer AI and Joint AI scenarios; no scenario-specific feature selection is applied.

Let the study horizon comprise months 1, …, T, with the final three months forming the out-of-sample evaluation window. Define the set of forecast origins $\mathcal{T}$ as the months preceding January, February and March 2025. For each origin $\tau \in \mathcal{T}$ and each scenario s, the model is re-estimated on the training window ending at τ and produces the one-step-ahead monthly forecast

$$\hat{y}^{(s)}_{i,\tau+1} = f^{(s)}_{\tau}\left(\mathcal{D}^{(s)}_{i,1:\tau} \right) \tag{14}$$

so that estimation uses only information available at the forecast origin (out-of-sample backtesting). Under this protocol, each of the 87 SKUs contributes three out-of-sample forecast–actual pairs per scenario, evaluated against the realized demand orders. The training windows span multiple seasonal cycles, and calendar, promotion and weather features are included to allow the models to capture recurring seasonal structure. The three-month evaluation window is therefore the out-of-sample test period, not the full period from which the models learn. Machine-learning models are retrained at every origin; the feature map and hyperparameter configuration remain fixed across origins and scenarios.

To assess the prediction accuracy, we used both absolute and relative error measures. Absolute metrics evaluate the magnitude of errors regardless of scale, while relative metrics normalize errors by dividing by actual values. Table 1 shows the metrics that were computed for each dataset and forecast horizon.

Stability was assessed by examining the standard deviation and interquartile range of errors across SKUs and over time, as well as the coefficient of variation of monthly MAE. All metrics were computed at the manufacturer aggregation level and averaged over the three-month period.

To assess robustness, we complement the descriptive error metrics with inferential tests based on the SKU-level error structure. We use cluster-bootstrap confidence intervals, resampling SKUs to preserve within-SKU dependence across months, for the volume-weighted total absolute error reductions. We also report paired Wilcoxon signed-rank tests at the SKU level to assess whether improvements are uniformly distributed across the portfolio. Together, these analyses distinguish volume-weighted robustness from per-SKU uniformity and therefore directly support the interpretation of Proposition 3.

Table 1
Evaluation metrics.

Name	Acronym	Description
Total Absolute Error	TAE	the sum of absolute errors over the forecast horizon, capturing the aggregate deviation between forecasted and actual quantities
Mean Absolute Error	MAE	the average absolute error per observation
Total Signed Error	TSE or bias	the sum of signed errors, indicating over- or under-forecasting tendencies
Mean Error	ME	the average signed error
Root Mean Squared Error	RMSE	the square root of the mean squared error, emphasizing larger deviations
Mean Absolute Percentage Error	MAPE	average absolute percentage error, sensitive to low volumes; if the metric diverges due to a zero-baseline value, it defaults to Absolute Error (AE)
Symmetric MAPE	sMAPE	a bounded alternative to MAPE that symmetrically scales the error
Weighted MAPE	WMAPE	the sum of absolute errors divided by the sum of actual quantities, mitigating distortions from small denominators
Coefficient of determination	R^2	proportion of variance explained by the forecast relative to a naive mean model
Mean Squared Logarithmic Error	MSLE	the squared difference between log-transformed quantities, penalizing large relative errors, especially underestimation

With forecast errors $e_{i,t} = \hat{y}_{i,t} - y_{i,t}$ over the n = 261 SKU–month cells of the evaluation window (87 SKUs × 3 months), accuracy is measured by

$$TAE = i,t\sum |ei,t|, MAE = n1i,t\sum |ei,t| \tag{15}$$

$$TSE = i,t\sum ei,t, ME = n1i,t\sum ei,t \tag{16}$$

$$RMSE = n1i,t\sum ei,t2 \tag{17}$$

$$sMAPE = \frac{100}{n}\sum_{i,t}\frac{2|e_{i,t}|}{|y_{i,t}| + |\hat{y}_{i,t}|} \tag{18}$$

$$WMAPE = 100\cdot\frac{\sum_{i,t}|e_{i,t}|}{\sum_{i,t}y_{i,t}} \tag{19}$$

$$R^2 = 1 - \frac{\sum_{i,t}e^2_{i,t}}{\sum_{i,t}\left(y_{i,t} - \bar{y}\right)^2} \tag{20}$$

TAE and MAE capture aggregate and average absolute deviation; TSE and ME the direction of bias; RMSE emphasizes large deviations; sMAPE bounds relative errors symmetrically; WMAPE weights relative error by realized volume and is therefore the primary relative measure in a portfolio in which error consequences scale with volume; R^2 measures variance explained relative to a mean benchmark. MAPE and MSLE, together with the stability measures (standard deviation, interquartile range and the coefficient of variation of monthly MAE), are defined in Appendix A.1.

3.5. Administration/AI Forecast Implementation

Typically, retailers and manufacturers rarely share data, so it is

important to mention two crucial intermediaries involved in the project: a) the forecast models were implemented and administered by the neutral AI provider Circly [30], based in Vienna, Austria, specializing in FMCG supply chains (two co-authors are actively involved at Circly), and b) the project was facilitated as part of a Joint Forecasting working group named the 'Efficient Consumer Response' (ECR) Austria [31]. As part of the project, data-sharing agreements were established to protect confidentiality and ensure compliance with competition law. The involvement of neutral third parties ensured that neither partner had undue influence over the modelling process, thereby facilitating trust around the data sharing.

4. Results

The dataset combined information from a leading snack manufacturer in Austria, supplying a broad portfolio of savory products, and a large grocery retailer with hundreds of stores across Central Europe. It covered 87 products (SKUs) and included three months (January-March 2025) of recent sales and shipment data, including both regular weeks and promotion weeks, thereby providing a realistic challenge for forecasting models. As described in Section 3.4, all models were evaluated over the January–March 2025 out-of-sample window using a rolling-origin design. To ensure transparency, we report the forecasting models used, the full set of error metrics, and the backtesting procedure in detail (Section 3.4). The cumulative error distributions reported in Fig. 2 additionally allow readers to assess whether performance gains are concentrated in a small number of SKUs or extreme observations, rather than being driven by a few outliers.

Table 2 summarizes absolute error metrics when forecasts are compared against their respective test datasets. The 'Joint AI' forecasting model consistently delivers lower errors than either the 'Manufacturer' or the 'Manufacturer AI' model. For example, total absolute error (TAE) decreases from roughly 128k units in the baseline to about 77k units under joint forecasting, a reduction of 40%. Mean absolute error similarly falls from 492 units to 295 units. Root mean squared error (RMSE) drops by more than one-third, indicating fewer large deviations.

Table 2
Absolute error metrics (test data).

Dataset	TAE	MAE	TSE (bias)	ME	RMSE
Manufacturer / Solo forecast / manufacturer's shipments	128 498.00	492.33	15 040.00	57.62	733.78
Manufacturer AI / Solo AI forecast manufacturer's shipments	139 299.40	533.71	60 662.20	232.42	1 150.04
Joint AI / retailer outbound	76 991.80	294.99	21 340.80	81.77	480.97

Table 3 presents relative error metrics. The 'Joint AI' forecasting model achieves the lowest symmetric MAPE (sMAPE) and weighted MAPE (WMAPE), indicating that on average it produces smaller percentage errors and performs well on high-volume items. Its MAPE is markedly lower than that of the 'Manufacturer' and 'Manufacturer AI' forecasting models, since its data is less sparse with fewer zeros, reducing extreme errors. The coefficient of determination (R^2) of 0.84 reflects a high proportion of variance explained, more than tripling the Manufacturer AI scenario's 0.27 and well above the Manufacturer baseline's 0.70. Lower values of MSLE further indicate fewer severe relative errors.

Table 3
Relative error metrics (test data).

Dataset	MAPE	sMAPE	WMAPE	R^2	MSLE
Manufacturer / Solo forecast / manufacturer's shipments	1 709.53	52.39	39.85	0.70	1.77
Manufacturer AI / Solo AI forecast / manufacturer shipments	2 051.48	47.26	43.02	0.27	1.67
Joint AI / retailer outbound	41.05	28.86	25.20	0.84	0.20

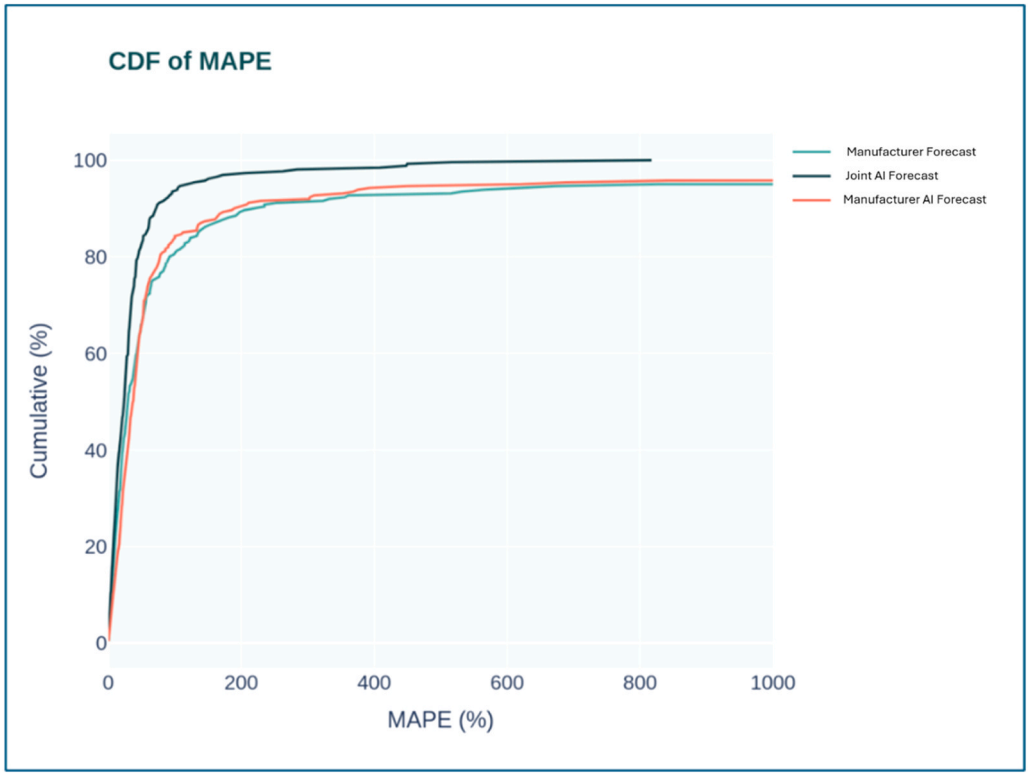

Fig. 2. Cumulative distribution of MAPE across SKUs and months. This plot was zoomed to a maximum of 1000 MAPE (%).

Fig. 2 shows the cumulative distribution function (CDF) of MAPE across SKUs and dates: the 'Joint AI' curve rises steeply and reaches the 100% mark at lower MAPE values than both the 'Manufacturer' and 'Manufacturer AI' scenarios, indicating the absence of extreme percentage errors and a consistently higher share of predictions with lower relative errors. It also should be noted that the Manufacturer and Manufacturer AI curves do not reach 100% within the plotted range even at 1000% MAPE, indicating that both scenarios contain a small number of SKU-month observations with extremely large relative errors (driven by near-zero actual values); the Joint AI curve, by contrast, reaches 100% by roughly 800% MAPE, confirming that collaboration eliminates the most extreme forecasting failures rather than only improving typical-case accuracy.

4.1. Prediction accuracy: Performance Against Demand

While the 'Joint AI' forecasting model excels when evaluated against its own test data, it is also crucial to assess how well the forecasts anticipate the retailer's actual demand, i.e. the orders placed by the retailers' warehouses to the manufacturer. This comparison is methodologically important: Tables 2–3 evaluate each scenario against its own respective test data (manufacturer shipments for the Manufacturer and Manufacturer AI scenarios, retailer outbound volumes for Joint AI), meaning the underlying target series differ in provenance and potentially in noise structure. Evaluating all three scenarios against a single common benchmark — actual demand orders — addresses this issue: if Joint AI's advantage in Tables 2–3 were driven merely by predicting an inherently smoother downstream series rather than by a genuine collaboration effect, that advantage should narrow or disappear once all scenarios are judged against the same target. As Table 4 shows, it does not.

Table 4 summarizes the absolute error metrics when forecasts are compared against the common demand benchmark. Joint AI achieves the lowest total absolute error, mean absolute error and RMSE in point-estimate terms. However, the inferential results show that the comparison needs to be interpreted carefully. Joint AI reduces total absolute error by 21.2% relative to Manufacturer AI on the common demand benchmark, and this reduction is statistically robust (cluster-bootstrap 95% CI: 4.4–35.7%). Relative to the Manufacturer baseline, the point estimate is also lower, but the confidence interval includes zero (9.2%, 95% CI: −1.9–19.3%). We therefore interpret the Manufacturer-baseline comparison as directionally positive, but not as a statistically established advantage. MAE and RMSE likewise favor the 'Joint AI' approach. Interestingly, the Manufacturer scenario shows considerably lower RMSE and ME than Manufacturer AI, suggesting that the Manufacturer AI model is more affected by large deviations and positive bias. The three models tend to over-forecast, with the 'Manufacturer AI' case exhibiting higher bias.

To assess whether these differences are statistically reliable rather than attributable to sampling variation, we complement the descriptive metrics with formal inference computed on the SKU-level errors underlying Fig. 2. Cluster-bootstrap confidence intervals (5000 replications, resampling SKUs to respect within-SKU correlation across months) place the reduction in total absolute error achieved by Joint AI over Manufacturer AI at 44.7% (95% CI: 27.8–57.3%) on the test-data basis of Table 2, and at 21.2% (95% CI: 4.4–35.7%) under the stricter common demand benchmark of Table 4; both intervals exclude zero. Per-SKU paired Wilcoxon signed-rank tests (n = 87, one-sided, Holm-corrected) show the advantage to be directional but not uniformly distributed across the portfolio (Joint AI vs. Manufacturer AI: W = 1625, p = 0.166, rank-biserial r = 0.15, 56% of SKUs improved): the collaborative gain is concentrated where volumes — and therefore the operational and financial consequences of forecast errors — are largest, consistent with Proposition 3. The corresponding volume-weighted reduction relative to the Manufacturer baseline on the demand benchmark is positive in point estimate (9.2%) but its interval includes zero (95% CI: −1.9–19.3%), so we do not claim a statistically established advantage of Joint AI over the statistical baseline on that basis. Formal definitions of the test statistic, effect size, and bootstrap procedure are provided in Appendix A.1.

Relative error metrics in Table 5 reveal a more nuanced picture. Weighted MAPE (WMAPE) and sMAPE both improve markedly in the 'Joint AI' scenario (37.49 & 50.09% respectively) compared with 'Manufacturer AI' (47.56 & 51.75%) and 'Manufacturer' (41.29 & 54.98%). However, MAPE remains high across all models, reflecting challenges with low-volume SKUs where small denominators inflate the percentage error. The 'Joint AI' again achieves the highest R^2 and lowest MSLE, indicating that it explains more variability in demand and produces fewer large relative errors. The 'Manufacturer' MSLE is highest, suggesting that although it may have lower MAPE for very low demand items, it struggles to capture overall variance.

Fig. 3 shows, in the top panel, monthly aggregated demand against the predictions of each forecasting scenario over the three-month test horizon, and, in the bottom panel, the corresponding residuals (predictions minus demand). The Joint AI forecast tracks demand closely across all three months, with residuals remaining consistently small and stable. The Manufacturer AI forecast, by contrast, systematically over-predicts demand, with residuals staying positive and substantial throughout the horizon — indicating a persistent upward bias rather than random error. The Manufacturer forecast shows the opposite pattern in the first two months, under-predicting demand before crossing over toward the end of the horizon, reflecting greater month-to-month volatility rather than a consistent directional bias. However, a consistent bias (as in Manufacturer AI) can in principle be corrected with a simple adjustment factor, whereas the volatility seen in the Manufacturer forecast is harder to correct systematically — a point we will address in the stability discussion below.

These results underscore an important trade-off: joint forecasting is most effective for high-volume items and when evaluated against its own downstream signal. When compared to demand orders, improvements persist but are less dramatic. The high MAPE values reflect the effect of dividing by small actual quantities. Nevertheless, the 'Joint AI' scenario shows the strongest overall point-estimate performance, indicating that incorporating downstream data indeed enhances demand visibility.

To make the results easy to interpret, we grouped forecasts into three categories: highly reliable (>80% accuracy), reliable (60–80% accuracy) and non-predictable (<60% accuracy). Here, accuracy simply

Table 4
Absolute error metrics (demand).

Dataset	TAE	MAE	TSE (bias)	ME	RMSE
Manufacturer / Solo forecast / manufacturer's shipments	128 950.00	494.06	24 010.00	91.99	726.17
Manufacturer AI / Solo AI forecast / manufacturer shipments	148 555.40	569.18	72 144.20	276.41	1 175.28
Joint AI / retailer outbound	117 099.20	448.66	14 482.80	55.49	657.44

Table 5
Relative error metrics (demand).

Dataset	MAPE	sMAPE	WMAPE	R^2	MSLE
Manufacturer / Solo forecast / manufacturer's shipments	2 875.47	54.98	41.29	0.72	2.23
Manufacturer AI / Solo AI forecast / manufacturer shipments	3 915.10	51.75	47.56	0.27	2.17
Joint AI / retailer outbound	3 354.91	50.09	37.49	0.77	1.93

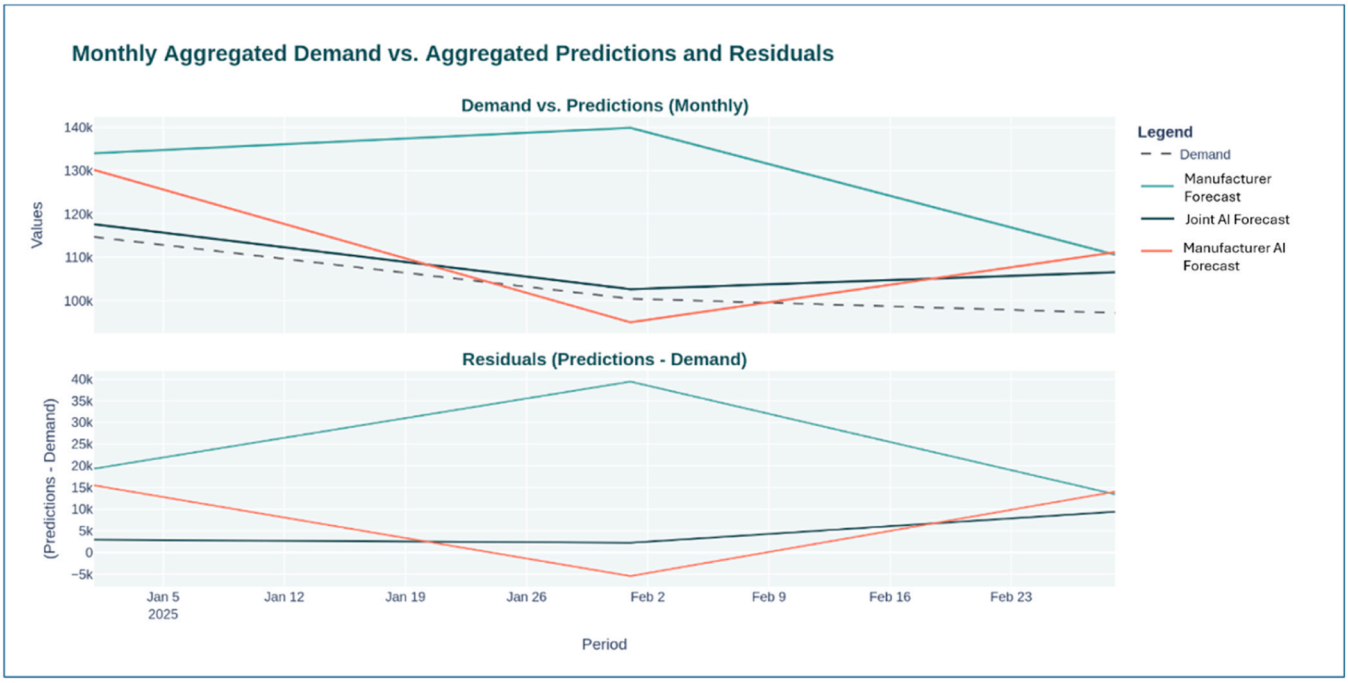

Fig. 3. Total predicted quantities.

means how close the forecast was to actual sales: for example, whether the model predicted 1000 units when 1050 were sold (close) or 600 (far off). Managers in practice use these categories to decide whether a forecast can be trusted for operational planning. To provide finer resolution, each of the "highly reliable" and "reliable" bands is further split into two sub-ranges (100–90% and 90–80% within "highly reliable"; 80–70% and 70–60% within "reliable"), so the figure displays five bars grouped into the three headline categories described above, plus the non-predictable category. Fig. 4 shows the results across two dimensions. The categories along the bottom represent the three accuracy ranges. The height of the bars reflects the share of total sales volume that falls into each category, emphasizing improvements where they matter most: for high-volume products. An error on a small, niche SKU may affect a few pallets, but the same error on a top-selling item can disrupt promotions, overload transport or tie up millions in working capital.

The message from the results is clear: Both manufacturer-only

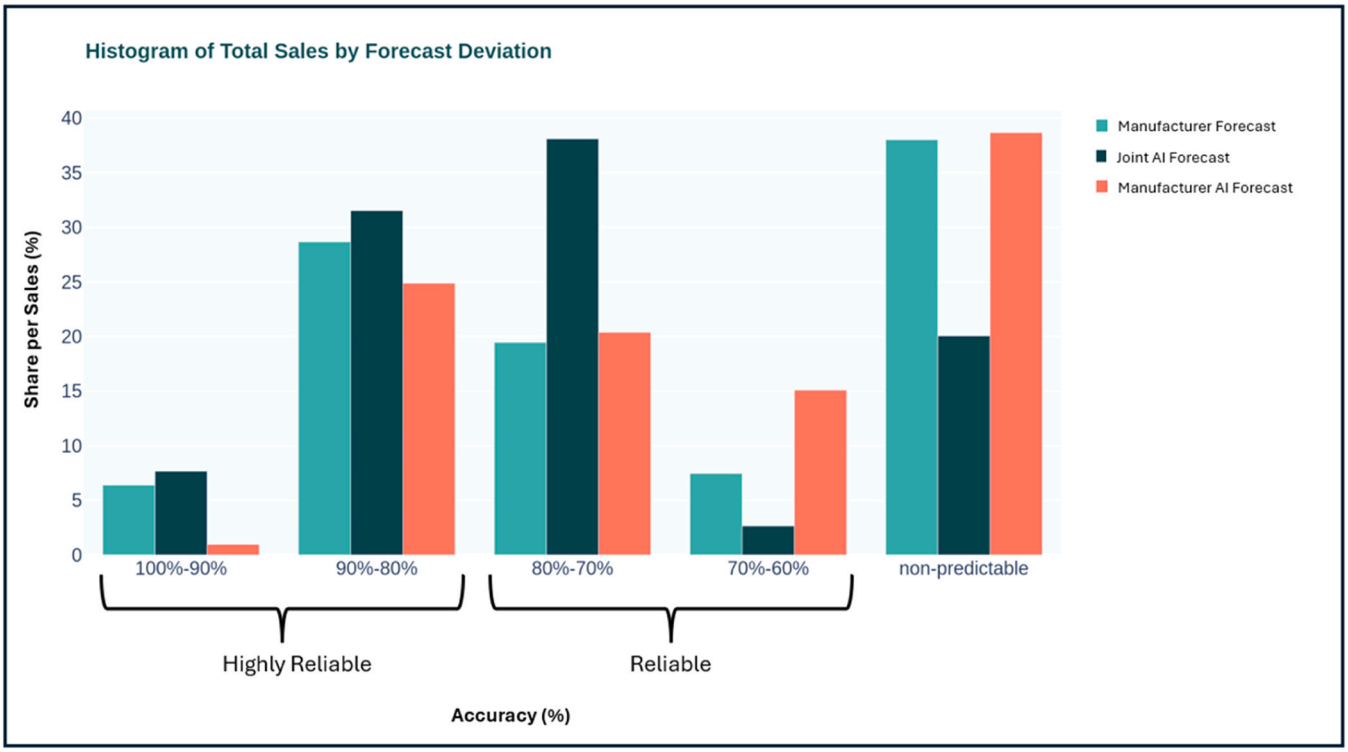

Fig. 4. Forecast accuracy by sales volume, categorized into highly reliable, reliable & non-predictable.

forecasts left the majority of sales volumes below 60%, meaning they could not be relied upon for operational decisions. In contrast, Joint AI did not merely cross this benchmark but decisively surpassed it. As Fig. 4 shows, approximately three-quarters of all sales volumes were forecast with more than 70% accuracy, with almost 40% reaching the highly reliable category range above 80%. For practitioners, this means forecasts that are not just mathematically better, but usable in day-to-day decision-making.

In quantitative terms, the improvement is substantial, but its interpretation depends on the evaluation level. On the test-data basis, Joint AI reduces total absolute error by 44.7% relative to Manufacturer AI, and this effect is statistically robust (cluster-bootstrap 95% CI: 27.8–57.3%). On the stricter common-demand benchmark, the corresponding reduction relative to Manufacturer AI remains positive and statistically robust at 21.2% (95% CI: 4.4–35.7%). At the individual SKU level, however, the advantage is directional rather than statistically significant: paired Wilcoxon tests show that Joint AI improves 56.3% of SKUs relative to Manufacturer AI, but the effect does not reach conventional significance ($p = 0.166$). This pattern supports Proposition 3: the collaborative advantage is not a uniform per-SKU shift, but a volume-concentrated effect that is strongest where sales volumes, inventory exposure and operational consequences are highest.

The largest improvements emerged in high-volume SKUs, which consistently moved into the reliable and highly reliable categories. This is critical because these products dominate overall sales and logistics flows; improving their forecast accuracy has disproportionate impact on production planning, safety stock and transport efficiency. By contrast, low-volume or highly volatile SKUs remained more difficult to predict, even with Joint AI. While errors were reduced, forecasts for these items often stayed below 60%. This underlines the need for complementary strategies such as safety stocks, manual overrides or specialized models to manage unpredictability in the long tail of the product portfolio.

4.2. Statistical validation and robustness checks

4.2.1. Data basis and replication

The inferential analysis uses the same SKU-level data underlying Tables 2, 4 and 5. Reconstructing the manuscript's aggregation convention reproduced the reported descriptive metrics to the reported decimal. All inferential tests therefore operate on the same data used in the descriptive results.

4.2.2. Bootstrap and paired-test inference

The first two intervals exclude zero: both the 44.7% reduction and the stricter common-benchmark reduction of 21.2% are statistically robust volume-weighted effects. The third interval includes zero: on the common demand benchmark, the improvement over the Manufacturer baseline is positive in point estimate but not statistically distinguishable from no improvement at the 95% level.

At the level of individual SKUs, the advantage is directional (median per-SKU error difference −10.6 and −22.1 units; Joint AI better on 56–59% of SKUs) but does not reach conventional significance. Taken together with Table 6, the finding is precise and instructive: the collaborative advantage is not a uniform per-SKU shift but a volume-concentrated effect — large where volumes are large, negligible or absent in the long tail. This is exactly the pattern Proposition 3 predicts and Fig. 4 of the manuscript displays; the inferential analysis quantifies it rather than contradicting it. Appendix A.2 provides supplementary visual evidence for this pattern by plotting the paired per-SKU mean absolute error differences between Joint AI and Manufacturer AI.

Table 6
Cluster-bootstrap 95% confidence intervals for volume-weighted TAE reductions (5000 replications; SKUs resampled).

Comparison (TAE reduction of Joint AI)	Point	95% CI
vs. Manufacturer AI — test-data basis (Table 2)	**44.7%**	**[27.8%, 57.3%]**
vs. Manufacturer AI — common demand benchmark (Table 4)	**21.2%**	**[4.4%, 35.7%]**
vs. Manufacturer baseline — common demand benchmark (Table 4)	9.2%	[−1.9%, 19.3%]

Table 7
Paired per-SKU Wilcoxon signed-rank tests.

Comparison (per-SKU mean absolute error)	W	p (Holm)	r (effect)	SKUs improved
Joint AI vs. Manufacturer AI	1625	0.166	0.151	56.3%
Joint AI vs. Manufacturer	1587	0.166	0.171	58.6%

4.2.3. Visual bootstrap evidence

Fig. 5 visualizes the cluster-bootstrap distribution for the common-demand benchmark comparison between Joint AI and Manufacturer AI. The distribution is centred clearly above zero, reinforcing the statistical robustness of the 21.2% volume-weighted reduction.

5. Discussion and implications

The results provide clear evidence that collaboration is a primary driver of forecasting performance improvements, particularly in volume-weighted terms. The improvement is consistent across evaluation metrics and evaluation targets, but the per-SKU analysis shows that the effect is not uniformly distributed across the portfolio; rather, it is concentrated in high-volume products, in line with Proposition 3. In this section, we interpret these findings and articulate the implications of the study to the forecasting, supply chain collaboration and AI-enabled operations literature.

The first implication of this study is the reconceptualization of forecasting as an interorganizational capability rather than a firm-internal function. Traditional forecasting research implicitly treats demand prediction as a task that can be optimized within organizational boundaries through better models, more computing power, or improved internal data [1,32,33]. Our findings challenge this assumption and show that even advanced AI models trained on manufacturer shipment data fail to deliver substantial improvements over traditional statistical methods. In contrast, when the same AI architecture is trained on downstream retailer data through a collaborative arrangement, forecast accuracy, stability and operational relevance improve markedly. This pattern suggests that the primary bottleneck in forecasting is not computational capability, but data boundary constraints imposed by organizational silos. This aligns with and extends research on supply chain collaboration by demonstrating that information sharing fundamentally alters the quality of demand signals, rather than merely refining existing ones [34,35].

The second implication concerns the role of artificial intelligence in forecasting. Much of the recent literature portrays AI as a disruptive force capable of overcoming long-standing forecasting limitations through superior pattern recognition and nonlinear modelling [36,37]. Our findings provide a more nuanced understanding as the manufacturer AI scenario demonstrates that AI alone does not resolve forecasting deficiencies when trained on distorted upstream signals. Despite its ability to capture nonlinear patterns, the AI model remains constrained by the quality and timeliness of the data it receives. Only when AI is embedded in a collaborative forecasting arrangement, i.e. where downstream data reflects actual consumer pull, does its full potential materialize. This suggests that AI should be regarded not as a standalone forecasting solution, but as an amplifier of collaborative data structures [38]. We argue that, at least for the algorithmic approaches examined here, advanced analytics may magnify the value of shared data, but cannot compensate for structurally inferior information. This reframes the question this study can speak to from "Which algorithm?" to "Under which organizational conditions can a given algorithm perform

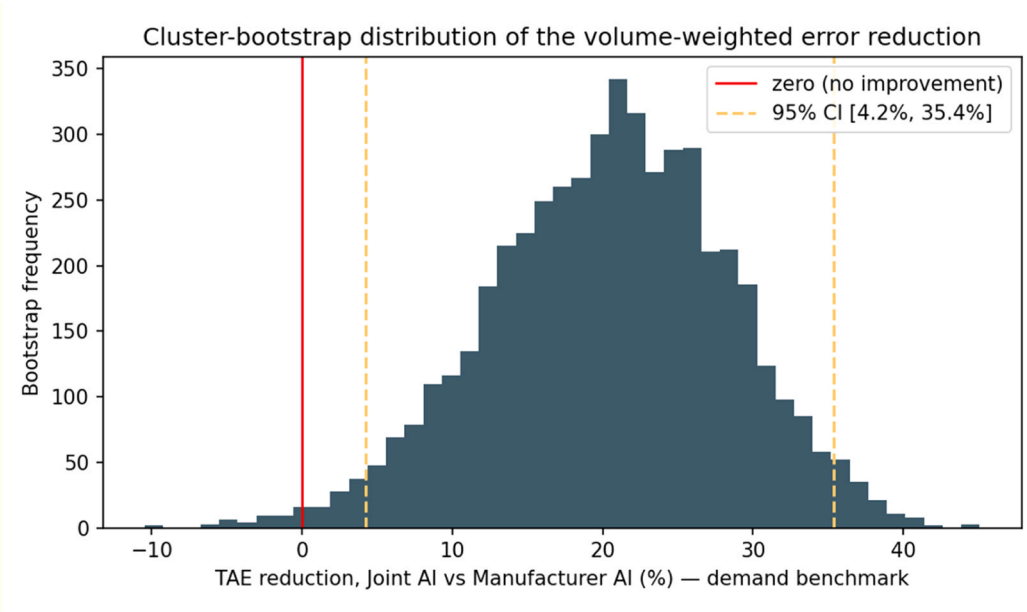

Fig. 5. Cluster-bootstrap distribution of the volume-weighted TAE reduction (Joint AI vs. Manufacturer AI, demand benchmark). The 95% interval [4.4%, 35.7%] excludes zero.

effectively?", while leaving open whether more sophisticated architectures (e.g., deep learning models) trained on manufacturer-only data could narrow the gap we observe.

The third implication emerges from the volume-weighted analysis of forecast performance. Prior forecasting research often evaluates models using average error metrics, implicitly assuming that all forecast errors are equally consequential [39,40]. Our results challenge this assumption by showing that the benefits of joint forecasting are disproportionately concentrated in high-volume SKUs, which dominate operational risk and cost. Under solo forecasting regimes, high-volume products remain largely in low-reliability categories, exposing firms to inventory imbalances, stockouts, and inefficient capacity utilization. In contrast, joint AI forecasting shifts a substantial share of high-volume sales into higher reliability ranges. This finding highlights that collaboration reshapes not only average accuracy, but also the economic distribution of forecasting errors, thereby reframing forecasting performance as a value-weighted phenomenon, where improvements matter most at the core of the product portfolio (it may also help to explain why collaborative forecasting can deliver outsized operational benefits even when improvements in conventional error metrics appear moderate).

The fourth implication is the focus on forecast stability as a critical performance dimension. Beyond reducing average errors, the joint AI forecast also produces more stable residuals over the forecast horizon (Fig. 3), which is suggestive of lower error variance across time and products, though we do not quantify this with formal stability metrics here. This stabilizing effect is particularly important in FMCG environments, where frequent replanning and firefighting erode operational efficiency [39,41]. From a theoretical standpoint, stability enhances the trustworthiness of forecasts and provides a more reliable foundation for decision-making. This insight highlights and emphasizes the role of stability as a core performance outcome alongside accuracy, in particular in collaborative contexts where multiple actors depend on shared forecasts.

The fifth implication reframes joint forecasting as an organizational and governance accomplishment, i.e. the performance gains observed here depended on institutional arrangements that made collaboration viable: neutral third-party facilitation, formal data-sharing agreements, and competition-law safeguards that enabled the partners to share downstream signals without relinquishing control over ordering decisions. This resonates with work emphasizing that voluntary data sharing across organizational boundaries hinges on governance structures that distribute risk and build trust [13,42], and it extends that literature by showing that such arrangements are not merely contextual preconditions but active determinants of forecasting performance — forecasting quality emerges as a joint outcome of technology, data, and governance [17,43].

This governance lens surfaces three considerations that temper and qualify the benefits documented above. First, collaboration carries costs that a purely technical account overlooks: investment in data-sharing infrastructure, neutral administration, and legal safeguards, alongside the higher computational burden of retraining machine-learning models on a rolling basis relative to a traditional ARIMA baseline. These must be weighed against the operational savings joint forecasting enables — reduced safety stock, fewer stockouts, and less costly replanning — and while a full cost-benefit quantification lies beyond our scope, the magnitude of the accuracy and stability gains for high-volume SKUs suggests the balance is favorable for the core of the portfolio, though this remains an empirical question for future research. Second, the benefits are likely conditional on organizational readiness. The dyad studied here had already established a dedicated forecasting function and normalized data-sharing through an industry working group (ECR Austria); firms with less mature planning processes or weaker analytical capabilities may realize smaller gains, or require a longer adjustment period, consistent with the view that AI-enabled collaboration is as much an organizational capability as a technical one [10]. The results should therefore be read as evidence of what joint AI forecasting can achieve under reasonably mature conditions, rather than a guarantee that collaboration alone suffices regardless of starting capabilities.

Third, data sharing introduces risks that constitute an ongoing governance requirement rather than a one-time setup cost. Sharing granular downstream data exposes the retailer to the possibility that competitively sensitive information — store-level sales patterns or promotional response — could be used beyond its intended forecasting purpose, or could shift negotiating leverage toward the party better able to exploit it. Reliance on a neutral intermediary mitigates but does not eliminate this exposure and introduces a dependency of its own, since the arrangement's viability rests on sustained trust in the intermediary's neutrality and data-handling [42]. The willingness of firms to deepen collaboration is thus likely to depend on how credibly these risks are governed over time — a dynamic that positions trust not as a static

precondition but as something continuously renewed through institutional design [13].

6. Conclusion

This study set out to examine whether joint AI-based forecasting can outperform non-collaborative forecasting approaches in an FMCG supply chain. Using a comparative research design that contrasts traditional statistical forecasting, solo AI-based forecasting and joint AI forecasting, the findings demonstrate a clear pattern: Joint AI delivers the strongest performance overall, with the most robust inferential support in volume-weighted comparisons against Manufacturer AI. Forecasting performance improves decisively when advanced analytics are embedded in collaborative, interorganizational data structures. In contrast, within the scope of the two algorithmic approaches examined here - a classical statistical method (ARIMA) and a gradient-boosting machine-learning model (XGBoost) - algorithmic sophistication alone was insufficient to overcome the structural limitations imposed by isolated, upstream demand signals. We do not claim that this finding extends to all possible algorithmic approaches; rather, it demonstrates that for a widely used and representative class of ML models, data provenance was the binding constraint.

Across the evaluation, Joint AI performs strongest overall, with statistically robust volume-weighted improvements over Manufacturer AI on both the test-data and common-demand benchmarks. The comparison with the traditional Manufacturer baseline is more nuanced: Joint AI shows a lower point estimate on the common-demand benchmark, but this difference is not statistically established at the 95% level.

Interestingly, our analysis shows that the performance gains are not evenly distributed across the product portfolio. High-volume SKUs benefit disproportionately from joint forecasting, shifting from low-reliability to high-reliability prediction ranges. These products dominate sales volumes, logistics flows and inventory exposure, meaning that improvements in their forecast accuracy translate into outsized operational and financial benefits. Solo forecasting challenges remain concentrated in low-volume and highly volatile items, underscoring that collaboration cannot fully resolve uncertainty for these segments. Beyond point accuracy, the residual evidence also points toward more stable joint AI forecasts (Fig. 3), suggesting reduced error variance across time and products. This stabilizing effect can be seen as relevant for operational planning, where volatile forecast performance can undermine trust and lead to costly replanning even when average errors appear acceptable.

The results, however, have to be viewed in the light of their limitations. First, the empirical scope of the study is necessarily narrow. The analysis focuses on a single manufacturer-retailer dyad within the FMCG sector, covering 87 SKUs from a single product category (savory snacks), and the out-of-sample evaluation covers a single three-month window (January–March 2025), even though the models were trained on 33 months of data spanning multiple seasonal cycles. While this context offers a demanding testbed, the narrow product scope and single evaluation window mean the findings may not generalize to portfolios with different demand characteristics (e.g., higher perishability, longer replenishment cycles, or greater seasonal variation), nor has the reported advantage of Joint AI been tested across other seasons or across multiple, non-overlapping evaluation periods. Future studies could examine whether similar patterns emerge in other industries, product categories, and multi-tier supply chains, and could evaluate performance across repeated, rolling out-of-sample windows spanning full annual cycles and multiple years to establish whether the advantage observed here is stable over time or specific to this setting.

Second, the research design bounds the strength and generality of the causal claim. The study compares only two algorithmic approaches — ARIMA and XGBoost, chosen to represent classical statistical and machine-learning paradigms — so the conclusion that collaboration matters more than algorithmic choice should be interpreted within this scope; it remains an open question whether more sophisticated architectures trained on manufacturer-only data (e.g., deep learning models capable of extracting weaker signals from noisier inputs) could partially close the observed gap. Moreover, as an observational rather than randomized comparison, the study cannot entirely rule out that factors correlated with data source — such as differences in the granularity or noise characteristics of the manufacturer and retailer data streams — contribute to the observed differences. We mitigate this concern by holding the algorithm and feature set constant and by evaluating all scenarios against a common ground truth (Section 4.1, Table 4), where Joint AI's advantage persists; nonetheless, a fully controlled design — for instance, feeding the same downstream signal into all approaches while varying only the collaborative-access mechanism — together with tests across a broader range of algorithms, would provide stronger causal identification and establish how far the finding generalizes.

Third, although the revised analysis reports cluster-bootstrap confidence intervals and paired Wilcoxon tests, analytical robustness remains bounded by the proprietary nature of the forecasting platform. Specifically, we could not conduct feature-level interpretability analyses, such as SHAP values or gain-based importance scores, nor controlled ablation analyses of hyperparameter settings, training-window length or feature selection. The reported inferential tests therefore establish robustness for the observed error reductions under the implemented models, but not robustness to all possible model configurations. Future research with access to platform internals could quantify feature contributions, test sensitivity to modelling choices and compare the collaborative-data effect across a broader range of model architectures. Relatedly, while the forecast-stability metrics are defined in Appendix A.1, the stability advantage posited in Proposition 4 is evidenced here qualitatively, through the residual patterns in Fig. 3, rather than through a separate quantitative stability comparison across the three scenarios; such a dedicated stability analysis is a natural extension of the present work.

Fourth, this study deliberately isolates the technical forecasting effect of collaboration and therefore leaves the surrounding organizational, economic, and behavioral conditions largely unexamined. It does not provide a formal cost-effectiveness or return-on-investment analysis, as data on platform fees, implementation costs, and administrative overhead were not disclosed; nor does it examine how power asymmetries, incentive alignment, and contractual arrangements shape the willingness to share data, or how organizational capabilities and planning maturity moderate the benefits realized. Both partners here had already established dedicated forecasting functions and prior experience with structured data-sharing, leaving open whether firms with less mature planning processes would realize comparable benefits or whether a capability threshold must first be reached. Understanding how trust, resistance, and organizational routines interact with AI-enabled collaboration likewise remains an important avenue for theory development. Future research could combine forecasting performance with firm-level cost data and examine planning maturity, incentive structures, and behavioral dynamics as moderating variables across multiple dyads with varying organizational readiness.

Within the scope of the single dyad, product category, and two algorithmic approaches examined here, our findings suggest that collaboration, rather than algorithmic sophistication alone, was the more decisive lever for forecast performance. Embedding AI-based forecasting within an interorganizational data-sharing arrangement allowed this manufacturer-retailer partnership to move from an unreliable, siloed forecasting practice toward a more stable and operationally relevant one. Whether this pattern extends to other industries, partnerships, and algorithmic approaches remains an open question for future research. We hope that this study sparks ideas and enables discussions to further advance the accuracy and collaboration in the sphere of AI-enabled demand forecasting.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Weisz Eric Ryan: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Lorenzo Bruno Prataviera:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **David M. Herold:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Sebastian Kummer:** Writing – review & editing, Supervision. **Maria Fernandez Pinar:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Markus Bures:** Validation, Methodology, Data curation.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the author(s) used ChatGPT (OpenAI) to support language refinement, structural suggestions, and clarity of expression. All outputs were critically reviewed and edited by the author(s), who take full responsibility for the content of the publication. While AI may have helped shape sentences, it had no role in shaping the arguments, thus any remaining errors, questionable judgments or moments of academic overconfidence are entirely the fault of the authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.sca.2026.100235.

References

[1] M.Z. Babai, J.E. Boylan, B. Rostami-Tabar, Demand forecasting in supply chains: a review of aggregation and hierarchical approaches, Int. J. Prod. Res. 60 (1) (2022) 324–348.
[2] M.J. Marzolf, J.W. Miller, S.T. Peinkofer, Retail & wholesale inventories: A literature review and path forward, J. Bus. Logist. 45 (1) (2024) e12367.
[3] J.T. Mentzer, J. Schroeter, Integrating logistics forecasting techniques, systems, and administration: the multiple forecasting system, J. Bus. Logist. 15 (2) (1994) 205.
[4] M. Barratt, A. Oliveira, Exploring the experiences of collaborative planning initiatives, Int. J. Phys. Distrib. & Logist. Manag. 31 (4) (2001) 266–289.
[5] L.B. Prataviera, A. Creazza, F. Dallari, M. Melacini, How can logistics service providers foster supply chain collaboration in logistics triads? Insights from the Italian grocery industry, Supply Chain Manag. Int. J. 28 (2) (2023) 242–261.
[6] T. Boone, B. Fahimnia, R. Ganeshan, D.M. Herold, N.R. Sanders, Generative AI: Opportunities, challenges, and research directions for supply chain resilience, Transp. Res. Part. E Logist. Transp. Rev. 199 (2025) 104135.
[7] L. Condé, C. Münch, Resilient by Design: Exploring the Social Abilities and Actor-Network Roles of Artificial Intelligence in Supply Chain Management, J. Bus. Logist. 46 (4) (2025) e70032.
[8] E. Weisz, D.M. Herold, S. Kummer, Revisiting the bullwhip effect: how can AI smoothen the bullwhip phenomenon? Int. J. Logist. Manag. 34 (7) (2023) 98–120.
[9] A.M. Geske, D.M. Herold, S. Kummer, Using sustainable technology to drive efficiency: Artificial intelligence as an information broker for advancing airline operations management, Sustain. Technol. Entrep. 4 (3) (2025) 100111.
[10] A.-M. Nitsche, B. Franczyk, C.-A. Schumann, K. Reuther, A decade of artificial intelligence for supply chain collaboration: Past, present, and future research agenda, Logist. Res. 17 (1) (2024).
[11] L.B. Prataviera, N. D'souza, C.L. Charlton, Exploring artificial intelligence for third-party logistics service providers: a dynamic capabilities perspective, Int. J. Phys. Distrib. & Logist. Manag. (2026) 1–35.
[12] M.A. Mediavilla, F. Dietrich, D. Palm, Review and analysis of artificial intelligence methods for demand forecasting in supply chain management, Procedia CIRP 107 (2022) 1126–1131.
[13] E. Weisz, D.M. Herold, N.K. Ostern, R. Payne, S. Kummer, Artificial intelligence (AI) for supply chain collaboration: implications on information sharing and trust, Online Inf. Rev. 49 (1) (2025) 164–181.
[14] S. Bevilacqua, A. Ferraris, R. Kozel, Z. Vincurova, Exploring the artificial intelligence integration in top management team decision-making: an empirical analysis, Bus. Process. Manag. J. (2025).
[15] Q. He, Z. Yang, Generative AI-driven knowledge management in manufacturing firms: a five-stage framework for dynamic knowledge optimization and digital innovation, J. Knowl. Manag. (2025) 1–21.
[16] H.L. Lee, V. Padmanabhan, S. Whang, The bullwhip effect in supply chains, MIT Sloan Manag. Rev. 38 (1997) 93–102.
[17] T. Paul, N. Islam, S. Rakshit, S. Mondal, A. Jeyaraj, The Impact of Artificial Intelligence (AI)-Enabled Collaborative Approach: Achieving Sustainable Supply Chain Performance, J. Bus. Logist. 46 (4) (2025) e70030.
[18] C.D. Smith, J.T. Mentzer, User influence on the relationship between forecast accuracy, application and logistics performance, J. Bus. Logist. 31 (1) (2010) 159–177.
[19] R. Carbonneau, K. Laframboise, R. Vahidov, Application of machine learning techniques for supply chain demand forecasting, Eur. J. Oper. Res. 184 (3) (2008) 1140–1154.
[20] X. Wan, P.T. Evers, Supply chain networks with multiple retailers: A test of the emerging theory on inventories, stockouts, and bullwhips, J. Bus. Logist. 32 (1) (2011) 27–39.
[21] R.G. Richey Jr, S. Chowdhury, B. Davis-Sramek, M. Giannakis, Y.K. Dwivedi, Artificial intelligence in logistics and supply chain management: A primer and roadmap for research, J. Bus. Logist. 44 (4) (2023) 532–549.
[22] H. Bousqaoui, I. Slimani, S. Achchab, Comparative analysis of short-term demand predicting models using ARIMA and deep learning, Int. J. Electr. Comput. Eng. (2021). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104386900&doi=10.11591%2fijece.v11i4.pp3319-3328&partnerID=40&md5=b89f9cd47bf733d256abb5af0fb709d6.
[23] D. Preil, M. Krapp, Artificial intelligence-based inventory management: a Monte Carlo tree search approach, Ann. Oper. Res. (2022). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104876328&doi=10.1007%2fs10479-021-03935-2&partnerID=40&md5=c7e6d7dd2afab6b33b080847e20993c8.
[24] B. Pecar, B. Davies, A new technology paradigm for collaboration in the supply chain, Int. J. Serv. Oper. Inform. (2007). https://www.scopus.com/inward/record.uri?eid=2-s2.0-42149190535&doi=10.1504%2fIJSOI.2007.015330&partnerID=40&md5=4eeb81eab454c941ebee0bb48bef0cc1.
[25] J. Van Belle, T. Guns, W. Verbeke, Using shared sell-through data to forecast wholesaler demand in multi-echelon supply chains, Eur. J. Oper. Res. 288 (2) (2021) 466–479.
[26] U. Nweje, M. Taiwo, Leveraging Artificial Intelligence for predictive supply chain management, focus on how AI-driven tools are revolutionizing demand forecasting and inventory optimization, Int. J. Sci. Res. Arch. 14 (1) (2025) 230–250.
[27] T. Stank, L.W. Saunders, A. Scott, C.W. Autry, T.L. Esper, Theory will take you only so far"(Nolan, 2023): In search of greater insight through quantitative, observation-based research, J. Bus. Logist. 45 (3) (2024) e12383.
[28] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
[29] C. Mandl, S. Nadarajah, S. Minner, S. Gavirneni, Data-driven storage operations: Cross-commodity backtest and structured policies, Prod. Oper. Manag. 31 (6) (2022) 2438–2456.
[30] Circly. (2025). *How it works*. https://www.circly.at/en/wie-es-funktioniert/.
[31] ECR. (2025). *ECR Arbeitsgruppe joint forecasting*. https://ecr-austria.at/arbeitsgruppen/joint-forecasting/.
[32] A. Jain, Sharing demand information with retailer under upstream competition, Manag. Sci. 68 (7) (2022) 4983–5001.
[33] S. Taghiyeh, D.C. Lengacher, A.H. Sadeghi, A. Sahebi-Fakhrabad, R.B. Handfield, A novel multi-phase hierarchical forecasting approach with machine learning in supply chain management, Supply Chain Anal. 3 (2023) 100032.
[34] G. Scarton, N. Benini, M. Formentini, AI-enhanced demand forecasting: an organizational information processing view, Int. J. Phys. Distrib. & Logist. Manag. (2025) 1–24.
[35] Y.-C. Tsao, Y.-K. Chen, S.-H. Chiu, J.-C. Lu, T.-L. Vu, An innovative demand forecasting approach for the server industry, Technovation 110 (2022) 102371.
[36] R. Brau, J. Aloysius, E. Siemsen, Demand planning for the digital supply chain: How to integrate human judgment and predictive analytics, J. Oper. Manag. 69 (6) (2023) 965–982.
[37] E. Revilla, M.J. Saenz, M. Seifert, Y. Ma, Human–artificial intelligence collaboration in prediction: A field experiment in the retail industry, J. Manag. Inf. Syst. 40 (4) (2023) 1071–1098.
[38] C. Baah, D. Opoku Agyeman, I.S.K. Acquah, Y. Agyabeng-Mensah, E. Afum, K. Issau, D. Ofori, D. Faibil, Effect of information sharing in supply chains: understanding the roles of supply chain visibility, agility, collaboration on supply chain performance, Benchmark. Int. J. 29 (2) (2022) 434–455.
[39] T. Aichner, V. Santa, Demand forecasting methods and the potential of machine learning in the fmcg retail industry. Serving the customer: The role of selling and sales, Springer, 2023, pp. 215–252.
[40] K. Chaowai, P. Chutima, Demand forecasting and ordering policy of fast-moving consumer Goods with promotional sales in a small trading firm, Eng. J. 28 (4) (2024) 21–40.

[41] D. Adebanjo, R. Mann, Identifying problems in forecasting consumer demand in the fast moving consumer goods sector, Benchmark. Int. J. 7 (3) (2000) 223–230.
[42] I. Susha, B. Rukanova, A. Zuiderwijk, J.R. Gil-Garcia, M.G. Hernandez, Achieving voluntary data sharing in cross sector partnerships: Three partnership models, Inf. Organ. 33 (1) (2023) 100448.
[43] S. Guo, J. Zhang, Y. Wang, W. Hu, F. Wei, D. Bai, Forecasting information sharing strategies in competitive smart connected platforms, Electron. Commer. Res. (2024) 1–47.