



RESEARCH ARTICLE

From forecast accuracy to inventory performance: Assessing machine-learning models for intermittent export demand

Asmuliardi Muluk^{1*}, Taufik Nazir¹, Stephany Christie Lase¹ and Dika Aulia Ardi²

¹ Department of Industrial Engineering, Faculty of Engineering, Universitas Andalas, Padang, Indonesia

² PT Digdaya Solusi, Padang, Indonesia

*Corresponding Author. Email: asmuliardi@eng.unand.ac.id (A. Muluk)

Article Information	Abstract
Received: 5 August 2026 Accepted: 7 October 2026 Published: 20 October 2026 AMS 2020 Classification: 90B05, 62M20, 68T05	Machine-learning forecasts are usually judged by accuracy, yet procurement decisions consume forecasts through an inventory policy whose lead times, review cycles, and lot sizes can overturn accuracy rankings. This paper quantified that value in replenishment decisions for an Indonesian cassia exporter's intermittent export demand. Ten approaches — the trading partner's company-described ex-ante annual plan, intermittent-demand estimators, ARIMA, and three machine-learning configurations — were embedded in a sequential simulation of the company's periodic-review policy over 55 months of grade-level demand, with a three-month lead time, lot-size constraints, and safety stocks calibrated to each method's realised errors. Each method was charged its cheapest configuration meeting a common achieved-fill requirement. Across sixteen demand units the plan and a Tweedie gradient-boosting model were not separated: the latter cost between 10% more and 4% less across four requirements, and the other eight alternatives at least 9% more, narrowing to 1% under a common warm-up. Resampling at the 0.98 requirement separated only two alternatives from the plan. Because the plan can carry the buyer's purchasing intentions, this shows no added value of history-based machine learning over such a plan in this portfolio, not a general absence of value. Separating the buffer rule from the forecast mattered more than the choice of model: replacing the fixed percentage mark-up with an error-calibrated buffer raised aggregate fill from 0.84 to 0.99 at unchanged cost, and from 0.75 to 0.99 at a fifth less. Policy simulation at a common service requirement should be the model-selection gate for intermittent-demand inventory systems. Keywords: Intermittent demand, demand forecasting, inventory simulation, periodic review policy, machine learning, forecast value

1. Introduction

Raw-material procurement in export-oriented agro-industry is particularly difficult when procurement lead times are long, order quantities are constrained by shipping lots, and demand is intermittent, with zero-demand months interrupted by large transactions. The company examined in this study, an Indonesian cassia cinnamon exporter referred to as PT XYZ for confidentiality, must commit to raw-material purchases three months before demand materialises. Orders are placed in lots of approximately twenty-five tonnes using an annual demand plan supplied by an overseas trading partner. Because the plan provides only yearly totals for each product grade, the monthly distribution of demand remains unknown when purchasing decisions are made.

A common response is to replace such a coarse annual plan with a statistical or machine-learning (ML) forecasting model. ML-based demand forecasting has been applied to multi-SKU retail settings [1], supply chains for time-critical products [2][3], logistics operations [4], and models incorporating exogenous variables such as exchange rates and search-interest indices [5]. Reviews of this literature, however, identify a recurring limitation: model performance is usually established through accuracy metrics calculated on hold-out samples [6][7], while improvements in operational decisions are often assumed rather than demonstrated [8]. This limitation has long

been recognised in the intermittent-demand literature. Previous studies have shown that forecast-accuracy rankings can differ from the inventory outcomes generated by the forecasts [9][10] and have argued that forecasting methods should be evaluated through their effects on inventory performance [11][12]. Nevertheless, empirical studies that feed competing forecasts through a company's actual inventory policy, including its lead time, review period, and lot-size constraints, remain uncommon, particularly for intermittent demand in emerging-economy export settings.

This gap has practical consequences. Forecast errors generate asymmetric and policy-dependent costs [13], and a method that performs well according to mean absolute percentage error (MAPE) may still increase inventory cost once its errors interact with safety-stock calculations and lot-size rounding. The problem is more severe for intermittent series because MAPE becomes problematic when actual demand is zero [14][15]. Specialised methods such as Croston's estimator [16], the Syntetos–Boylan approximation (SBA) [17], and the Teunter–Syntetos–Babai method (TSB) [18] were developed to address characteristics of intermittent demand that conventional forecasting and evaluation approaches do not handle well [19].

This study evaluates the relationship between forecast accuracy and inventory value using fifty-five months of grade-level export demand for two cassia product categories and the trading partner's actual ex-ante annual forecasts for four consecutive years. Because the partner is also the buyer, its annual plan can reflect purchasing intentions that no model fitted to the demand history can observe, so the comparison is in part one between information sets as well as between methods. Ten forecasting approaches — the partner's annual plan, statistical and intermittent-demand benchmarks, and ML configurations including a zero-inflation-aware gradient-boosting model — are embedded in a sequential simulation of the company's periodic-review (R, S) replenishment policy. Each method receives safety stock calibrated to its own realised forecast-error distribution. The study addresses three research questions:

RQ1. Does higher forecast accuracy translate into lower total inventory cost at a common achieved-service requirement once the (R, S) policy, three-month lead time, and lot-size constraints are taken into account?

RQ2. How does the relationship between accuracy and inventory value vary across the intermittency spectrum of the demand portfolio and with the level of aggregation at which forecasts are consumed?

RQ3. Which accuracy metrics, if any, rank forecasting methods consistently with their simulated inventory cost?

The study makes three contributions. First, it provides a reproducible evaluation protocol that combines rolling-origin forecasting, error-calibrated safety stocks, sequential policy simulation over an observable horizon, comparison under a common achieved-service requirement rather than at a common service target, and cost-ratio sensitivity analysis. This protocol measures forecast value directly in an intermittent-demand setting with realistic procurement constraints. Second, the study presents evidence from sixteen evaluated demand units, drawn from a portfolio that spans the full Syntetos–Boylan–Croston classification plane. The results show that ML forecasting improves materially on the annual plan in no segment of the portfolio and that an annual plan supported by an error-calibrated buffer is a stronger benchmark than is commonly assumed in forecasting applications. Third, the study quantifies the weak relationship between standard accuracy metrics and inventory outcomes, with implications for how companies select forecasting models.

The remainder of the paper is organised as follows. Section 2 reviews the related literature. Section 3 describes the case and the data. Section 4 presents the forecasting methods, simulation design, and evaluation metrics. Section 5 reports the experimental results. Section 6 discusses the implications and limitations, and Section 7 concludes the paper.

2. Literature review

2.1. Forecasting intermittent demand

Intermittent demand series, in which many periods record no demand, do not satisfy the assumptions underlying conventional smoothing and ARIMA-type methods. Croston [16] separated the estimation of demand sizes from the estimation of the intervals between demands. Syntetos and Boylan [17] identified a bias in Croston's estimator and proposed the corrected SBA variant. Teunter, Syntetos and Babai [18] reformulated the problem around demand probability, producing the TSB method, which also links forecasting to obsolescence risk. Boylan and Syntetos [19] provide a comprehensive treatment of these methods. Empirical work on large spare-parts datasets indicates that measurable differences between these estimators do not translate directly into differences in inventory performance [9][18].

Machine learning has entered this area along two routes. The first is the global cross-learning model, trained on pooled data from many series, which performed well on the intermittent-heavy M5 retail competition data [20]. The second is the application study, in which gradient boosting, random forests, or hybrid architectures are fitted to a single company's demand using engineered features and exogenous variables [1][4][5]; Seyam et al. [2] review this body of work across time-critical industries. Recent applications often combine a classification stage, which predicts whether demand occurs, with a regression stage, which predicts its size. This structure mirrors the model

level the size-interval decomposition that Croston introduced at the estimator level. The present study includes such a hybrid model together with a gradient-boosting model that uses a Tweedie objective, a loss function intended for zero-inflated and right-skewed targets and the objective of the winning entry in the M5 competition [20].

2.2. From accuracy to inventory value

The limitations of accuracy as a proxy for operational value have been documented repeatedly. Syntetos, Nikolopoulos and Boylan [10] proposed evaluating forecasting competitions through accuracy-implication metrics such as stockholding and service outcomes, and reported that the two rankings diverge. Petropoulos, Wang and Disney [11] simulated inventory performance for the methods used in the M3 competition and reached a similar conclusion on a larger scale. Sanders and Graman [13] quantified the cost of forecast error in a warehouse setting and showed that the cost surface is asymmetric with respect to bias. Kourentzes, Trapero and Barrow [12] optimised forecasting models directly against inventory objectives rather than accuracy criteria. On the measurement side, Hyndman and Koehler [14] introduced scaled errors because percentage errors are unstable when demand is zero, and Kolassa [15] showed that evaluating point forecasts with an inappropriate functional can reward degenerate forecasts. The results reported below illustrate this effect for adjusted MAPE applied to intermittent series.

Three elements remain uncommon in this literature and motivate the present study. The first is the use of a company's actual ex-ante forecast as the benchmark rather than a statistical benchmark constructed by the researcher. The second is a demand portfolio that covers the intermittency spectrum, which allows the location of forecast value to be identified. The third is an inventory policy simulated with the operational constraints that practitioners face, including a long lead time, a two-month review period, and lot-size rounding. Studies of forecast-driven decision frameworks in other operational settings, such as prediction-driven staffing [21][22], report the same general point: value is realised at the decision layer rather than in the accuracy table. The growing interest in autonomous, agent-based replenishment [23] makes this question more relevant, because an autonomous ordering loop inherits any gap between accuracy and value.

The studies closest to the present one can be set against it one by one along five dimensions: the setting, the methods compared, the benchmark, the number of series, and the inventory policy through which forecasts are judged.

(i) Syntetos, Nikolopoulos and Boylan [10] study the fast-moving items of a pharmaceutical company. They compare statistical forecasts with the judgementally adjusted forecasts that planners derived from them, the statistical forecast serving as the benchmark, and judge both by the inventory volumes and service levels they imply. The present study examines intermittent series instead, and its benchmark is the company's own ex-ante plan.

(ii) Petropoulos, Wang and Disney [11] use 334 monthly industry series from the M3 competition, which are fast-moving, and compare eleven statistical methods ranging from the naive method and exponential smoothing to ARIMA, the Theta method and combinations, without intermittent-demand estimators or machine learning. Their policy is an order-up-to rule reviewed every month, with lead times of 1, 3, 6 and 12 months, a safety stock proportional to the standard deviation of forecast errors, target service levels of 90%, 95% and 99%, backlogging, and no lot sizing.

(iii) Sanders and Graman [13] simulate a labour-intensive warehouse with cost data from a case study. Rather than comparing forecasting methods, they vary the bias and the standard deviation of forecast errors and measure the resulting organisational cost, which in that setting arises from labour rather than from a replenishment policy; bias proves the costlier component.

(iv) Kourentzes, Trapero and Barrow [12] use 229 weekly series of 173 observations from a UK manufacturer of household cleaning and personal hygiene products. They estimate a local-level exponential smoothing model with a conventional loss and with inventory-based losses, the conventional estimation being the benchmark, and evaluate an order-up-to policy with lead times of 3 and 5 weeks, service levels of 90%, 95% and 99%, and lost sales; the inventory-based losses give forecasts whose accuracy is worse by up to 9% and whose bias is lower by up to 62%.

(v) Ramaekers and Janssens [24] simulate intermittent demand for a single item, with occurrence following a Markov process with probabilities of 0.2 and 0.4 and gamma-distributed sizes, in a 16-point fractional factorial design. They compare Croston's method, simple exponential smoothing and a four-period moving average under (R, s, S) and (R, s, Q) policies with a lead time of one period, optimise the policy parameters jointly with the forecasting method by tabu-search simulation-optimisation, and find that the best method depends on the policy.

Against these, the present study uses real intermittent demand from one company: sixteen grade-level series spanning the intermittency spectrum, ten methods including three machine-learning configurations, the company's

own ex-ante annual plan as the benchmark, and the company's (R, S) policy with a two-month review period, a three-month lead time, and lot rounding to shipping lots.

3. Case description and data

PT XYZ, an anonymised Indonesian exporter, ships cassia cinnamon in two product categories. Broken cassia consists of crushed bark and is shipped in container lots of approximately twenty-five tonnes. Stick cassia consists of whole quills and is shipped in smaller mixed lots. Raw material is procured from farm-gate suppliers approximately three months before export shipment ($L = 3$ months), and the purchasing plan is reviewed every two months ($R = 2$ months). The company's current planning input is an annual forecast for each product grade, supplied by its trading partner before the plan year begins and containing no monthly information. Safety stock is currently set as a fixed percentage mark-up, between 10% and 30%, on ordered quantities.

The dataset covers June 2020 to December 2024, comprising 55 months of realised export quantities recorded by product grade and destination. These records were aggregated to the eighteen grade-level demand units at which the partner states its annual forecasts, consisting of six broken grades and twelve stick grades. Two stick units record no demand in any month of the horizon. The records do not show why: one grade carries a zero annual plan in every year, and the other carries a plan of three tonnes for 2024 only, against which nothing was shipped. Both are grades in the partner's list that were not traded in the period rather than gaps in the data. They are retained in the description of the portfolio but are excluded from every metric and cost aggregation, because scaled error measures are undefined for a constant-zero history. Their exclusion does not affect the representativeness of the volume-weighted results, since the two units carry no demand; adding them to the cost aggregation moves no cost index by more than 0.7 points. All results below are therefore computed on sixteen units.

Two distinct series were transcribed from the company's planning documents: the partner's annual forecast for each grade and year, and the annual realised total stated in the same documents. Only the second series was verified against the transaction records, as a transcription check: seventy of the seventy-two unit-year totals matched exactly, and the two remaining discrepancies were resolved in favour of the transaction records. Forecast values were never altered, and no realised quantity was substituted for a forecast at any point in the analysis. The plans are described by the company as issued before each plan year, and three properties of the transcribed forecasts are consistent with that description: only two of the fifty-seven unit-year pairs with non-zero activity coincide with the realised total; the median absolute percentage deviation between plan and realisation is 33%; and the plan understates realised demand at the portfolio level in every year, by between 12% and 44%. A plan revised with knowledge of realisations would not display this systematic, one-sided deviation. The limitation that remains is documentary rather than statistical: the company retains the plans but not a dated issue record for each one, so publication timestamps could not be recovered. The benchmark is therefore described throughout as the company-described ex-ante plan, and its ex-ante status rests on the company's account of its planning process together with the statistical evidence above, not on an immutable record. Because the partner is also the buyer, the plan can carry purchasing intentions that are contained neither in the demand history nor in the exogenous series available to the other nine methods. The comparison is therefore partly one between information sets: it asks whether monthly models built on history can match or add to an annual plan informed by the buyer's intentions, not only whether one forecasting technique is better than another.

Figure 1 shows the sixteen evaluated units on the Syntetos–Boylan–Croston (SBC) classification plane [25], using the average inter-demand interval (ADI) and the squared coefficient of variation of non-zero demand sizes (CV^2) computed on the pre-2024 history. The portfolio covers the full range of the plane. It contains one large smooth unit and one large erratic unit, which together account for 65% of 2024 demand, two lumpy units, and twelve intermittent units with demand in 51% to 2% of months. Broken cassia accounts for 92% of the 2,605 tonnes demanded in 2024. This spread allows RQ2 to be examined within a single company, although two segments contain only one unit each, which bounds the strength of any segment-level claim.

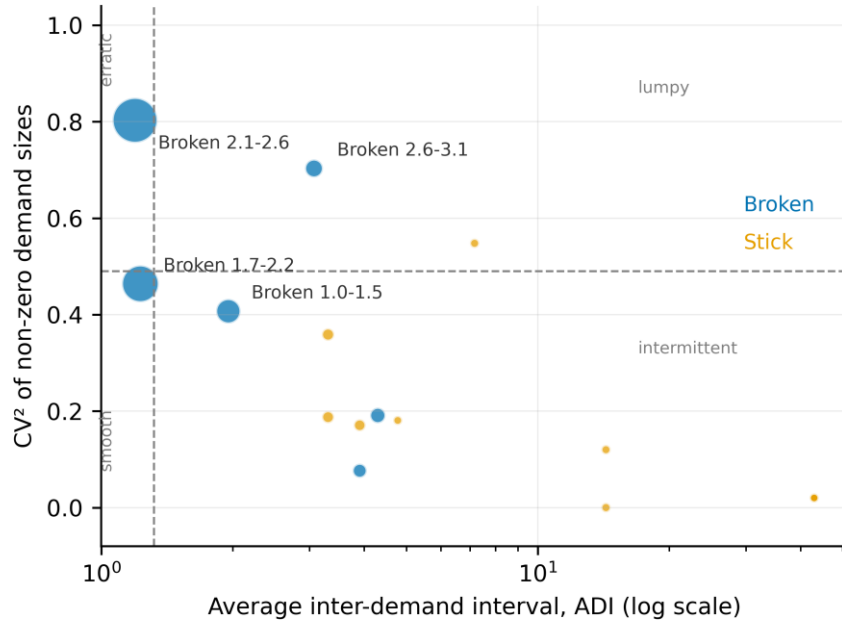


Figure 1. The 16 evaluated demand units on the SBC classification plane (bubble area proportional to mean monthly demand; dashed lines at $ADI = 1.32$ and $CV^2 = 0.49$). The two units with no demand in the horizon are omitted, since ADI and CV^2 are undefined for them.

4. Methodology

Figure 2 summarises the evaluation pipeline. Rolling-origin forecasting supplies both an accuracy evaluation and a sequential (R, S) simulation in which each method's safety stock is calibrated to its own realised errors. The two evaluations are then compared with each other.

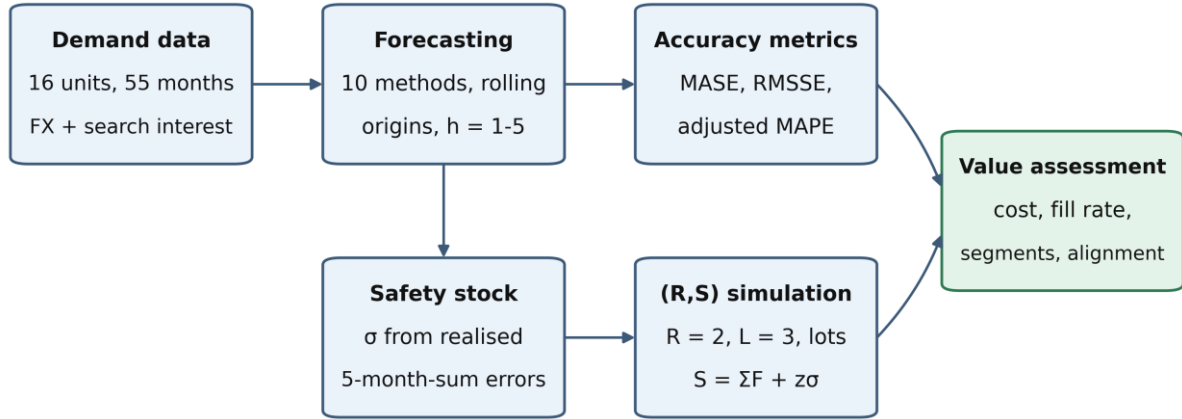


Figure 2. Overview of the evaluation pipeline: rolling-origin forecasting feeds an accuracy evaluation and a sequential (R, S) simulation with per-method error-calibrated safety stocks.

4.1. Forecasting methods and the information set

Ten forecasting methods are compared, and each is paired with two buffer rules, giving the policy set described in Section 4.3. The incumbent partner-annual benchmark distributes the partner's ex-ante annual forecast uniformly across the twelve months of the target year, which is the most faithful monthly interpretation of a plan that contains no intra-year information. Naive repeats the last observed month. SES is single exponential smoothing with the smoothing constant fitted by minimising one-step squared error on the available history. The intermittent-demand specialists are Croston [16], SBA [17], and TSB [18], implemented with the smoothing parameter fixed at 0.1, within the ranges recommended in the literature reviewed by Kourentzes [26], 0.1 to 0.3 following Croston and 0.05 to 0.2 following Syntetos and Boylan; Section 5.7 reports the sensitivity of the results to this choice. Croston and SBA are initialised with the first non-zero demand as the demand size and the number of months up to that demand as the inter-demand interval, and both estimates are updated at every later demand; TSB updates its occurrence probability every month and its demand size at every demand, with the same smoothing constant for

both. ARIMA selects the orders (p, d, q) , with $p \leq 2$, $q \leq 2$ and $d \leq 1$, by AIC at every origin. Three ML configurations are trained globally across all units on pooled data: RF (random forest), XGB-Tweedie (gradient boosting with a Tweedie loss, variance power 1.3; a power between 1 and 2 corresponds to a compound Poisson–gamma distribution, which places a probability mass at zero on a right-skewed size distribution, the data type for which the objective was effective in the M5 competition [20]), and Hybrid, a two-stage classifier-regressor in which the estimated occurrence probability multiplies the conditional-size regression. The architecture, loss function, and hyperparameters of all three were fixed before the 2024 test window was examined, and Section 5.7 reports variance powers from 1.1 to 1.9; they are listed in full in the released configuration files.

The information set is defined at the level of the individual feature rather than by calendar year. At an origin t , no input dated later than month $t - 1$ enters any model. Forecasts are issued for the five months t to $t + 4$, which is the protection interval implied by $L + R$. The machine-learning forecasts are direct: a separate model is trained for each horizon $j = 0, \dots, 4$, and for a target month $m = t + j$ its demand features are the lags $m - j - 1$ to $m - j - 4$, a three-point moving average of the months $m - j - 1$ to $m - j - 3$, and a six-month non-zero rate ending in month $m - j - 1$. The most recent input is therefore dated $t - 1$ at every horizon, which is the information the local methods use at the same origin. The exogenous variables, the monthly IDR/USD exchange rate and three Google Trends search-interest indices, are lagged by the same amount, so their most recent value is also dated $t - 1$; both series are published within days of the end of their reference month, so this lag is consistent with their release calendars. Local methods are re-fitted at every origin. The ML models are annual vintages, which reflects realistic retraining practice, and the vintage is selected from the forecast origin rather than from the target month: a model fitted on data up to 31 December of year Y issues the forecasts of every origin in year $Y + 1$, whatever month those forecasts target. Selecting the vintage from the target year would let an origin late in one year use a model trained to the end of that year, which is a look-ahead path; an automated test asserts that the latest training observation precedes every origin it serves, and that the most recent feature of every direct model is dated $t - 1$, and it is run as part of the released pipeline. Section 5.7 reports the effect of the annual schedule itself, and of a shared design in which one model serves the horizons $j \leq 3$ with lags from $m - 4$. All forecasts are truncated at zero.

4.2. Accuracy metrics

Let $y_{u,m}$ denote demand for unit u in month m and $f_{u,m}^{(t)}$ the forecast issued at origin t . The unit of evaluation is the triple (unit, origin, horizon); each target month in the outcome window, January to November 2024, is therefore scored once per origin from which it is forecast, and horizons are pooled unless stated otherwise. Let $\mathcal{E}_u$ be the set of scored pairs (t, m) for unit u , and let $n_u = 43$ be the number of months of pre-2024 history, June 2020 to December 2023, on which the scaling denominators are estimated. Scaled errors follow Hyndman and Koehler [14]:

$$\text{MASE}_u = \frac{\frac{1}{|\mathcal{E}_u|} \sum_{(t,m) \in \mathcal{E}_u} |f_{u,m}^{(t)} - y_{u,m}|}{\frac{1}{n_u - 1} \sum_{i=2}^{n_u} |y_{u,i} - y_{u,i-1}|} \quad (1)$$

$$\text{RMSSE}_u = \sqrt{\frac{\frac{1}{|\mathcal{E}_u|} \sum_{(t,m) \in \mathcal{E}_u} (f_{u,m}^{(t)} - y_{u,m})^2}{\frac{1}{n_u - 1} \sum_{i=2}^{n_u} (y_{u,i} - y_{u,i-1})^2}} \quad (2)$$

RMSSE is the root-mean-square counterpart of MASE. Because the order-up-to level is set from the forecast of a five-month sum, MASE is also computed on five-month sums, with the same denominator multiplied by five so that the two are commensurate:

$$\text{MASE}_u^{(5)} = \frac{\frac{1}{|\mathcal{T}_u|} \sum_{t \in \mathcal{T}_u} \left| \sum_{j=0}^4 (f_{u,t+j}^{(t)} - y_{u,t+j}) \right|}{\frac{5}{n_u - 1} \sum_{i=2}^{n_u} |y_{u,i} - y_{u,i-1}|} \quad (3)$$

where $\mathcal{T}_u$ is the set of origins whose five target months all lie in the outcome window. Bias is reported as the mean error. For comparability with the original company study [27], the adjusted MAPE used there is also reported:

$$\text{aMAPE}_u = \frac{100}{|\mathcal{E}_u|} \sum_{(t,m) \in \mathcal{E}_u} a_{u,m}^{(t)} \quad (4)$$

where $a_{u,m}^{(t)} = |f_{u,m}^{(t)} - y_{u,m}| / y_{u,m}$ when $y_{u,m} \geq \varepsilon$, and otherwise $a_{u,m}^{(t)} = 1$ if $f_{u,m}^{(t)} > \varepsilon$ and 0 if not, with $\varepsilon = 0.1$ kg, so that a zero-demand month scores 100% when it receives a positive forecast and 0% when it does not. It is included as the incumbent measurement practice examined in RQ3, not as a recommended metric. All scaled measures are undefined when the in-sample denominator is zero, which is the case for the two all-zero units; those units are excluded, leaving $n_{\text{valid}} = 16$ for every accuracy summary.

4.3. Inventory policy and simulation

Each pair of demand unit and policy is simulated in monthly steps, with reviews every second month. At a review in month t the order-up-to level is

$$S_t = \sum_{j=0}^4 f_{u,t+j}^{(t)} + B_t \quad (5)$$

where B_t is the safety buffer. Two buffer rules are compared. The error-calibrated rule sets

$$B_t = z \sigma_t, \quad \sigma_t = \sqrt{\frac{1}{|\mathcal{W}_t|} \sum_{s \in \mathcal{W}_t} e_s^2}, \quad e_s = \sum_{j=0}^4 (f_{u,s+j}^{(s)} - y_{u,s+j}) \quad (6)$$

where $\mathcal{W}_t$ holds the last twelve origins $s \leq t - 5$, so that only error windows completed before t are used, and z is the buffer multiplier, $z = \Phi^{-1}(0.95) = 1.645$ for the common 95% cycle-service target. When fewer than six windows exist, σ_t falls back to the standard deviation of historical five-month demand sums. Because the mean error is generally non-zero, σ_t is a root-mean-square rather than a standard deviation, and it therefore absorbs bias as well as dispersion; Section 5.2 reports the two components separately. The fixed-markup rule reproduces the company's current practice, $B_t = \gamma \sum_{j=0}^4 f_{u,t+j}^{(t)}$ with $\gamma \in \{0.10, 0.20, 0.30\}$, and is applied to the partner-annual forecast.

The order quantity is

$$q_t = Q_u \text{ceil}\left(\frac{\max(0, S_t - \text{IP}_t)}{Q_u}\right) \quad (7)$$

where IP_t is the inventory position, stock on hand plus stock on order minus backorders, and Q_u is the shipping lot, 25,025 kg for broken and 20 kg for stick; the order arrives after the three-month lead time. Demand is filled from stock and unmet demand is backordered. Costs are normalised: the holding cost h is 1 per kg-month on ending on-hand inventory, and the backorder cost b applies per kg-month to ending backorders, with the ratio $b:h$ swept over 2, 5, 10, 20, and 50 [13]. The total cost over the outcome window $\mathcal{O}$ is

$$C = \sum_u [\sum_{m \in \mathcal{O}} (h I_{u,m} + b K_{u,m}) + k h (I_{u,m_{\text{end}}} + P_{u,m_{\text{end}}})] \quad (8)$$

where $I_{u,m}$ and $K_{u,m}$ are the ending on-hand stock and backorders of unit u in month m , $P_{u,m_{\text{end}}}$ is the pipeline standing at the last month m_{end} of the window, and k is the terminal charge in months described below, zero in the primary configuration.

The design separates the window in which decisions are made from the window over which their consequences are counted. Reviews fall on odd months and the protection interval is five months, so with demand recorded to December 2024 the last review whose entire protection interval is observed is July 2024. The primary configuration therefore places no order at a review after July 2024, and accumulates cost and service over January to November 2024, the month in which the July decision's protection interval closes. Each method operates the policy from the first review in January 2023, starting from an on-hand stock equal to its own order-up-to level at that review, and the 2023 reviews form a warm-up whose costs are not counted. The stock entering 2024 is therefore the product of each method's own 2023 decisions, and because orders arrive three months after they are placed, the outcomes of January to March 2024 are determined by reviews held up to November 2023. The decisions judged in the primary configuration are therefore those of January 2023 to July 2024, and each has a fully observed consequence. Two checks bound the effect of the warm-up: operating every method on the annual plan throughout 2023, so that the state entering 2024 is common and every difference in outcome arises from decisions taken in 2024, and counting outcomes from April 2024, the first month in which an order placed in 2024 can arrive. Two further sensitivities bound what remains: extending the outcome window to December, and charging a terminal holding cost of k months on the on-hand inventory and pipeline standing at the end of the window, with $k \in \{0, 2.5, 5\}$. The terminal charge is a sensitivity to the value of leftover stock, not a substitute for an observable horizon.

Two comparisons of cost are made, and they answer different questions. The first fixes the buffer multiplier at a common cycle-service target and compares cost, which is how model selection is usually reported; because the calibrating σ absorbs bias and intermittent errors are not normal, a common target does not deliver a common achieved service, so this comparison is service-similar rather than service-matched. The second imposes a common service requirement instead: the multiplier z is swept from 0 to 3 in steps of 0.1, with the values 1.2816, 1.6449, 2.0537, 3.5 and 4 added, for every method, and each method is charged the cost of its cheapest configuration that attains at least a given achieved aggregate fill rate. The feasibility rule — reach the service requirement first, then compare cost — is fixed before any cost is compared, and it is this comparison on which the cost claims in this paper rest. The requirement is a floor rather than an exact match: because the sweep is discrete, the selected configurations land at different achieved fill rates, so the achieved fill and cycle service of every selected configuration are reported beside its cost and the residual service difference stays visible. A robustness check that

interpolates cost along each method's lower envelope at exactly the required fill is reported alongside. Pseudo-code for the simulation loop is provided in the Appendix, and the full implementation is released with the paper.

4.4. Value metrics and alignment analysis

For each configuration the analysis records total normalised cost, its holding and backorder components, the volume-weighted fill rate, the demand served from stock in the month it occurs divided by total demand, summed over units and months of the outcome window, and the achieved cycle service, defined as the proportion of replenishment cycles beginning in the evaluation window that pass without any shortage. Reporting both is necessary because the buffers are set from a cycle-service target while cost is driven by quantities served, and the two measures need not move together. RQ1 is addressed through the total-cost index across all valid units under a common achieved-service requirement, with the partner-annual benchmark set to 100, and the common-target comparison is reported beside it. Uncertainty is quantified by a bootstrap over units with 1,000 resamples that preserves each unit's full time path. The sixteen units are the portfolio the company operates rather than a sample drawn from a larger population, so this interval measures how much of the ordering depends on portfolio composition; it is not a population inference, and an interval that spans parity means the data do not resolve the direction of a difference rather than that two methods are equivalent. No equivalence margin was specified in advance and none is claimed. RQ2 disaggregates the index by SBC segment. For RQ3, the ten methods are ranked within each unit by each accuracy metric and by simulated total cost, using average ranks for ties, and Spearman's rank correlation is computed between the two orderings; units for which either ranking is constant are dropped and n_{valid} is reported with every summary. Costs are aggregated on a per-kilogram basis, which assumes a homogeneous holding cost and shortage consequence across grades; aggregate figures therefore weight units by volume, which is an operational rather than an economic weighting, and no unit-specific margin or holding rate was available to convert one into the other.

5. Experimental results

5.1. Forecast accuracy

Table 1 reports median accuracy across the sixteen valid demand units over the evaluation window. Three observations are relevant to the analysis that follows. First, the differences among the leading methods are small. XGB-Tweedie records the lowest median MASE at 0.52, followed by the naive method, SES, and TSB at 0.55 to 0.56, with the partner-annual plan at 0.68; SBA, Croston and the random forest record the highest values, 0.80 to 0.81. The gap between Croston or SBA and TSB, which share the smoothing constant, is concentrated in units whose demand stopped or thinned before or during the outcome window: Croston and SBA are updated only when a demand occurs and continue to forecast the last smoothed size-to-interval ratio through a run of zeros, whereas TSB updates its occurrence probability every month and decays [18]. On eight of the sixteen units Croston and TSB differ by less than 0.1 in MASE. Second, the ranking depends on the metric. On adjusted MAPE the naive method scores best on twelve of the sixteen units, which reflects the metric's treatment of zero-demand months rather than forecast quality [15]; on MASE the best-scoring method is XGB-Tweedie on eight units. Third, the ordering by accuracy bears little resemblance to the ordering by cost reported below, which is the substance of RQ3 [10][11].

Table 1. Median accuracy across the 16 valid demand units over the outcome window, January-November 2024.

Method	MASE	RMSSE	MASE (5-mo sums)	Adjusted MAPE (%)
XGBoost (Tweedie)	0.52	0.54	0.34	90.7
Naive	0.55	0.50	0.45	47.6
SES	0.55	0.34	0.43	92.6
TSB	0.56	0.34	0.38	94.5
ARIMA	0.61	0.36	0.43	91.8
Partner annual	0.68	0.44	0.37	89.1
Hybrid cls-reg	0.72	0.51	0.39	88.2
SBA	0.80	0.52	0.62	92.1
Croston	0.80	0.53	0.59	91.6
Random Forest	0.81	0.48	0.36	84.4

5.2. Cost at a common service target

Figure 3 reports total cost by method when every method is given an error-calibrated buffer at the same 95% cycle-service target. The partner-annual benchmark attains the lowest total cost, index 100, at an aggregate fill rate of 0.99 and a mean achieved cycle service of 0.97. XGB-Tweedie follows at 107, at a fill rate of 0.99 and a cycle

service of 0.95. The remaining methods cost between 1.7 and 3.9 times as much: SES 165, SBA 176, TSB 188, ARIMA 191, Croston 194, the random forest 207, the hybrid 209, and naive 388.

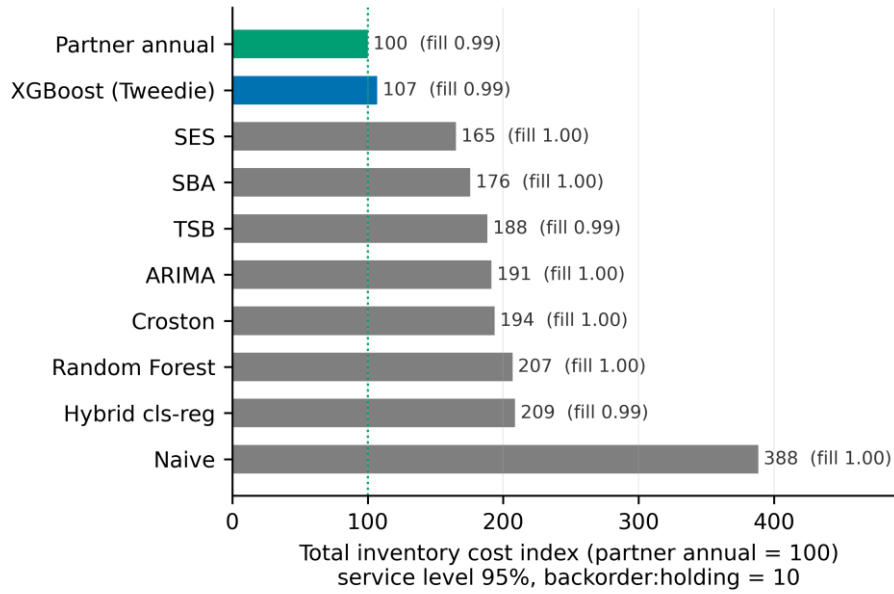


Figure 3. Total inventory cost index across the 16 valid units at a common 95% service target (partner annual set to 100) with volume-weighted fill rates; backorder-to-holding ratio 10, lot rounding on, decisions to July 2024 and outcomes counted to November 2024.

Table 2. Cost index at a common 95% service target (partner annual set to 100) across backorder-to-holding cost ratios; lot rounding on. *

<i>b:h</i>	Partner	XGBoost	SES	SBA	TSB	ARIMA	Croston	RF	Hybrid	Naive
2	100.0	104.1	167.8	180.5	191.0	196.5	198.9	212.3	212.0	399.1
5	100.0	105.2	166.8	178.7	190.0	194.5	196.9	210.3	210.8	395.0
10	100.0	107.0	165.1	175.7	188.3	191.3	193.6	207.0	208.8	388.4
20	100.0	110.3	161.8	170.1	185.1	185.3	187.4	200.7	205.0	375.8
50	100.0	119.3	153.2	155.2	176.7	169.2	171.0	184.1	195.0	342.6

* *Partner*: partner annual plan; *XGBoost*: XGBoost with a Tweedie objective; *Hybrid*: two-stage classifier-regressor; *RF*: random forest.

The two cheapest policies are also the two with the lowest achieved fill rates, 0.988 for the annual plan and 0.992 for XGB-Tweedie against 0.994 or more for every other method, and the mean achieved cycle service ranges from 0.94 to 1.00 across methods. A common buffer multiplier does not produce a common achieved service, because σ is a root-mean-square that absorbs bias and because intermittent errors are not normal. The comparison in this subsection is therefore service-similar, and it is reported for continuity with how model selection is usually presented rather than as the basis of any conclusion. Section 5.3 imposes the service requirement.

A bootstrap over units with 1,000 resamples places the XGB-Tweedie index at 107 with a 95% interval of [92, 140], which spans parity. Every other method's interval lies entirely above 100, the closest being SBA at [115, 255]. Because the sixteen units are the operating portfolio rather than a sample, that interval describes how much of the ordering depends on which units the portfolio contains. Spanning parity means the data do not resolve whether XGB-Tweedie is cheaper or more expensive than the annual plan at this operating point; it is not evidence that the two are equivalent.

The result is also sensitive to how the end of the horizon is treated. Charging a terminal holding penalty of 2.5 or 5 months on the on-hand inventory and pipeline standing at the end of the outcome window lowers the XGB-Tweedie index to 104.5 and 102.6, because the annual plan ends the window with more stock, while extending the outcome window to the end of December lowers it to 101.6. Across these treatments the index ranges from 102 to 107, so once decisions stop at the last fully observable review the result is not sensitive to how leftover stock is valued. The warm-up has a bounded effect of the same order. January to March account for 31% to 36% of each method's January–November cost. When every method is operated on the annual plan through 2023, the XGB-Tweedie index is 103.7 and every other index moves towards parity, because the costs of January to March then become common to all methods; the annual plan and XGB-Tweedie remain first and second, and SES is next at 140. Counting outcomes from April 2024 only gives an XGB-Tweedie index of 102.3, with SES next at 166.

The mechanism behind the ranking follows from how the buffers are set. Because each method's safety stock is calibrated to its own errors, all methods reach broadly similar service, and the cost differences arise from the stock each must hold. The annual plan distributes demand uniformly across months and is therefore often wrong about timing, but its errors over the five-month protection interval, which is the quantity that determines the buffer, are comparatively stable. Methods that track monthly variation accumulate larger protection-interval errors on this portfolio and hold more stock as a result. The naive method is the extreme case, because after a demand spike it projects that spike five months forward and holds a large surplus.

Decomposing the protection-interval errors identifies which component drives the buffers. For five of the ten methods, including the two leading ones, the root-mean-square error is dominated by its systematic part rather than by dispersion: the median ratio of absolute bias to RMS across units is 0.86 for the annual plan and 0.93 for XGB-Tweedie, while the naive method records 0.07 and is almost entirely dispersion. Both leading methods underforecast the five-month sum in the median unit, so their calibrated buffers correct a level offset as much as they absorb noise. Because σ is defined as a root-mean-square, this systematic component enters the safety stock directly, which is the mechanism by which a method with lower monthly error can still require more stock.

5.3. Cost under a common service requirement (RQ1)

Figure 4 addresses the service confound by sweeping the buffer multiplier for every method and plotting the cost index against the achieved aggregate fill rate. The partner-annual and XGB-Tweedie frontiers lie close together: the annual plan is the cheaper of the two at most fill rates, the two nearly touch around a fill of 0.98, and XGB-Tweedie is the cheaper only in a narrow band around 0.99, which is why it undercuts the plan at that requirement. The other eight methods sit above both, with SES and TSB closest over most of the range.

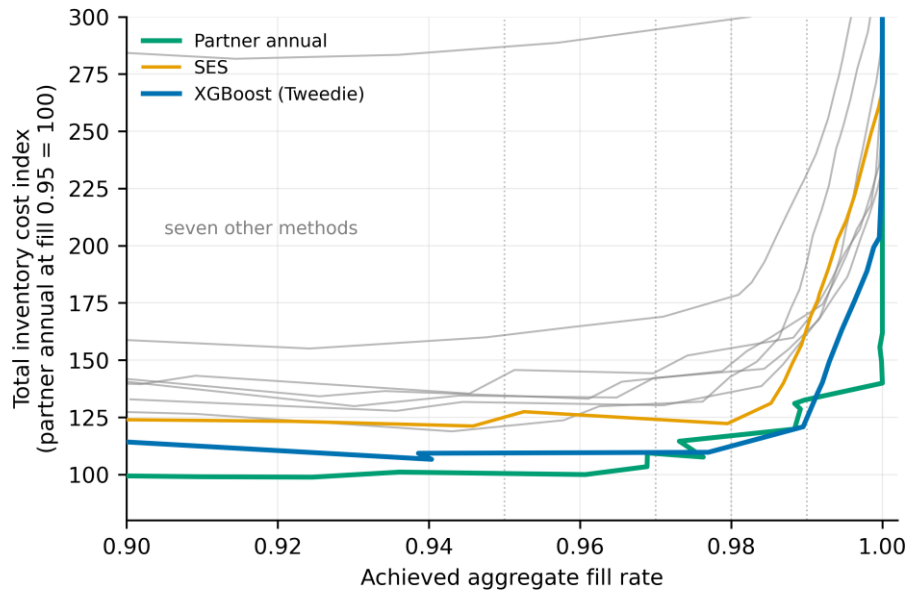


Figure 4. Service-cost frontier: total cost index against achieved aggregate fill rate as the buffer multiplier is swept from 0 to 4 for every method. The index is relative to the cost of the cheapest partner-annual configuration reaching a fill of 0.95, which is the reference of the first requirement in Table 3; Table 3 normalises each requirement by its own partner-annual cost, so the two scales coincide only at 0.95. Dotted lines mark the four service requirements reported in Table 3.

Table 3 reads the frontier at four service requirements. Charging each method the cost of its cheapest configuration that reaches at least the required fill, the annual plan and XGB-Tweedie are not separated: XGB-Tweedie costs 10% more at a fill of 0.95, 2% more and 1% more at 0.97 and 0.98, and 4% less at 0.99. Every other method costs at least 9% more at every requirement: SES 9% to 22% more, the others 16% or more, and the naive method 2.5 to 2.9 times as much. The multipliers needed to reach a given service differ substantially between methods, which is the quantitative form of the same point: the annual plan needs $z = 1.5$ to reach a fill of 0.98 while the random forest needs only 0.5, and the annual plan is still cheaper at that service because its protection-interval errors are smaller in absolute terms.

Table 3. Cost under a common achieved-fill requirement.*

Panel A. Required fill 0.95 and 0.97								
Method	0.95 index	0.95 fill	0.95 cycle	0.95 z	0.97 index	0.97 fill	0.97 cycle	0.97 z
Partner	100.0	0.961	0.891	1.1	100.0	0.976	0.906	1.3
XGBoost	109.7	0.977	0.875	1.4	102.0	0.977	0.875	1.4
SES	122.3	0.979	0.859	0.7	113.6	0.979	0.859	0.7
RF	130.3	0.971	0.906	0.4	121.0	0.971	0.906	0.4
SBA	133.7	0.965	0.844	0.7	135.8	0.984	0.875	0.9
TSB	123.7	0.958	0.844	0.4	122.5	0.976	0.906	0.6
ARIMA	133.1	0.961	0.875	0.5	134.9	0.979	0.891	0.7
Croston	144.3	0.970	0.844	0.7	141.2	0.974	0.844	0.8
Hybrid	164.4	0.959	0.922	0.6	157.0	0.971	0.922	0.7
Naive	288.7	0.957	0.906	0.5	278.8	0.983	0.922	0.6
Panel B. Required fill 0.98 and 0.99								
Method	0.98 index	0.98 fill	0.98 cycle	0.98 z	0.99 index	0.99 fill	0.99 cycle	0.99 z
Partner	100.0	0.988	0.938	1.5	100.0	1.000	0.984	1.8
XGBoost	100.8	0.990	0.906	1.5	96.2	0.991	0.938	1.6
SES	109.4	0.985	0.891	0.8	121.1	0.991	0.922	1.2
RF	115.7	0.984	0.922	0.5	120.5	0.992	0.953	0.8
SBA	121.9	0.984	0.875	0.9	119.7	0.992	0.906	1.1
TSB	124.5	0.983	0.906	0.8	137.7	0.990	0.969	1.2
ARIMA	128.3	0.982	0.891	0.8	124.5	0.992	0.938	1.0
Croston	133.3	0.988	0.891	0.9	121.8	0.991	0.906	1.0
Hybrid	149.0	0.981	0.922	0.8	171.6	0.991	0.969	1.4
Naive	250.3	0.983	0.922	0.6	268.6	0.992	0.984	1.0

* Each method is charged the cost of its cheapest configuration reaching at least the required fill (partner annual set to 100 at each requirement; panel A reports the requirements 0.95 and 0.97, panel B 0.98 and 0.99). The requirement is a floor, not an exact match, so the achieved fill, achieved cycle service and buffer multiplier z of each selected configuration are given beside its cost. An exact-match interpolation is in the released result tables. Partner: partner annual plan; XGBoost: XGBoost with a Tweedie objective; Hybrid: two-stage classifier-regressor; RF: random forest.

Because the requirement is a floor, Table 3 reports the achieved fill of each selected configuration so that the residual difference is visible rather than hidden by the word "matched". The residual matters most at the loosest requirement: at a required fill of 0.95 the annual plan settles at 0.961 while XGB-Tweedie and SES settle at 0.977 and 0.979, so part of the premium those two pay at that level buys service the plan does not deliver. At the 0.99 requirement the residual runs the other way: the annual plan reaches 1.000 against 0.991 for XGB-Tweedie, so part of the advantage of XGB-Tweedie at that level reflects service the plan delivers beyond the requirement. Interpolating cost along each method's lower envelope at exactly the required fill, which removes the residual at the cost of pricing configurations that were never simulated, gives the same picture: XGB-Tweedie is at 107.8 at a fill of 0.95, 104.9 and 100.9 at 0.97 and 0.98, and 89.0 at 0.99, SES at 111 to 122, and every other method at 115 or above.

Bootstrapping over units widens the intervals considerably. At a required fill of 0.98 the interval spans parity for XGB-Tweedie [63, 162], SES [81, 219], the random forest [86, 235], SBA [92, 218], TSB [83, 268], ARIMA [93, 215] and Croston [92, 228]; only the hybrid and the naive method have lower bounds above parity. The defensible reading is therefore that portfolio resampling does not resolve the ordering among the annual plan and seven of the nine alternatives at this requirement, while it does place the hybrid and the naive method consistently above them. The point estimates place the annual plan and XGB-Tweedie ahead of every other method at every requirement, but sixteen units, two of which carry most of the volume, are not enough to establish that ordering against resampling.

The warm-up affects the size of these premiums more than their sign. With every method operated on the annual plan through 2023, the annual plan is the least expensive policy at every requirement, but the premiums narrow because the costs of January to March become common to all methods: the nearest alternative costs 1% to 4% more, and it is TSB, ARIMA or XGB-Tweedie depending on the requirement. Counting outcomes from April 2024 with each method's own warm-up, XGB-Tweedie costs between 4% less and 6% more than the plan and every other method at least 12% more; combining the common warm-up with the April start leaves the annual plan the least expensive at every requirement, XGB-Tweedie being unable to reach a fill of 0.99 in that configuration. Across these treatments no alternative costs less than 95% of the annual plan at any requirement.

5.4. The buffer rule versus the forecast

The comparisons above equip every method, including the annual plan, with an error-calibrated buffer. That is not the company's current policy, which applies a fixed percentage mark-up. Simulating the actual policy separates the two interventions. With mark-ups of 10%, 20%, and 30% the annual plan attains cost indices of 146, 126, and 100, but at aggregate fill rates of only 0.67, 0.75, and 0.84 and achieved cycle service of 0.59, 0.75, and 0.77. The current policy is therefore not cheaper in any meaningful sense: it operates far below the 95% target. Within that set a larger mark-up lowers cost and raises service together, because at these buffer levels the backorder cost dominates: the 30% mark-up is both the cheapest and the most reliable of the three, and it still reaches a fill of only 0.84. Replacing the fixed mark-up with an error-calibrated buffer, holding the forecast unchanged, raises aggregate fill to 0.99: from 0.84 at the same cost as the 30% mark-up, or from 0.75 at a fifth less than the 20% mark-up. Table 4 reports the comparison. On this portfolio the buffer rule, not the choice of forecasting method, is the larger operational lever.

Table 4. The company's current fixed mark-up compared with an error-calibrated buffer, the forecast held constant at the partner annual plan; backorder-to-holding ratio 10, lot rounding on.

Buffer rule applied to the annual plan	Cost index	Aggregate fill rate	Mean cycle service
Fixed mark-up, 10%	146.5	0.67	0.59
Fixed mark-up, 20%	125.7	0.75	0.75
Fixed mark-up, 30%	100.2	0.84	0.77
Error-calibrated, 95% target	100.0	0.99	0.97

5.5. Where forecasting pays: the intermittency spectrum (RQ2)

Figure 5 disaggregates the cost index by SBC segment at the common service target. On the single large erratic unit, which is the largest in the portfolio and records demand in 84% of months with high variability in demand size, XGB-Tweedie attains an index of 100.5, level with the plan, at a fill rate of 1.00 against the incumbent's 0.97. The comparison on that unit moves with the treatment of the horizon: the same unit gives 110 and 118 under terminal penalties of 2.5 and 5 months, and 82 over the calendar year. On the twelve intermittent units XGB-Tweedie is level with the plan at 99, on the two lumpy units the closest method is XGB-Tweedie at 131, and on the single smooth unit every alternative is more expensive, from 141 upwards, mainly because 2024 demand fell below the historical pattern and level-tracking methods carried the surplus.

These are case observations rather than segment estimates. Two of the four segments contain a single unit, and the erratic result is one unit in one year whose sign depends on the treatment of the horizon. What the portfolio supports is a weaker but still useful statement: no forecasting method examined here improves materially on the annual plan in any segment; XGB-Tweedie is level with it on the erratic unit and across the intermittent tail, and more expensive on the lumpy and smooth units.

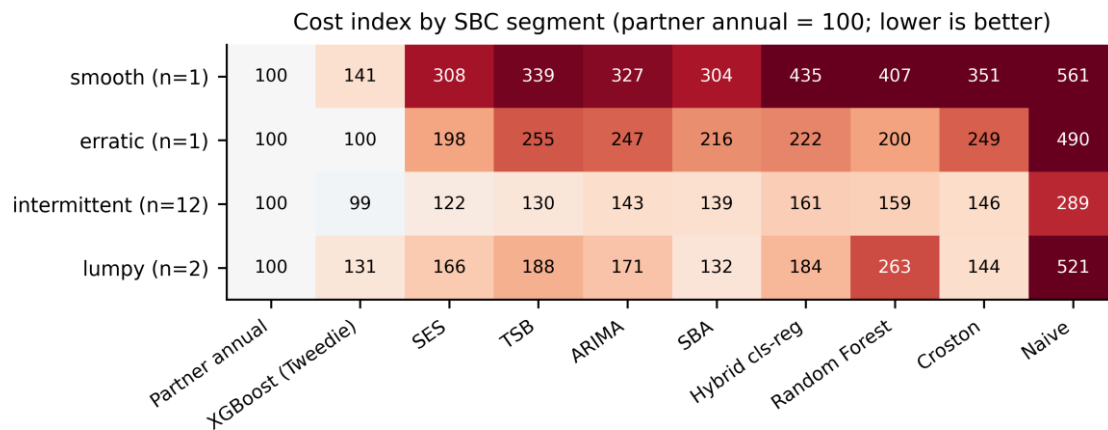


Figure 5. Cost index by SBC demand segment at a common service target (partner annual set to 100; lower is better). Two segments contain a single unit and are case observations rather than estimates.

5.6. Do accuracy metrics predict cost? (RQ3)

Figure 6 compares, within each unit, the ranking of methods by accuracy with their ranking by simulated cost, using average ranks for ties; all sixteen valid units yield a defined correlation for every metric. Scaled errors are weakly informative at the median, 0.41 for MASE and 0.23 for RMSSE, but the metric currently used in practice

is not: adjusted MAPE records a median Spearman correlation of -0.19 . Alignment is also weakest where the volume lies. On the two largest units, which carry 65% of 2024 demand and 71% of the pre-2024 mean monthly demand used as the weight, correlations across metrics range from -0.43 to 0.30 , and the volume-weighted mean correlations fall between -0.09 and 0.14 . Matching the metric's horizon to the protection interval, by computing MASE on five-month sums, does not help: its median is 0.15 and its volume-weighted mean -0.09 . On this portfolio no accuracy table, including those reported in this study, identifies the cost-minimising method for each unit; only the policy simulation does.

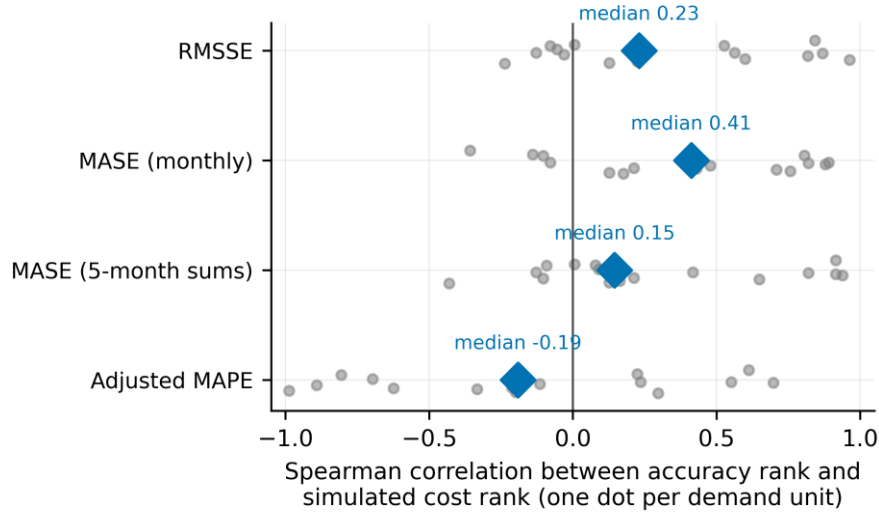


Figure 6. Within-unit Spearman correlation between method rankings by accuracy and by simulated total cost (one dot per unit; diamonds mark medians).

5.7. Sensitivity to design choices

Six further checks bound the influence of design decisions. Varying the backorder-to-holding ratio from 2 to 50 keeps the annual plan first and XGB-Tweedie second at every ratio, with the XGB-Tweedie index rising from 104 to 119; below them the order changes, TSB falling from fifth to seventh place at a ratio of 50 and the random forest and the hybrid exchanging places between ratios (Table 2). Varying the service target from 90% to 98% moves XGB-Tweedie from 97.5, below the plan, to 123; at the 90% target XGB-Tweedie reaches a fill of 0.94 against 0.97 for the plan, so there the lower cost again buys lower service. Lot rounding raises absolute costs by between 1.5% and 9.2%, and it falls hardest on the two cheapest policies, the annual plan and XGB-Tweedie, at 8.5 and 9.2% against 1.5% to 5.8% elsewhere; measured against every method other than XGB-Tweedie the incumbent's position would therefore be slightly stronger without the lot constraint.

The third check concerns the machine-learning design. Serving the horizons $j \leq 3$ from one model whose most recent input is dated $m - 4$, instead of the direct design, gives XGB-Tweedie 104 at the common target, the random forest 233 and the hybrid 219, against 107, 207 and 209 with the direct design; at the 0.98 requirement the shared design gives XGB-Tweedie 106 against 101. Training XGBoost with a squared-error objective on the same features and hyperparameters, which isolates the loss function from the algorithm family, gives an index of 190 at the common target and 130 at the 0.98 requirement, with a median MASE of 0.79 against 0.52 for the Tweedie objective. Tweedie variance powers from 1.1 to 1.7 give indices of 103 to 107 at the common target and 96 to 110 at the 0.98 requirement, and a power of 1.9 gives 130 and 121; the power of 1.3 used throughout is therefore not the most favourable tested setting. At the common target XGB-Tweedie remains the closest alternative to the annual plan at every power from 1.1 to 1.7, but under the fill requirements its position relative to the plan and to SES depends on the power.

The remaining three checks concern estimation. Repeating the machine-learning runs under three random seeds moves the cost index of XGB-Tweedie between 104 and 107 and of the random forest between 207 and 214, and leaves the hybrid unchanged at 209; this variation is small relative to the bootstrap interval reported above, leaves the annual plan and XGB-Tweedie first and second, and at most exchanges the random forest and the hybrid, which are within a few points of each other. The smoothing constant of the intermittent estimators matters more than the choice among them: raising α from 0.05 to 0.30 moves Croston from 265 to 163, SBA from 254 to 129, and TSB from 250 to 171. These methods therefore improve at α of 0.2 or 0.3, so the value used throughout, $\alpha = 0.1$, is not their best tested setting and the comparison is not favourable to them. It lies within the ranges recommended in the literature [26] and was not tuned on the evaluation window, as was every other method's configuration. Even at their best tested setting all three remain above both the annual plan and XGB-Tweedie, so the ranking does not depend on this choice. Finally, refitting the machine-learning models at every forecast origin instead of once a year

improves point accuracy but not inventory performance: the median MASE of XGB-Tweedie falls from 0.52 to 0.49 while its cost index rises from 107 to 111, the random forest moves from 207 to 205 and the hybrid from 209 to 194. Continuous retraining lowers the cost of the random forest and the hybrid but leaves both far above the annual plan, and for the closest competing method it moves cost in the opposite direction to the accuracy gain. This is the same disconnect that RQ3 measures across methods, observed within a single method as its estimation frequency changes.

6. Discussion

6.1. Interpretation

The findings are consistent with previous studies showing that forecast rankings based on accuracy can diverge from rankings based on inventory outcomes [10] [11] [12]. The distinctive result in this case is that the company's incumbent annual plan, which contains no monthly demand information, matches the best machine-learning configuration and costs less than every other method once its safety stock is calibrated to its own errors. Part of that result reflects what the plan knows: it can carry the buyer's purchasing intentions, which the history-based methods cannot see, so the comparison measures the value of monthly models against an intention-informed annual plan rather than against a naive annual average. The relevant managerial question is therefore not simply whether a forecasting model can outperform the annual plan, but where monthly information creates enough operational value to justify the additional modelling effort.

Separating the two interventions sharpens that question. Moving from the company's fixed percentage mark-up to an error-calibrated buffer, without changing the forecast at all, raises aggregate fill to 0.99, from 0.84 at the same cost as the 30% mark-up or from 0.75 at a fifth less than the 20% mark-up. Changing the forecast, with the buffer rule held constant and a common service requirement imposed, produces no consistent cost improvement: only XGB-Tweedie is level with the annual plan, cheaper at the strictest requirement and more expensive at the loosest, and every other method is more expensive at every requirement examined. On this portfolio the buffer design is the larger lever, and the forecasting model the smaller one.

The strength of that statement should be read against its uncertainty. The point estimates place the annual plan and XGB-Tweedie ahead of every other method at every requirement, but resampling the portfolio widens the intervals enough that its ordering against seven of the nine alternatives is not resolved by sixteen units at the 0.98 requirement, two of which carry most of the volume. What the data do resolve is the gap to the hybrid and the naive method, which is large and stable under every treatment examined. The paper therefore claims that these two methods are clearly more expensive and that the rest are not separated, not that a winner has been identified.

Where XGB-Tweedie is level with the plan, on the single high-volume erratic unit and across the intermittent tail, the comparison depends on how leftover stock is valued: on the erratic unit the index is 100.5 in the primary window, 82 over the calendar year, and 110 and 118 when ending stock is charged 2.5 and 5 months of holding. The choice of objective function also had a substantial effect. Using the same features and hyperparameters with a squared-error objective produced a cost index of 190, while the two-stage hybrid produced 209, against 107 for the Tweedie objective. These results suggest that, for this dataset, matching the loss function to a zero-inflated and right-skewed target mattered more for inventory performance than the choice of algorithm family, although the same uncertainty caveats apply.

The RQ3 findings also have implications for forecasting practice and application-oriented research. The original company study [27] that motivated this analysis used adjusted MAPE and an analytic safety-stock formula to conclude that its hybrid ML system could reduce safety stock by approximately 276 tonnes at a 95% service level. The sequential simulation does not reproduce that estimated reduction. Although the model's accuracy advantage was measurable, it did not translate into lower cost once the forecasts were processed through the actual replenishment policy.

This result does not invalidate the accuracy analysis itself. It shows that translating forecast-error statistics directly into expected safety-stock savings can misrepresent the operational value of a forecasting method. In this case, replacing the one-time analytic calculation with a sequential simulation of the replenishment process changed the economic ranking of the methods.

6.2. Managerial insights

The results support three actions for PT XYZ and companies operating under comparable inventory conditions. First, modelling resources should be allocated according to both demand characteristics and operational importance, measured here by volume rather than by margin, since unit-specific economic weights were not available. The SBC classification can be used to direct any further modelling effort to the large units, where a monthly model is at least level with the plan and where cost differences are largest in absolute terms, while the annual plan with error-calibrated buffers remains appropriate for the intermittent tail.

Second, the company should replace fixed-percentage safety stocks with buffers calibrated to realised forecast errors over the policy's protection interval. In the simulation, this calibration allowed the methods to achieve broadly comparable aggregate service. The improvement therefore came from aligning the safety stock with the error behaviour of each forecast rather than applying the same percentage adjustment to every method and unit.

Third, policy simulation should be used as the final model-selection criterion. A forecasting method should enter operational use only when it reduces simulated cost at a common achieved-service requirement under the company's actual lead time, review cycle, and lot-size constraints, with the service target treated as a feasibility condition to be met before any cost is compared. Accuracy metrics can remain useful diagnostics, but the results show that neither percentage-based nor scaled-error metrics reliably identify the lowest-cost method for each unit. The released implementation provides a basis for applying this evaluation procedure to future models and updated demand data.

6.3. Limitations

Seven limitations define the scope of the findings. First, the analysis covers one company and one evaluation window: ordering decisions from January to July 2024, four review cycles, with their consequences observed to November 2024. Second, the erratic and smooth segments each contain only one unit; although these units dominate the portfolio by volume, the segment-level results are case observations rather than estimates. Third, the ML models are faithful reconstructions rather than the company's original models, and they are annual vintages while the local methods are re-fitted at every origin; the alternative schedule reported in Section 5.7 does not change the conclusions, but the reconstruction itself remains an approximation of the system in use.

Fourth, the service requirement is imposed as a minimum on the portfolio-level aggregate fill rate: the selected configurations land above the requirement by differing margins, individual units may sit above or below it, and no method is matched on cycle service at the same time. Table 3 reports the achieved fill and cycle service of every selected configuration so the residual is visible, and an interpolated exact-match check leaves the conclusions unchanged. Fifth, the cost parameters are normalised per kilogram and evaluated across a range of backorder-to-holding ratios rather than estimated from the company's financial records, which assumes homogeneous holding cost and shortage consequence across grades; aggregates therefore weight units by volume, an operational rather than an economic weighting.

Sixth, decisions stop at the July 2024 review and outcomes are counted to November so that every judged decision has an observable consequence, but the stock standing at the end of the outcome window is still valued by assumption rather than observed; the terminal-charge sensitivity bounds that effect; it moves the portfolio index of XGB-Tweedie by less than 5 points but is wide enough to turn the erratic-unit comparison from level to unfavourable. The stock entering 2024 is likewise set by each method's own warm-up decisions, and the common warm-up check bounds that effect. Extending the demand record into 2025 would remove this ambiguity and is the single most valuable addition to the dataset. Seventh, the ex-ante status of the annual plans rests on the company's description of its planning process together with the statistical evidence in Section 3; dated issue records for individual plans could not be recovered, so the benchmark is described as a company-described ex-ante plan throughout; and because the plan can carry the buyer's purchasing intentions, the comparison is partly one between information sets, so the results show that history-based models did not add value over this plan in this setting, not that they are inferior forecasting techniques. These limitations do not appear to favour a particular forecasting method systematically, but they bound the generalisability of the segment-level results and mean that the reported cost indices should not be interpreted as precise effect sizes outside the evaluated setting.

7. Conclusions and future research

This study examined whether higher forecast accuracy produces lower inventory cost for intermittent export demand, and addressed the question by simulating the company's replenishment policy on real data over a window in which every ordering decision has an observable consequence. Across the sixteen valid demand units, a company-described ex-ante annual plan combined with an error-calibrated safety stock was matched only by XGB-Tweedie, which cost between 10% more and 4% less across the achieved-service requirements examined, while the other eight alternatives cost at least 9% more, a margin that narrowed to 1% when every method shared a common warm-up. This result is specific to the information sets compared. The plan is issued by the buyer and can carry its purchasing intentions, which no history-based model observes, so the evidence shows that history-based monthly models did not add value over an intention-informed annual plan in this portfolio; it does not show that machine-learning forecasting lacks value where no such plan exists, where the models and the plan draw on the same information, or in other portfolios. Resampling the portfolio does not resolve the ordering among the annual plan and seven of the alternatives at the 0.98 requirement, but it does place the hybrid and the naive method consistently above them. On the single high-volume erratic unit XGB-Tweedie was level with the plan, and the comparison there turned on the valuation of leftover stock. Separating the buffer rule from the forecast showed that replacing the company's fixed percentage mark-up with an error-calibrated buffer, without changing the forecast, raised aggregate fill to 0.99, from 0.84 at the same cost as the 30% mark-up or from 0.75 at a fifth less

than the 20% mark-up; on this portfolio the buffer design was the larger operational lever. No accuracy metric ranked the methods consistently with simulated cost, and the adjusted MAPE used in practice was effectively unrelated to it. The evaluation protocol, which combines origin-safe rolling-origin forecasting, per-method error-calibrated buffers, sequential (R, S) simulation over an observable horizon, comparison under a common achieved-service requirement, and bootstrap uncertainty, is released in full and can be applied to other settings with periodic review and long lead times.

Five extensions follow from these findings. Replication across companies and years would establish how general these findings are, and would supply the segment sample sizes this single portfolio cannot. Giving the machine-learning models the annual plan as an input would compare the methods on a common information set and so separate the contribution of the model from that of the information it receives. Extending the demand record beyond the protection interval of the final reviews would remove the terminal-horizon ambiguity that currently bounds the closest comparison. Probabilistic forecasting, evaluated through the same policy simulation rather than through scoring rules alone, could strengthen the link between predictive distributions and buffer setting [15]. Finally, the simulation testbed developed here provides a basis for studying autonomous replenishment agents [23], which must decide not only order quantities but also when to defer to human judgement. This question is the subject of ongoing work.

Acknowledgments

The authors thank the case company for access to its operational data under a confidentiality arrangement.

Funding

None.

Data Availability

The demand data are commercially confidential and are not redistributed. The aggregate result tables computed from them, which are the evidence behind every table and figure in this paper, are released under a separate file namespace with recorded checksums, together with the complete simulation and analysis code. A synthetic data generator with the same panel structure is provided so that the pipeline can be executed end to end; it demonstrates the pipeline and does not reproduce the confidential case numbers, and it cannot overwrite the released tables. Regenerating the reported numbers requires the confidential transaction file and therefore the data owner. Code and released tables are available at <https://github.com/Ardi2020/value-of-forecast-intermittent-inventory> and archived at <https://doi.org/10.5281/zenodo.21783340>. Raw data may be requested from the corresponding author; access depends on the case company's consent.

Conflict of Interest

The authors declare that there is no conflict of interest to disclose.

Author Contributions

Asmuliardi Muluk: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision. **Taufik Nazir:** Resources, Project administration, Writing – review & editing. **Stephany Christie Lase:** Data curation, Software, Validation, Writing – review & editing. **Dika Aulia Ardi:** Resources, Data curation, Validation.

Declaration of using AI tools

During the preparation of this work the authors used Claude (Anthropic; Claude Opus 5, accessed July–October 2026) for linguistic correction of the manuscript text. The research design, the analysis code, the verification of all cited sources, the numerical results, and the final wording were produced and verified by the authors, who take full responsibility for the content of this article.

References

- [1] S. A. Rajagukguk, E. Heriyanni, A. F. Maulida, and E. Putri, "Cross-modal attention transformer for multi-SKU demand forecasting in retail environments," in *Information Systems for Intelligent Systems (Lecture Notes in Networks and Systems)*. Springer, 2026, pp. 281–290, doi: 10.1007/978-3-032-12993-2_27.
- [2] A. Seyam, S. S. Mathew, M. E. Barachi, C. Zhang, and J. Shen, "The application of machine learning and deep learning on demand forecasting across time-critical industries: A systematic review," *Adv. Eng. Inform.*, vol. 74, Sep. 2026, Art. no. 104625, doi: 10.1016/j.aei.2026.104625.
- [3] I. Jackson, J. Namdar, M. J. Saénz, R. A. Elmquist III, and L. R. Dávila Novoa, "Revolutionize cold chain: An AI/ML driven approach to overcome capacity shortages," *Int. J. Prod. Res.*, vol. 63, pp. 2190–2212, 2025, doi: 10.1080/00207543.2024.2398583.

- [4] P. Eichenseer, L. Hans, and H. Winkler, "A data-driven machine learning model for forecasting delivery positions in logistics for workforce planning," *Supply Chain Analytics*, vol. 9, Mar. 2025, Art. no. 100099, doi: 10.1016/j.sca.2024.100099.
- [5] T. U. Aye, A. A. Neto, and E. R. da Silva, "Knowing your surroundings: Tactical retail demand forecasting through exogeneity-aware machine learning," *IEEE Eng. Manag. Rev.*, early access, 2025, doi: 10.1109/emr.2025.3645405.
- [6] R. I. A. Omer, R. Hassan, and M. S. A. Aziz, "A review and comparative analysis of predictive models for supply chain demand forecasting," in *Proc. 10th Int. Conf. Inf. Commun. Technol. Muslim World (ICT4M)*, Nov. 2025, pp. 1–7, doi: 10.1109/ict4m68001.2025.11363509.
- [7] K. Aymane, N. Messaoudi, and H. Fouraiji, "Improving supply chain performance by improving the quality of demand forecasting: Theoretical background and bibliometric analysis," in *Proc. 16th Int. Conf. Logistics Supply Chain Manag. (LOGISTIQUEA)*, May 2025, pp. 1–8, doi: 10.1109/logistiqua66323.2025.11122780.
- [8] M. A. Habib, M. J. Hossain, S. Sakib, and M. M. Alam, "Bridging prediction and decision: An application-driven review of load forecasting in energy systems," *Comput. Electr. Eng.*, vol. 134, Jun. 2026, Art. no. 111113, doi: 10.1016/j.compeleceng.2026.111113.
- [9] M. Z. Babai, A. Syntetos, and R. Teunter, "Intermittent demand forecasting: An empirical study on accuracy and the risk of obsolescence," *Int. J. Prod. Econ.*, vol. 157, pp. 212–219, Nov. 2014, doi: 10.1016/j.ijpe.2014.08.019.
- [10] A. A. Syntetos, K. Nikolopoulos, and J. E. Boylan, "Judging the judges through accuracy-implication metrics: The case of inventory forecasting," *Int. J. Forecast.*, vol. 26, no. 1, pp. 134–143, Jan. 2010, doi: 10.1016/j.ijforecast.2009.05.016.
- [11] F. Petropoulos, X. Wang, and S. M. Disney, "The inventory performance of forecasting methods: Evidence from the M3 competition data," *Int. J. Forecast.*, vol. 35, no. 1, pp. 251–265, Jan. 2019, doi: 10.1016/j.ijforecast.2018.01.004.
- [12] N. Kourentzes, J. R. Trapero, and D. K. Barrow, "Optimising forecasting models for inventory planning," *Int. J. Prod. Econ.*, vol. 225, Jul. 2020, Art. no. 107597, doi: 10.1016/j.ijpe.2019.107597.
- [13] N. Sanders and G. Graman, "Quantifying costs of forecast errors: A case study of the warehouse environment," *Omega*, vol. 37, no. 1, pp. 116–125, Feb. 2009, doi: 10.1016/j.omega.2006.10.004.
- [14] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, Oct. 2006, doi: 10.1016/j.ijforecast.2006.03.001.
- [15] S. Kolassa, "Evaluating predictive count data distributions in retail sales forecasting," *Int. J. Forecast.*, vol. 32, no. 3, pp. 788–803, Jul. 2016, doi: 10.1016/j.ijforecast.2015.12.004.
- [16] J. D. Croston, "Forecasting and stock control for intermittent demands," *J. Oper. Res. Soc.*, vol. 23, no. 3, pp. 289–303, Sep. 1972, doi: 10.1057/jors.1972.50.
- [17] A. A. Syntetos and J. E. Boylan, "The accuracy of intermittent demand estimates," *Int. J. Forecast.*, vol. 21, no. 2, pp. 303–314, Apr. 2005, doi: 10.1016/j.ijforecast.2004.10.001.
- [18] R. H. Teunter, A. A. Syntetos, and M. Z. Babai, "Intermittent demand: Linking forecasting to inventory obsolescence," *Eur. J. Oper. Res.*, vol. 214, no. 3, pp. 606–615, Nov. 2011, doi: 10.1016/j.ejor.2011.05.018.
- [19] J. Boylan and A. Syntetos, *Intermittent Demand Forecasting: Context, Methods and Applications*. Hoboken, NJ, USA: Wiley, 2021, doi: 10.1002/9781119135289.
- [20] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *Int. J. Forecast.*, vol. 38, no. 4, pp. 1346–1364, Oct. 2022, doi: 10.1016/j.ijforecast.2021.11.013.
- [21] Y. Hu et al., "Implementing a prediction driven framework for emergency department nurse staffing to optimize real time decisions," *npj Health Syst.*, vol. 2, May 2025, Art. no. 16, doi: 10.1038/s44401-025-00019-2.
- [22] M. F. Hadi and A. Muluk, "A machine-learning decision support system for demand-driven nurse staffing in outpatient clinics: Design and validation at an Indonesian public hospital," *Trans. Comput. Model. Intell. Syst.*, vol. 4, Jul. 2026, Art. no. 10092, doi: 10.65112/tcmis.10092.
- [23] S. Maitra, "Agentic automation driven reinforcement learning for inventory optimization," *J. Model. Manag.*, early access, pp. 1–36, Apr. 2026, doi: 10.1108/jm2-07-2025-0347.
- [24] K. Ramaekers and G. K. Janssens, "Optimal policies for demand forecasting and inventory management of goods with intermittent demand," *J. Appl. Oper. Res.*, vol. 6, no. 2, pp. 111–123, 2014.
- [25] A. A. Syntetos, J. E. Boylan, and J. D. Croston, "On the categorization of demand patterns," *J. Oper. Res. Soc.*, vol. 56, no. 5, pp. 495–503, May 2005, doi: 10.1057/palgrave.jors.2601841.
- [26] N. Kourentzes, "On intermittent demand model optimisation and selection," *Int. J. Prod. Econ.*, vol. 156, pp. 180–190, Oct. 2014, doi: 10.1016/j.ijpe.2014.06.007.
- [27] S. C. Lase, "Pengembangan sistem pendukung pengadaan bahan baku berbasis peramalan permintaan menggunakan machine learning (studi kasus: PT XYZ) [Development of a raw-material procurement support system based on machine-learning demand forecasting (case study: PT XYZ)]," Undergraduate thesis, Dept. Ind. Eng., Univ. Andalas, Padang, Indonesia, 2026.

Appendix: Simulation algorithm

Algorithm A1 states the sequential simulation for one demand unit u and one forecasting method; the symbols are those of Eqs. (5)–(8).

1. Initialisation. At the first review, t_0 , January 2023, set the on-hand stock to the method's order-up-to level S_{t_0} , the backorders to zero, and the pipeline to empty.
2. Monthly loop. For each month t from January 2023 to December 2024:
 - (a) receive the orders due in month t ;
 - (b) if t is a review month, every $R = 2$ months, and $t \leq t_{\text{last}}$, where t_{last} , July 2024, is the last review whose protection interval is observed: compute the five-month forecast F_t , the sum of $f_{u,t+j}^{(t)}$ over $j = 0, \dots, 4$, the buffer $B_t = z \sigma_t$ by Eq. (6) or $B_t = \gamma F_t$ under the fixed mark-up, the order-up-to level $S_t = \max(0, F_t + B_t)$, the inventory position IP_t , and the order q_t by Eq. (7), which arrives in month $t + L$ with $L = 3$;
 - (c) serve the backorders and then the demand $y_{u,t}$ from stock, and add any unmet demand to the backorders;
 - (d) record the ending stock $I_{u,t}$, the backorders $K_{u,t}$, and the pipeline $P_{u,t}$.
3. Cost. Compute C by Eq. (8) over the outcome window $\mathcal{O}$, January to November 2024, or January to December 2024 as a sensitivity, with terminal charge $k \in \{0, 2.5, 5\}$. The reviews of 2023 form a warm-up whose costs are not counted; in the common warm-up check every 2023 review uses the annual plan with $z = 1.645$.