INTERNATIONAL JOURNAL OF
INTELLIGENT OPTIMIZATION
AND LOGISTICS

Available online at www.ijiol.com

RESEARCH ARTICLE

Criticality-Aware Hybrid Ensemble Forecasting for Intermittent Aircraft Spare-Parts Demand: From Probabilistic Lead-Time Demand to Inventory Policy

Badr-Eddine Badi[a,*], **Oussama Bouazaoui**[a], **Ismail Lagrat**[a], **Driss Serrou**[b], **Mahmoud Ouhdan**[a]

[a] Advanced Systems Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco

[b] Laboratory for Scientific Research and Innovation, Higher School of Technology, Ibn Tofail University, Kenitra, Morocco

ABSTRACT

Demand for aircraft spare parts is intermittent, heterogeneous and subject to safety-critical shortages. We investigate an end-to-end pipeline using XGBoost, random forest and an LSTM, followed by class-wise dynamic forecast combination, negative-binomial (NB2) lead-time-demand calibration, AHP-based criticality tiers and a weekly reviewed (s, Q) simulation. The dataset is a mechanism-based synthetic fleet panel (150 parts over 313 weeks), not airline transaction data. Calibration uses only H-week validation targets ending within the 2023 validation period, so no test-year outcome enters the fitted dispersion. The three-model ensemble and six inventory policies are evaluated on three independently generated panels (seeds 2026–2028) at baseline schedule–demand coupling, and on three panels with weakened coupling. Models are implemented in TensorFlow/Keras (a three-input LSTM with a Tweedie objective) with an extended hyper-parameter search; numerical results depend on the software and hardware environment. For the seed-2026 strong-coupling run, ensemble RMSE is 3.698 versus 4.010 for TSB, while the proposed tiered NB2 inventory policy costs \$5.90 M against \$12.76 M for SBA/normal/uniform stocking (saving 53.7%; part-bootstrap 95% CI [27%, 68%]). Across all six panels the saving against P1 ranges from 12.2 to 63.0% (14.5 to 55.0% at baseline coupling), yet the proposed policy is the cheapest deployable option in only four panels and, against the best uniform baseline selected in test-year hindsight, ranges from −40.5 to +32.1% with no significant advantage. The ensemble's incremental accuracy over equal or statically combined forecasts is inconsistent. These controlled synthetic experiments support evaluating forecasting and stocking jointly but do not establish transferable airline savings.

KEYWORDS: Spare-parts Forecasting · XGBoost · LSTM · Criticality Classification · Inventory Simulation · Aviation Maintenance

* Corresponding author: badreddine.badi@uit.ac.ma.

Article history: Received 8 October 2026; Revised 19 October 2026; Accepted 18 November 2026; Published 23 November 2026.
Publication details: Volume 1, Issue 1, Article 0004.

© **2026 The Authors.** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

1. Introduction

Aircraft maintenance organisations must provision rotables, repairables and consumables against highly intermittent and heterogeneous component demand. An unavailable safety-critical part can disrupt aircraft operations, while overstocking costly slow movers ties up capital. Forecast error alone is therefore an incomplete criterion: the lead-time demand distribution, reorder rule and relative consequence of a shortage all affect the operational decision [1–3].

Croston-family methods remain essential comparators for intermittent demand [4–6], but generally do not use planned checks, aircraft utilisation or part attributes. Modern covariate-driven predictors permit that information to be incorporated; nevertheless, a lower point RMSE does not automatically imply lower shortage costs or superior service. Similarly, complex estimated combination weights need not outperform equal weights, an observation associated with the forecast combination puzzle [7].

We investigate an end-to-end, *synthetically evaluated* forecasting-to-inventory architecture: multi-horizon XGBoost, random forest and LSTM forecasts feed a dynamic ensemble, negative-binomial (NB2) lead-time-demand quantiles and criticality-dependent (s, Q) inventory decisions. Our central question is whether gains are attributable to the predictive learner, the distributional safety-stock rule or the assignment of service levels, and how far those gains depend on design assumptions. The study offers four contributions:

1. A documented, chronologically separated point-forecast comparison against rolling-origin intermittent-demand benchmarks, equal-weight and static-weight combinations.
2. Explicit equations for multi-horizon interpolation, NB2 validation-likelihood calibration and the translation of nominal cycle-service targets into reorder points.
3. A six-policy inventory comparison that contrasts forecast inputs, demand-distribution approximations and criticality tiers, with a documented shortage-penalty and holding-cost sensitivity grid.
4. End-to-end repeat fitting of all three forecasting members, NB2 dispersion and six inventory policies on independent synthetic seeds under both strong and weak maintenance coupling, with results reported separately for every panel.

No airline transaction dataset was available for external validation. The synthetic generator deliberately links scheduled checks to part demand; demonstrating performance when that link is weakened is consequently important. The numerical savings reported below describe this simulator under specified costs, not expected savings for a real operator. Probabilistic calibration respects the validation boundary in every reported run. The lack of airline data and the limited number of synthetic seeds remain important limitations.

The remainder is organised as follows. Section 2 reviews intermittent and probabilistic forecasting, forecast combination and stocking policies. Section 3 describes the task and synthetic benchmark; Section 4 specifies the forecasting-to-inventory methods. Section 5 reports prediction, supplementary robustness and inventory results. Section 6 discusses interpretation and validity, and Section 7 concludes.

2. Related work

Intermittent-demand models and aviation applications. The ADI–CV^2 classification partitions series into smooth, erratic, intermittent and lumpy demand regimes [8]. Croston separates positive demand size and occurrence interval [4]; the Syntetos–Boylan approximation (SBA) corrects its bias [5], and TSB updates the probability of demand directly [6]. ADIDA aggregates sparse observations into longer buckets [9]. Research based on airline records documents utilisation-driven intermittency and tests classical forecast families [1, 2, 10]. Maintenance-linked demand modelling is particularly pertinent when a future maintenance plan is known [11]; such relevance also creates the risk that a simulator with strongly encoded schedule–consumption dependence rewards models designed to exploit that dependence.

Probabilistic and neural forecasting. Sparse nonnegative demand requires explicit treatment of occurrence frequency, demand size and uncertainty. Count-based or state-space descriptions are discussed by Snyder et

al. [12], Seeger et al. [13] and Kolassa [14]. DeepAR learns probabilistic forecasts by sharing information across related time series [15]; the Temporal Fusion Transformer (TFT) exploits known-future inputs and attention-based multi-horizon representations [16]; and N-BEATS uses residual basis expansion for point forecasting [17]. Conformal time-series forecasting investigates distribution-free calibration of predictive intervals under stated assumptions [18]. These are relevant methodological alternatives, *not* additional models implemented or experimentally benchmarked in this study. In particular, nominal conformal coverage cannot be presumed under unrestricted temporal dependence or arbitrary dataset shift.

Why tree ensembles here? Previous work investigates neural predictors for irregular demand [19–21] and analyses the behaviour of recurrent forecasters [22]. Tree-based methods remain competitive on moderate-sized tabular datasets with heterogeneous covariates [23], while Tweedie-loss boosting provides a nonnegative response model suited to zero-heavy, right-skewed outcomes [24]. This motivates XGBoost and random forest as practical comparators to a modest LSTM rather than claiming that these methods universally dominate deep probabilistic models. A full DeepAR/TFT/N-BEATS comparison would require additional implementation and tuning and is outside the evidence supplied here.

Combining forecasts. Classical forecast combination is intended to reduce dependence on a single forecasting specification [25, 26]. The forecast combination puzzle describes why estimated optimal weights may perform worse than equal weights: correlated member errors limit diversification, while fitted weights add estimation variability and can become unstable [7]. Accordingly, the present work compares its moving weights to equal and validation-estimated static weights and evaluates their incremental contribution instead of assuming that adaptivity implies superiority.

Recent aviation evidence and related conference work. Del Olmo and Domingo [27] test aerospace rotable-part forecasting on actual MRO consumption, with operational features and LightGBM. Chen et al. [28] analyse repair capacity and queuing with airline engineering records; Brito et al. [29] review aviation spare-parts methods, while their later systematic review [30] considers maintenance-demand integration and potential resource-efficiency implications. These reviews synthesise published studies rather than directly validating the cost savings reported here. Our team has also published a closely related conference contribution [31], first available online in July 2026. This article adds a validation-boundary-controlled probabilistic calibration, independent synthetic generator seeds and detailed multi-policy sensitivity results. We do not claim independent industrial validation here.

From predictive distributions to inventory. Inventory safety stocks can be miscalibrated if forecast uncertainty or correlation across forecast errors is ignored [3, 32]. Stock-control and classification studies emphasise differentiated policies for spare parts [33, 34]; AHP offers a transparent, consistency-checked way to express judgemental criteria [35], while (s, Q) policies are established in inventory theory [36, 37]. In aviation, service priorities must distinguish the *cycle service-level target* encoded in a reorder-point quantile from the realised unit fill rate. The financial preference for availability is conditional on shortage penalties and lead-time variability rather than a universal economic constant.

3. Problem formulation and data

3.1. Notation and forecasting task

Let $y_{i,t}$ be the units of part i demanded in week t ($i = 1, \ldots, P$). At forecast origin $t - 1$, all information up to the end of week $t - 1$ is available, together with the *planned* maintenance calendar and planned flying programme for week t and beyond. For $H \in \{1, 4, 13\}$ we forecast the mean weekly rate

$$r_{i,t}^{(H)} = \frac{1}{H} \sum_{j=0}^{H-1} y_{i,t+j}. \tag{1}$$

Horizon $H = 1$ is used for accuracy comparison, and $H = 4, 13$ provide the lead-time-demand inputs to the inventory layer.

3.2. Synthetic data and why they are used

Operator consumption records were not available for this analysis. Consequently, the entire experimental dataset is synthetic. The generator (`generate_synthetic_data.py`) encodes selected fleet, maintenance and component-consumption mechanisms and is reproducible, but its parameters have not been calibrated to an independent airline population; realism and generalisability must not be inferred from the model structure alone:

- **Fleet.** 42 aircraft (A320, A321, B737-800, B787-9) with delivery dates, entry into service, four retirements, ageing, weekly flight hours and cycles, A- and C-check schedules (C-check: three weeks out of service), a seasonal utilisation cycle, and a 2020–21 utilisation collapse followed by recovery. The active fleet ranges from 28 to 41 aircraft.
- **Catalogue.** 150 parts mapped to ATA-100 chapters, built from 94 component archetypes in three classes (22 rotables, 86 repairables, 42 consumables), with unit cost (median \$6 010, range \$99–\$99 595), base lead time (median 25 days, range 4–104), safety-impact rating (1–5), no-go flag, installed count per aircraft, usage driver (flight hours or cycles) and Weibull wear-out shape.
- **Demand.** Expected weekly demand is an unscheduled component (usage × age-dependent wear-out hazard × climate effect × an AR(1) reliability frailty × campaign multipliers) plus a scheduled component proportional to planned A/C-checks. Arrivals are Poisson with Gamma over-dispersion and mild self-excitation, and each arrival carries a compound (lumpy) order size. Eight parts receive service-bulletin campaigns that multiply demand by 2.5–4 for 6–10 weeks, five in the training period and three inside the test year.
- **Exogenous shocks.** A supplier-disruption index inflates quoted lead times by up to 45% in 2022–23; unit cost drifts at 3% per year.

The panel covers 313 weeks from 2019-01-07 to 2024-12-30 (46 950 rows, 71 176 units in total). Mean demand is 1.52 units per part-week and 67.6% of part-weeks are zero (61.1% in the test year). By the classification rule above (computed on training data) the portfolio contains 0 smooth, 9 erratic, 111 intermittent and 30 lumpy items (Figure 1b). The latent expected demand is stored separately (`ground_truth_expected_demand.csv`) and is never a model input; it is used only to compute the oracle reference. The realised test-year R^2 of the latent conditional expectation is 0.56; stochastic counts limit the achievable explained variance even when the generator expectation is known. This is a simulator-specific reference rather than a transferable accuracy ceiling.

3.3. Chronological split

Weeks 1–208 (2019-01-07 to 2022-12-26) are used for training, the next 52 weeks (2023-01-02 to 2023-12-25) for validation (early stopping, hyper-parameter selection, ensemble hyper-parameters, dispersion estimation) and the final 53 weeks (2024-01-01 to 2024-12-30) for testing, i.e. approximately 67%/17%/17%. After selection, XGBoost, random forest and LSTM are refitted on training plus validation for the test year with the number of boosting rounds or epochs fixed from validation. Targets whose horizon crosses a split boundary are excluded from the corresponding fitting set.

4. Method

4.1. Features

Features (`FEATURE_GROUPS` in the code; 42 inputs in five groups) are computed at the forecast origin. *History:* lags 1, 2, 3, 4, 8, 13; rolling means (4, 13, 26, 52 weeks) and standard deviations (4, 13, 26) of lagged demand. *Intermittency:* share of zero weeks over 13 and 52 weeks, ADI and mean non-zero size over 52 weeks, weeks since last demand. *Operations:* planned flight hours per eligible aircraft, lagged flight hours and cycles per aircraft, ratio of planned to recent usage, eligible fleet size, mean age, lagged hub temperature, week-of-year sine/cosine. *Maintenance plan:* A- and C-checks planned in the current week, in the next 4 and in the next 13 weeks (the check schedule is a known plan, not an outcome). *Part attributes:* class, safety rating, no-go flag, installed count, ATA chapter, usage driver, log unit cost and last quoted lead time. The week-of-year features use

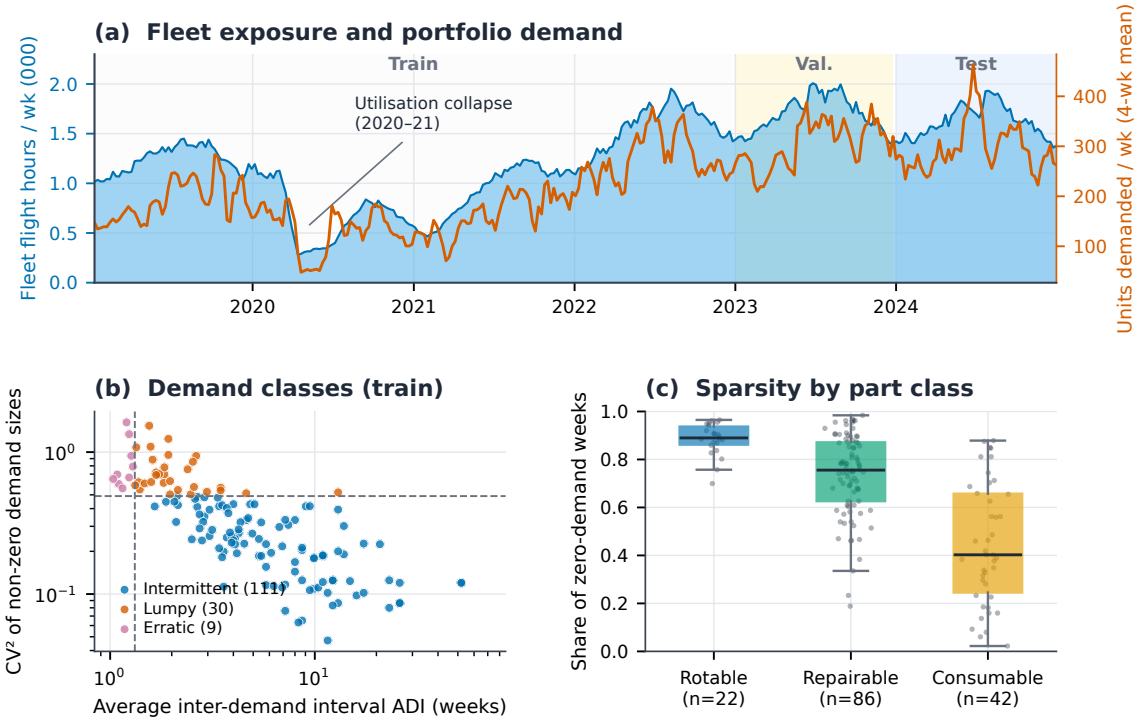

Figure 1. Synthetic benchmark. (a) Fleet flight hours and portfolio demand (4-week moving average) with the chronological split. (b) ADI–CV^2 plane with the Syntetos–Boylan–Croston cut-offs. (c) Share of zero-demand weeks by part class.

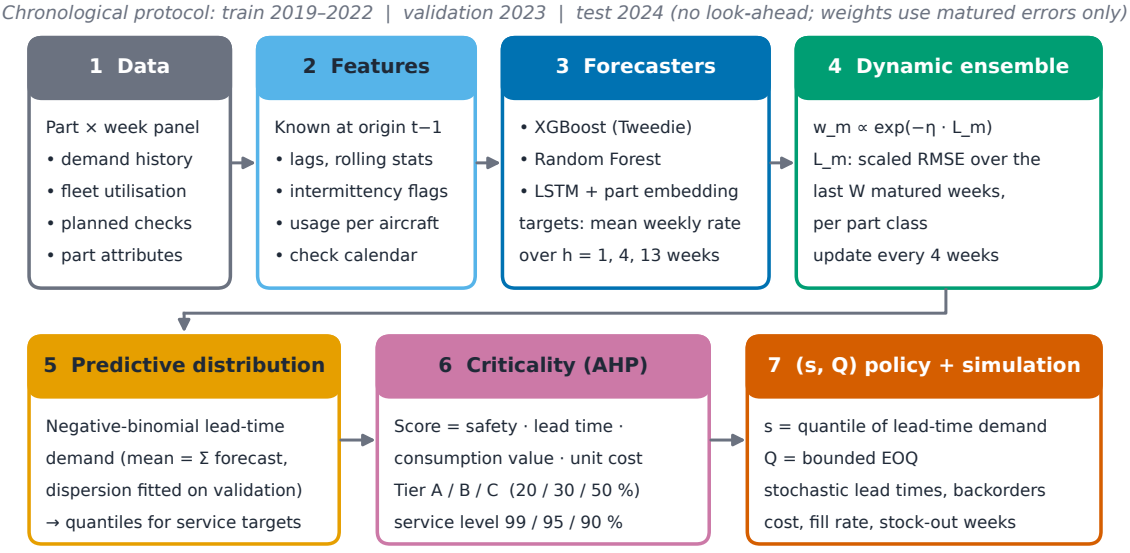

Figure 2. Framework. All inputs are known at the forecast origin; the ensemble uses only errors that have already matured; the predictive distribution and the criticality tier jointly determine the reorder point.

cyclical sine and cosine encodings. The forward check counts represent an assumed available planning calendar; their information advantage may disappear when actual maintenance plans are revised or less predictive of withdrawals.

4.2. Statistical benchmarks

Naive (last value), 13-week moving average, SES, Croston [4], SBA [5], TSB [6] and ADIDA with 4-week buckets [9] are evaluated as *rolling-origin* forecasts: the forecast for week t uses only data up to $t-1$ and the state is updated every week. Smoothing parameters are chosen on the validation year from $\{0.02, 0.05, 0.1, 0.2, 0.3\}$ (TSB: $\alpha \in$ same set, $\beta \in \{0.02, 0.05, 0.1, 0.2\}$); the selected values for every independent panel are in the accompanying result JSON files. Statistical methods produce one rate per week, which is used for every horizon.

4.3. Learners

XGBoost [24] with a Tweedie objective (variance power in $\{1.1, 1.3, 1.5\}$) so that predictions are non-negative means for zero-inflated data; hyper-parameters by random search [38] with 14 draws on validation RMSE at $H = 1$ in each run, then early stopping per horizon. **Random forest** [39] using three tuned 120-tree random forest configurations. **LSTM** [40]: a 26-week window of eight scaled per-step inputs (log-scaled demand, zero flag, flight hours and cycles per aircraft, A/C-checks, temperature, age) feeds one LSTM layer; its output is concatenated with 21 static or known-ahead covariates and an 8-dimensional part embedding, passes through a 64-unit dense layer and ends in three softplus outputs, one per horizon. It is trained with the Tweedie deviance (variance power 1.3), on demand scaled by each part's training mean, with Adam [41], early stopping on validation loss, with three LSTM random initialisations averaged for each independent panel. The TensorFlow/Keras LSTM uses the grid (32, 0.003), (64, 0.001), (64, 0.003), a maximum of 30 epochs and validation-based early stopping. All hyper-parameter choices were made on validation data only.

4.4. Dynamic ensemble

Let $f_m(\cdot)$ be the forecast of learner $m \in \{1, \dots, M\}$, $c(i)$ the part class of i (rotable, repairable, consumable) and $\mathcal{M}_H(\tau) = \{s \le \tau - H\}$ the origins whose H-week target is fully observed before week τ. At each update week τ (every four weeks) the scaled loss of learner m for class c over the last W matured origins is

$$\begin{aligned} L_{m,c}(\tau) &= \frac{\mathrm{RMSE}_{m,c}(\tau)}{\frac{1}{M}\sum_{k=1}^{M} \mathrm{RMSE}_{k,c}(\tau)}, \\ \mathrm{RMSE}_{m,c}(\tau) &= \left(\mathrm{mean}_{\,i:c(i)=c,\; s\in\mathcal{M}_H(\tau),\; s>\tau-H-W} \left(f_{m,i,s} - r_{i,s}^{(H)} \right)^2 \right)^{1/2}, \end{aligned} \tag{2}$$

and the weights, which stay fixed until the next update, are

$$w_{m,c}(\tau) = \frac{\exp\{-\eta\, L_{m,c}(\tau)\}}{\sum_{k=1}^{M} \exp\{-\eta\, L_{k,c}(\tau)\}}, \qquad \hat{r}_{i,t}^{(H)} = \sum_{m=1}^{M} w_{m,c(i)}(\tau)\, f_{m,i,t} \quad (\tau \le t < \tau + 4). \tag{3}$$

Normalising by the mean RMSE makes η scale-free, and $\eta = 0$ gives equal weights. The pair $(\eta, W) \in \{0, 2, 5, 10, 20, 40\} \times \{13, 26, 52\}$ is selected on validation RMSE at $H = 1$. Because $\mathcal{M}_H(\tau)$ excludes origins whose target is not yet observed, the weights never use future outcomes. Reference combinations are equal weights and static class-wise non-negative least squares fitted on validation.

4.5. Probabilistic lead-time demand

Let $r_{i,t}^{(h)}$ denote a forecast of the *mean weekly demand rate* over horizon $h \in \{1, 4, 13\}$. The implementation uses piecewise-linear interpolation in the rate, with constant endpoint extrapolation. For arbitrary v_1, v_4, v_{13} define

$$\mathcal{I}(\ell; v_1, v_4, v_{13}) = \begin{cases} v_1, & \ell \le 1, \\ v_1 + (\ell - 1)(v_4 - v_1)/3, & 1 < \ell \le 4, \\ v_4 + (\ell - 4)(v_{13} - v_4)/9, & 4 < \ell \le 13, \\ v_{13}, & \ell > 13. \end{cases} \tag{4}$$

Here $d_{i,t}^{\text{quote}}$ is the quoted delivery time in days; the protection period $\ell_{i,t} = d_{i,t}^{\text{quote}}/7 + 1$ includes a one-week review allowance. The weekly rate and mean demand for the protection period are

$$\bar{r}_{i,t} = \max\{0, \mathcal{I}(\ell_{i,t}; \hat{r}_{i,t}^{(1)}, \hat{r}_{i,t}^{(4)}, \hat{r}_{i,t}^{(13)})\}, \qquad \mu_{i,t} = \ell_{i,t}\bar{r}_{i,t}. \tag{5}$$

Interpolation is applied to mean *rates*, not to the three horizon-total forecasts. For $\ell > 13$ the 13-week rate is held constant, an approximation that merits recalibration for very long lead times.

To model overdispersed demand, assume $Y_{i,t}(\ell) \sim \text{NB2}(\mu_{i,t}, \alpha_{i,t})$ with $\text{Var}(Y \mid \mu, \alpha) = \mu + \alpha\mu^2$. For class c and horizon $h \in \{1, 4, 13\}$, let the validation observations be $z_{i,t}^{(h)} = \sum_{j=0}^{h-1} y_{i,t+j}$ and their fitted total means $m_{i,t}^{(h)} = h\hat{r}_{i,t}^{(h)}$. Dispersion is estimated separately by

$$\hat{\alpha}_{c,h} = \arg \max_{\exp(-6) \le \alpha \le \exp(4)} \sum_{\substack{i:c(i)=c \\ t \in \mathcal{V}_h}} \log \text{NB}\left(z_{i,t}^{(h)};\ k = \frac{1}{\alpha},\ p = \frac{k}{k + m_{i,t}^{(h)}}\right), \tag{6}$$

where $\mathcal{V}_h = \{t :\ t_{\text{val,start}} \le t,\ t + h \le t_{\text{val,end}}\}$ includes strictly those validation origins whose entire h-week target lies before the test boundary. The numerical optimiser searches on $\log \alpha$, and fitted means are floored at 10^{-3} to stabilise the likelihood. Class-specific dispersion for a protection period is

$$\alpha_{i,t} = \mathcal{I}(\ell_{i,t}; \hat{\alpha}_{c(i),1}, \hat{\alpha}_{c(i),4}, \hat{\alpha}_{c(i),13}). \tag{7}$$

Thus the implementation linearly interpolates *dispersion values*, not their logarithms, and uses constant extrapolation outside $[1, 13]$. The NB2 reorder point at nominal cycle-service probability β is

$$s_{i,t} = \left\lceil F_{\text{NB2}}^{-1}(\beta; \mu_{i,t}, \alpha_{i,t}) \right\rceil. \tag{8}$$

The comparison normal rule instead uses $s = \lceil \max\{0, \mu + z_\beta \sigma_i \sqrt{\ell}\} \rceil$, with σ_i estimated from validation one-week forecast errors. Calibration of one-week predictive counts is assessed using CRPS and empirical quantile coverage [14, 42].

Lead-time modelling and validation boundary. Reorder points are constructed at the quoted deterministic lead time, whereas physical delivery times in the inventory simulator have log-normal coefficient of variation 0.25. The policy therefore does not integrate the supplier-time distribution into its NB2 quantile; a complete treatment would use $\Pr(Y \le s) = \int \Pr(Y(L) \le s \mid L = \ell)\, \mathrm{d}F_L(\ell)$. Calibration targets must lie entirely inside the validation period; the selection mask is `mask_val & (tt <= n_trva-H)`. Hence the accepted validation-origin counts are 52, 49 and 40 per part for $H = 1, 4, 13$ respectively (rather than 52 for every horizon), and no 2024 outcome enters the dispersion fit. An assertion-based test and the fitted class-wise dispersions are included in the package. The simulated inventory confidence intervals are still conditional on these synthetic demand and cost mechanisms.

4.6. Criticality tiers

Four criteria are scored on $[0, 1]$: safety (rating/5 weighted 0.8 plus the no-go flag weighted 0.2), lead time, annual consumption value and unit cost (the last three as percentile ranks on training data). Weights come from an AHP pairwise-comparison matrix on Saaty's scale [35],

$$\mathbf{A} = \begin{pmatrix} 1 & 3 & 4 & 5 \\ 1/3 & 1 & 2 & 3 \\ 1/4 & 1/2 & 1 & 2 \\ 1/5 & 1/3 & 1/2 & 1 \end{pmatrix}, \tag{9}$$

whose principal eigenvector gives weights safety 0.55, lead time 0.23, consumption value 0.14, unit cost 0.08 (consistency ratio 0.019, below the customary 0.10). The score is ranked and the top 20% of parts form tier A, the next 30% tier B and the rest tier C, with target cycle service levels 0.99, 0.95 and 0.90 respectively. The matrix and tier cut-offs are modelling judgements, not elicited operator preferences; the cost-sensitivity analysis in Section 5.6 changes the shortage and holding assumptions but does not identify a unique optimal AHP matrix. Robustness to alternative expert judgements remains a validation need.

4.7. Inventory policy and simulation

Each part follows a continuous-quantity (s, Q) rule reviewed weekly: when the inventory position (on hand plus on order minus backorders) is at or below s, an integer multiple of Q is ordered to bring it above s. The batch size is $Q = \text{round}\left(\sqrt{2DS/(hc)}\right)$ with annual demand $D = 52\hat{r}^{(13)}$, ordering cost $S = \$150$, holding rate $h = 0.20$ per year and unit cost c, bounded between 1 and $\lceil 26\hat{r}^{(13)} \rceil$. Unmet demand is backordered. Realised lead times are the quoted lead time multiplied by a log-normal factor with coefficient of variation 0.25. Costs are holding ($hc/52$ per unit-week on hand), shortage (penalty per backordered unit-week: \$20 000 for tier A, \$2 500 for B, \$300 for C) and ordering. The simulation starts 26 weeks before the test year to remove start-up effects, runs 20 replications with common random numbers across policies, and reports costs for the 53 test weeks. The shortage penalties and holding rate are assumptions; Section 5.6 varies them. Six policy configurations provide informative comparisons, but not a complete factorial estimate of independent causal contributions from the three design choices:

Policy	Forecast engine	Lead-time distribution	Service level
P1	SBA	Normal	uniform 0.95 (practice baseline)
P2	TSB	Normal	tiers 0.99/0.95/0.90
P3	Hybrid ensemble	Normal	tiers
P4	Hybrid ensemble	Negative binomial	uniform 0.95
P5	Hybrid ensemble	Negative binomial	tiers (proposed)
P6	Latent expectation (oracle)	Negative binomial	tiers (bound)

4.8. Evaluation

Point accuracy uses RMSE, MAE, mean error (bias), R^2 and the mean per-part RMSSE (root mean squared error scaled by the RMS of the in-sample one-step naive error). Because MAE is minimised by the median, which is zero for most intermittent series, RMSE and RMSSE are the primary criteria [43, 44]. Uncertainty for comparisons is obtained by resampling parts with replacement (2 000 bootstrap draws) for ratios of test MSE and of total cost, and by the paired Wilcoxon signed-rank test on part-level MSE [45, 46]. All hyper-parameter and weighting decisions use validation data only, and the test year is evaluated once.

5. Results

5.1. Forecast accuracy for the primary panel

All results come from the code included with the manuscript. For the primary seed-2026, strong-coupling run, test RMSE of the dynamic ensemble is 3.698, against 4.010 for TSB (a 7.8% relative reduction). The component learners yield RMSE 3.728 (XGBoost), 3.846 (random forest) and 3.718 (LSTM). The equal-weight combination gives 3.709, and static NNLS gives 3.699. Thus the extra estimation effort of dynamic weighting is not justified by consistent one-week accuracy gains. The LSTM is the TensorFlow/Keras model described in Section 4, fitted with the extended hyper-parameter search; all figures and statistics derive from these fits.

5.2. Forecast horizons and demand strata

At $H = 4$ and $H = 13$ the dynamic ensemble has RMSE 2.160 and 1.480, respectively. Because lead-time orders depend on integrated forecast rates and service quantiles, no single one-week point accuracy score establishes inventory quality. Demand-class and tier stratification for the primary run is shown in Figure 4. Few observations in small strata limit inference.

5.3. Forecasts, weights and feature contribution

The test-year forecasts and NB2 bands for the primary run are plotted in Figure 5. Validation selects $\eta = 20$ and $W = 13$ weeks for the dynamic ensemble in this run. Figure 6 reports the corresponding weights and parameter grid. In the independent-seed tests, selected weights change materially, including a zero-η (equal-weights) optimum. These observations support the forecast-combination argument: correlated member errors and noisy short-window estimates can erase any incremental benefit from dynamic adaptation [7, 26]. The XGBoost

Table 1. One-week test accuracy for seed 2026 (TensorFlow/Keras, extended search). Test MSE ratios and part-bootstrap 95% intervals are relative to the dynamic ensemble.

Model	RMSE	MAE	Bias	R^2	RMSSE	MSE ratio [95% CI]	p
Naive	5.185	2.024	+0.005	0.013	1.359	1.97 [1.79, 2.13]	<0.001
Moving average (13 wk)	4.121	1.759	+0.043	0.377	1.048	1.24 [1.13, 1.35]	<0.001
SES	4.017	1.724	+0.045	0.407	1.025	1.18 [1.09, 1.27]	<0.001
Croston	4.021	1.707	-0.003	0.406	1.029	1.18 [1.09, 1.28]	<0.001
SBA	4.016	1.692	-0.052	0.408	1.029	1.18 [1.09, 1.28]	<0.001
TSB	4.010	1.714	+0.032	0.409	1.028	1.18 [1.09, 1.27]	<0.001
ADIDA	4.022	1.725	+0.047	0.406	1.026	1.18 [1.09, 1.29]	<0.001
XGBoost	3.728	1.661	+0.145	0.490	1.012	1.02 [1.00, 1.04]	<0.001
LSTM	3.718	1.633	+0.045	0.492	1.002	1.01 [1.00, 1.03]	0.472
Random forest	3.846	1.740	+0.297	0.457	1.027	1.08 [1.05, 1.12]	<0.001
Ensemble, equal	3.709	1.659	+0.162	0.495	1.004	1.01 [1.00, 1.01]	<0.001
Ensemble, NNLS static	3.699	1.645	+0.123	0.497	1.001	1.00 [0.99, 1.01]	0.064
Ensemble + TSB	3.711	1.651	+0.130	0.494	1.003	1.01 [1.00, 1.01]	0.482
Ensemble, dynamic	3.698	1.649	+0.138	0.498	1.002	1 (ref.)	–
Oracle (latent mean)	3.459	1.524	+0.072	0.561	0.962	–	–

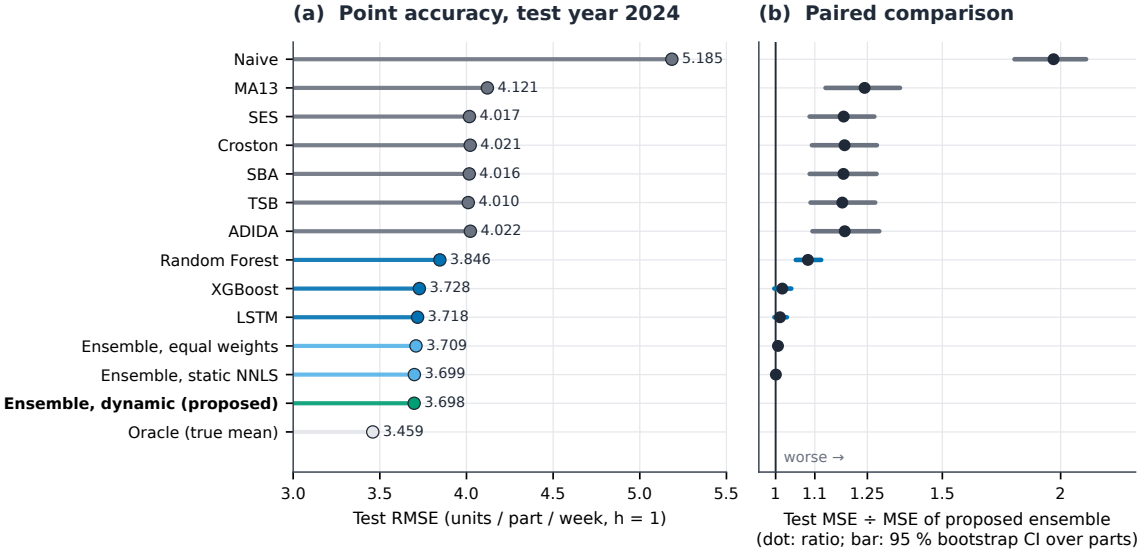

Figure 3. Seed-2026 one-week test RMSE and per-part bootstrap MSE ratios.

Table 2. Test RMSE for mean weekly demand rates at horizons $H = 1, 4, 13$; seed 2026.

Model	$H = 1$	$H = 4$	$H = 13$
TSB	4.010	2.397	1.713
SBA	4.016	2.384	1.691
Croston	4.021	2.394	1.700
ADIDA	4.022	2.407	1.732
XGBoost	3.728	2.225	1.612
LSTM	3.718	2.159	1.509
Random forest	3.846	2.369	1.713
Dynamic ensemble	3.698	2.160	1.480
Latent oracle	3.459	1.739	1.029

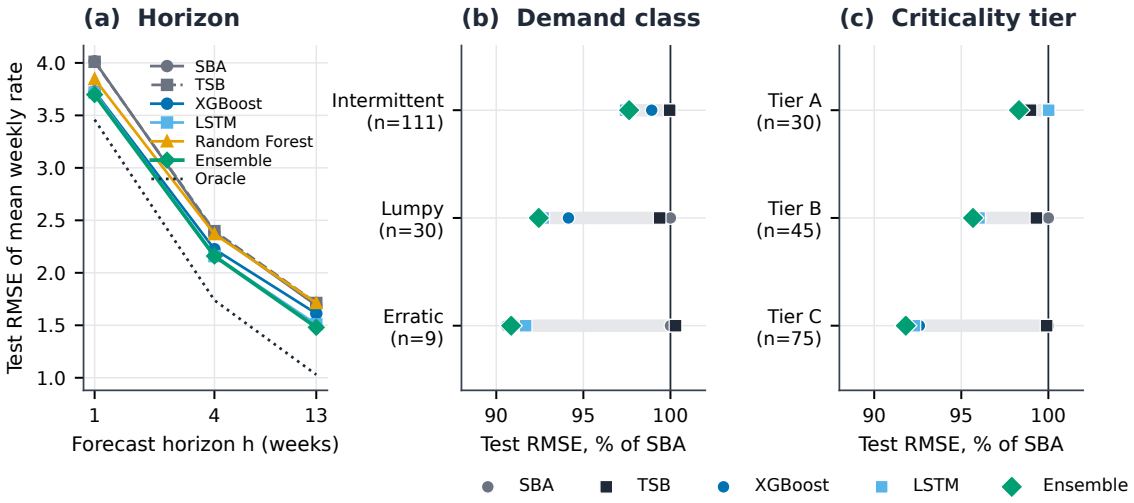

Figure 4. Seed-2026 test errors by horizon, demand class and criticality tier.

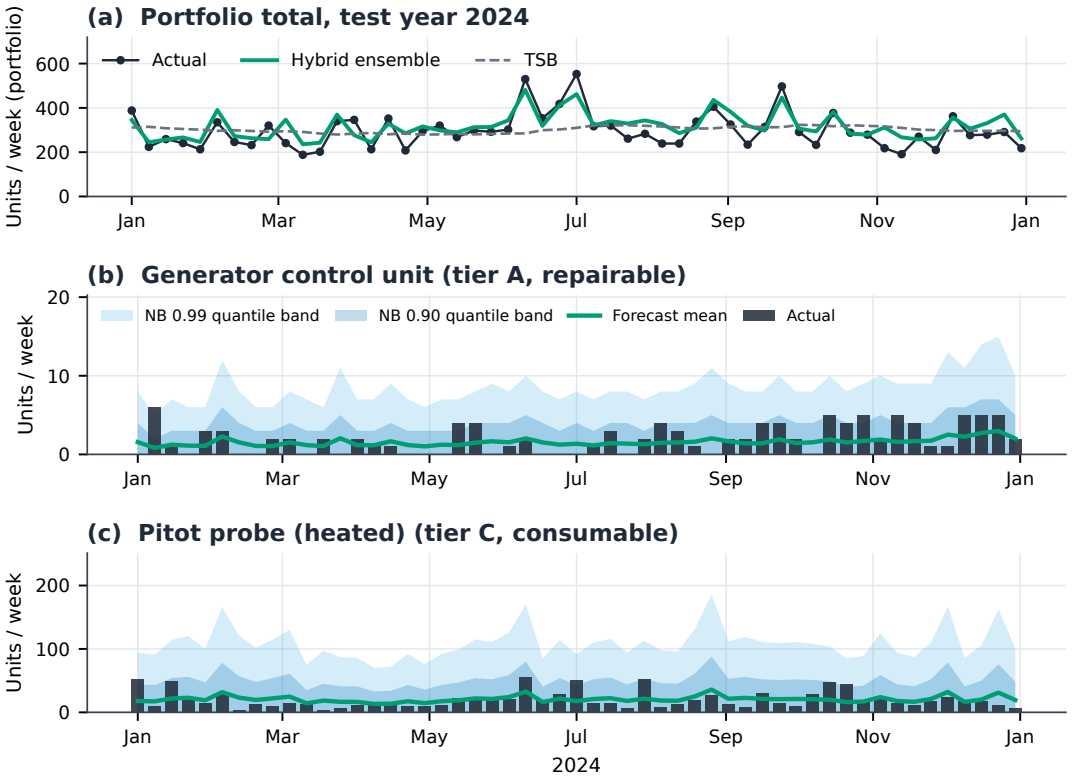

Figure 5. Seed-2026 portfolio and selected-part forecasting illustrations; predictive bands come from the NB2 layer.

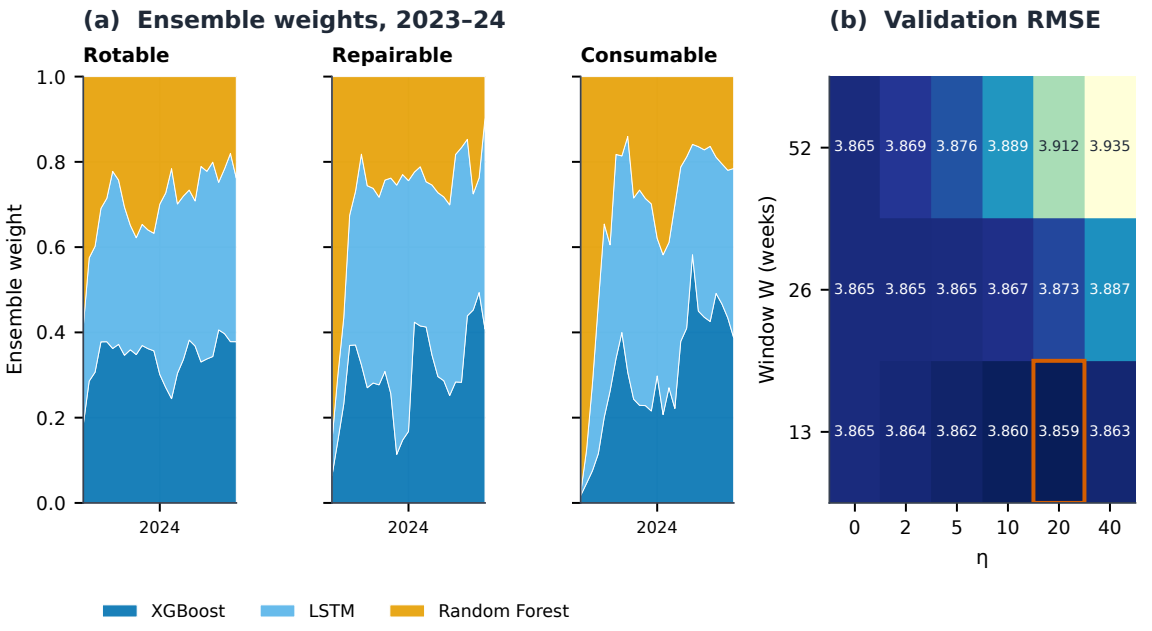

Figure 6. Time-varying model weights and validation hyperparameter sensitivity for the seed-2026 run.

feature-group deletion test and TreeSHAP contributions [47] are displayed in Figure 7; they are descriptive properties of a synthetic schedule-linked generator, not causal proof that equivalent features will improve airline forecasting.

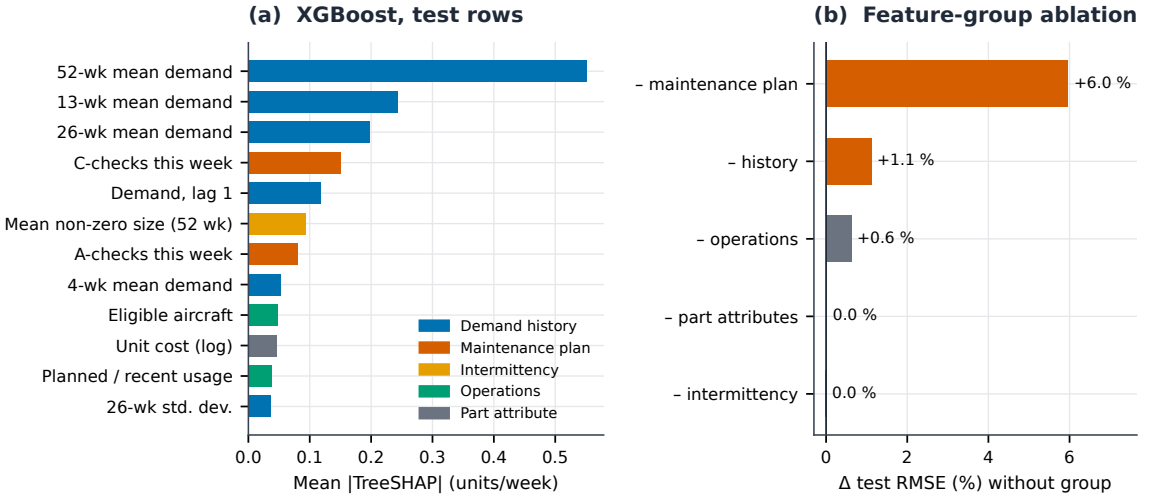

Figure 7. Seed-2026 TreeSHAP contributions and removed-feature-group RMSE sensitivity.

5.4. Leakage-free NB2 calibration and one-week probabilistic scores

Class-wise dispersion parameters were fitted for $H = 1, 4, 13$ using only outcome windows fully contained in the 2023 validation period. The NB2 layer of the hybrid yields one-week CRPS 1.077 compared with 1.193 for the normal approximation, and empirical NB2 coverage 0.971 at nominal 0.95. These coverage values are properties of the simulator and discrete predictive quantiles; the wide-tailed stock protection changes costs but does not establish calibration under actual airline lead-time variation.

Table 3. 2024 one-week predictive scores using strictly in-validation NB2 dispersion fits, seed 2026.

Engine	Distribution	CRPS	Coverage 0.50	0.90	0.95	0.99
Hybrid ensemble	NB2	1.077	0.696	0.940	0.971	0.993
	Normal	1.193	–	0.897	0.918	0.953
TSB	NB2	1.136	0.688	0.921	0.958	0.988
	Normal	1.250	–	0.887	0.915	0.950
SBA	NB2	1.142	0.684	0.910	0.953	0.986
	Normal	1.244	–	0.882	0.909	0.946

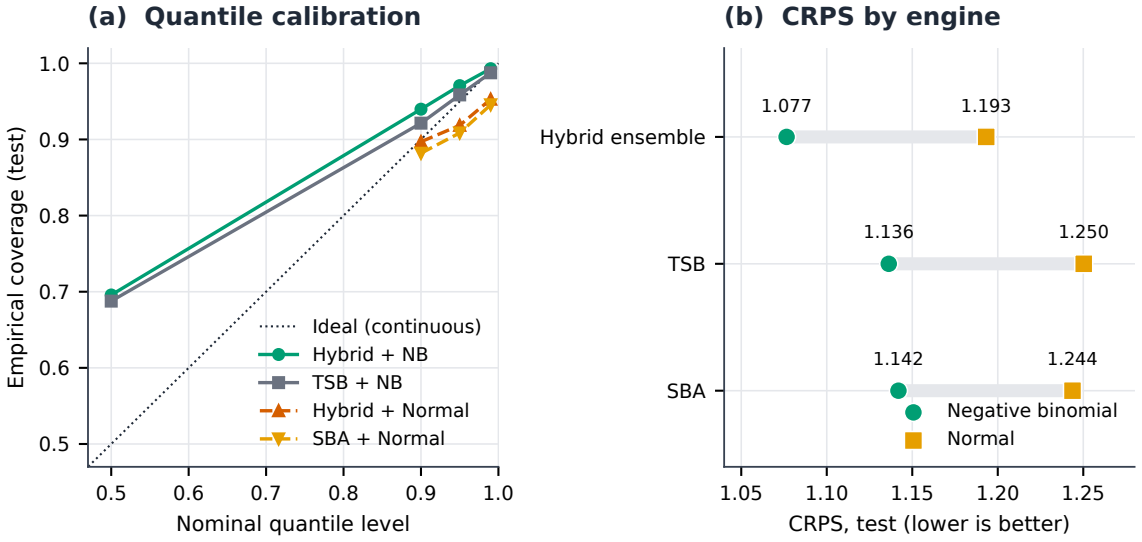

Figure 8. Empirical quantile coverage and CRPS under validation-boundary-controlled calibration.

5.5. Full end-to-end multiple-seed robustness

We generated an *entire* independent synthetic panel for each of seeds 2026, 2027 and 2028 and ran all three forecasting models, the dynamic and static ensembles, class-wise NB2 calibration and six inventory policies on each of them. We repeated the same end-to-end procedure at schedule-coupling strength $\gamma = 0.25$, where the scheduled demand contribution is $k_i[\gamma C_{i,t} + (1-\gamma)\bar{C}_i]$ rather than $k_i C_{i,t}$. The generator mechanisms, 2019–2024 chronology, validation-boundary rule, policy settings and 20 common-random-number inventory replications are held fixed; model random seeds and generated demand differ by run. Each run uses 14 XGBoost search trials, three 120-tree RF configurations and three TensorFlow/Keras LSTM seeds over a three-configuration hyper-parameter grid. The study is therefore an *end-to-end robustness assessment* across generators and coupling strengths.

Table 4. Full three-model end-to-end results on independently generated synthetic panels (20 inventory replications each). All inventory calibration masks enforce $t + H \leq t_{\text{validation,end}}$. Cost savings are not field estimates.

Coupling	Seed	TSB RMSE	Hybrid RMSE	Equal RMSE	P5/P1 saving (%)	P5/best uniform (%)
1.00	2026	4.010	3.698	3.709	+53.7	+22.2
1.00	2027	3.734	3.434	3.430	+14.5	-18.7
1.00	2028	4.234	3.866	3.884	+55.0	+11.6
0.25	2026	3.453	3.416	3.413	+12.2	-40.5
0.25	2027	3.336	3.302	3.302	+63.0	+32.1
0.25	2028	3.855	3.837	3.837	+38.5	0.0

Table 5. Independent-run uncertainty of 2024 cost saving against P1 (within-seed, part-bootstrap 95% CIs). With only three seeds per regime these are not confidence intervals for the population of generators.

Coupling	Seed	P1 total (USD M)	P5 total (USD M)	Saving [part-CI]
1.00	2026	12.76	5.90	53.7% [27%, 68%]
1.00	2027	6.40	5.47	14.5% [-9%, 32%]
1.00	2028	11.73	5.28	55.0% [20%, 72%]
0.25	2026	7.49	6.57	12.2% [-32%, 43%]
0.25	2027	15.42	5.71	63.0% [36%, 75%]
0.25	2028	11.71	7.20	38.5% [16%, 52%]

In the three strong-coupling runs the proposed P5 policy reduces cost relative to the fixed P1 baseline by 41.1% on average (sample SD 23.0 percentage points), ranging from 14.5% to 55.0%; the number of seeds is too small for a meaningful normality-based claim about generator-population uncertainty. Under weaker coupling the forecasting and inventory comparisons must be interpreted relative to each generated panel, not as results from the same observed demand sequence. The cheapest uniform service-level policy in the tables is selected using the *test-year costs*, so it is an optimistically strong reference and must not be interpreted as a prospective validation-selected benchmark. Table 6 reports every policy on every panel. P5 lowers cost against P1 in all six panels (12.2–63.0%; at $\gamma = 0.25$ the mean is 37.9% with sample SD 25.4 percentage points), but it is the cheapest of the five deployable policies in only four. The exceptions are seed 2027 at $\gamma = 1$, where P4 (NB2, uniform 0.95) is cheapest, and seed 2026 at $\gamma = 0.25$, where P2 (TSB, normal, tiers) is cheapest and P5 is the most expensive of P2–P5. Isolating single design changes, replacing the normal by the NB2 quantile (P3→P5), replacing uniform by tiered service targets (P4→P5) and replacing TSB by the hybrid forecast (P2→P3) each lowered cost in four of the six panels; the first two sets of exceptions coincide with the two panels above. Against the best uniform policy chosen in test-year hindsight, P5 ranges from −40.5 to +32.1%, and the part-bootstrap 95% interval excludes zero only for the two negative panels (seed 2027 at $\gamma = 1$ and seed 2026 at $\gamma = 0.25$); no panel shows a significant advantage. Over the 20-cell shortage/holding grid, the P5-versus-P1 saving per panel reaches as low as −100.2% (seed 2026, $\gamma = 0.25$) and as high as +84.8% (seed 2027, $\gamma = 0.25$), so the economic conclusion depends on the assumed penalties.

Table 6. Cost saving (%) against P1 for every policy on each of the six generated panels (20 replications). Bold marks the cheapest of the five deployable policies P1–P5; P6 uses the latent expectation and is a bound, not a deployable policy.

Coupling	Seed	P2	P3	P4	P5	P6 (bound)	Cheapest of P1–P5
1.00	2026	34.4	28.7	40.7	**53.7**	69.9	P5
1.00	2027	12.4	15.8	**22.5**	14.5	38.8	P4
1.00	2028	26.7	31.8	49.5	**55.0**	65.1	P5
0.25	2026	**26.0**	17.9	17.8	12.2	51.9	P2
0.25	2027	28.1	31.4	47.7	**63.0**	75.7	P5
0.25	2028	26.6	27.1	37.1	**38.5**	63.5	P5

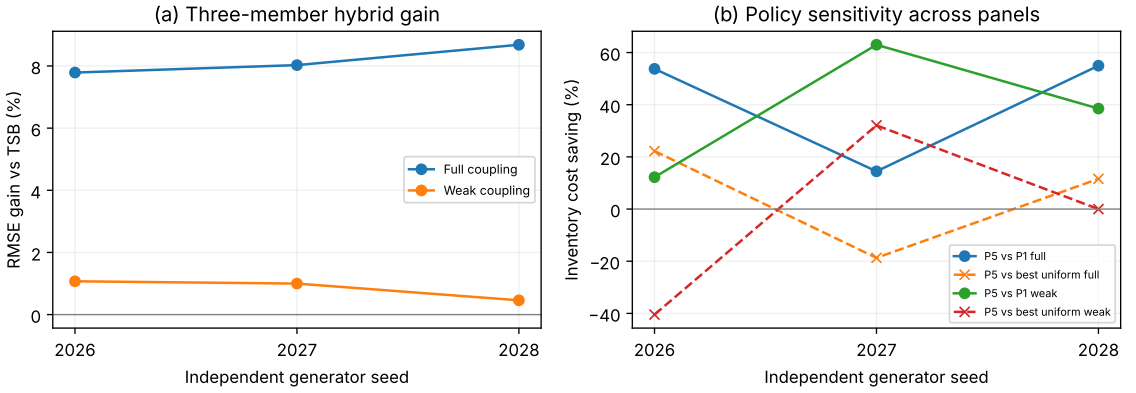

Figure 9. Independent end-to-end experiments: relative hybrid RMSE gain against TSB and percentage inventory savings against the fixed P1 and test-hindsight best uniform baseline. Each marker is one generated panel; the horizontal line marks zero.

5.6. Inventory policies and cost sensitivity

All six policies, the normal/NB2 distribution fits, bootstrap intervals, uniform-service sweeps, shortage/holding sensitivity grid, and tier fill rates use boundary-controlled dispersion estimates. Table 7 and Figure 10 report the primary seed-2026 run. P1 costs \$12.76 M and P5 costs \$5.90 M, a 53.7% saving (part-bootstrap 95% interval [27%, 68%]). Against the baseline using a *test-hindsight* optimal uniform service level, however, saving is 22.2% (interval [-3%, 40%]). This distinction makes the assumption-dependent nature of the result explicit. Over the prespecified shortage multiplier 0.25–4 and annual holding rate 10–40% grid, the P5-versus-P1 cost saving varies from -15.4% to 76.9%. The magnitudes should not be transferred to a real carrier without observed AOG penalties, supplier histories, maintenance plan revisions and cost-of-capital measurements.

Table 7. 2024 inventory simulation for seed 2026 (20 replications). Monetary quantities in million USD; the 95% cost-saving intervals resample parts.

Policy	Cost	Holding	Shortage	Ordering	Fill (%)	Stockout wk	Stock value	Saving vs P1
P1 SBA, normal, uniform	12.76	2.12	10.40	0.24	90.1	2.13	10.41	–
P2 TSB, normal, tiers	8.37	3.20	4.96	0.20	90.2	1.69	15.72	34% [14%, 47%]
P3 Hybrid, normal, tiers	9.10	2.82	6.06	0.22	90.4	1.81	13.85	29% [12%, 38%]
P4 Hybrid, NB2, uniform	7.56	3.17	4.19	0.21	97.3	0.82	15.53	41% [20%, 53%]
P5 Hybrid, NB2, tiers	5.90	3.74	1.96	0.21	95.3	1.03	18.33	54% [27%, 68%]
P6 Oracle, NB2, tiers	3.84	3.02	0.59	0.23	97.1	0.82	14.83	70% [50%, 79%]

6. Discussion

Scientific scope and generator dependence. The present work tests an integrated modelling *architecture* with synthetic data; it cannot infer airline-wide savings. The multi-seed experiments show that the forecast advantage

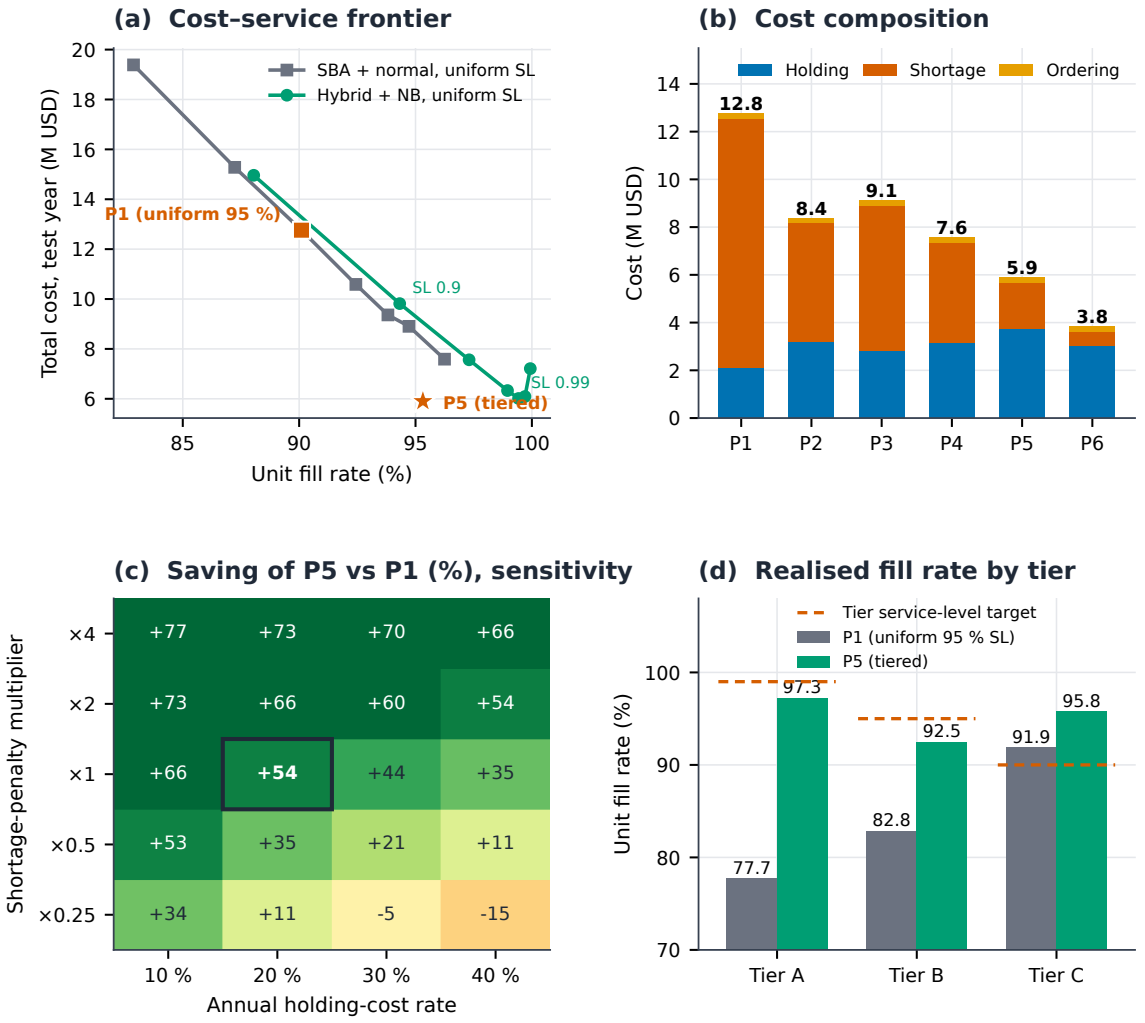

Figure 10. Seed-2026 inventory service–cost frontier, cost decomposition, penalty/holding sensitivity and realised tier fill rates. All distribution-dependent charts use boundary-controlled dispersion estimates.

and downstream stocking cost depend on the chosen generator and the strength of schedule-to-demand coupling. These are controlled stress tests, not an empirical external-validity claim. Recent aviation evidence using real repair and demand records provides an important contrasting context [27–30]. The related conference paper by overlapping authors [31] is cited above and disclosed in the Declarations.

NB2 calibration, tiering and dynamic weights. Reorder decisions respond to upper lead-time-demand quantiles, not solely to one-week RMSE. Strongly intermittent positive-count tails can be missed by normal approximations, while AHP tiers explicitly allocate higher service targets to critical components. In the six panels, replacing the normal by the NB2 quantile (P3→P5) and replacing uniform by tiered service targets (P4→P5) each lowered cost in four panels but not in seed 2027 at $\gamma = 1$ or seed 2026 at $\gamma = 0.25$, so neither benefit is uniform. The six-policy comparison is informative but is not a full factorial design and does not separate every causal mechanism. Dynamic weighting, by contrast, adds estimation noise when member errors are correlated; its improvement over equal or static ensembles is not stable across independent runs [3, 7, 32]. The simpler combination remains a credible operational competitor.

Uncertainty and managerial interpretation. Bootstrap intervals resampling 150 parts quantify within-panel heterogeneity; they do *not* quantify uncertainty across generator mechanisms or across real airlines. With only three seeds per coupling regime, the sample spread is descriptive. Inventory replications use the same realised demands and common random numbers. The β values used to choose reorder quantiles represent cycle-service targets, whereas reported unit fill rate is a realised metric. The shortage penalties \$20 000/\$2 500/\$300 per backordered unit-week are assumed, not operator-estimated. The deterministic quoted-lead-time NB2 quantile remains mismatched to a random-lead-time delivery simulator, potentially biasing safety stock.

Limitations and future validation. The TensorFlow/Keras architecture and Tweedie loss follow the specification above, but library versions and training hardware may change numerical results. The validation mask removes *dispersion-boundary leakage* from the NB2 calibration. The experiment does not address other forms of hidden data quality problems or supply disruption, actual maintenance schedule forecast errors, alternate AHP matrices, supplier lead-time mixture calibration, repair loops, or external tests. A real-operator replication and prospectively tuned cost/service policies are required before deployment. DeepAR, TFT, N-BEATS and conformal predictive distributions are relevant comparison methods but not experimental baselines tested here [15–18].

7. Conclusion

The evidence consists of six independent, end-to-end synthetic experiments, with leakage-free negative-binomial horizon calibration, three forecasting model families, dynamically and statically weighted combinations, and all six inventory policies evaluated under uniform assumptions. The proposed policy lowers simulated cost against the fixed P1 baseline in every panel (12.2–63.0%), but it is the cheapest deployable policy in only four of six panels and is never significantly better than the best uniform policy chosen in test-year hindsight. The hybrid forecast has lower RMSE than TSB in every panel, with small margins that vary with generated demand and maintenance coupling. No stable incremental gain from dynamic over simple static combination is established. The deterministic-lead-time approximation and assumed costs further constrain management interpretation. These findings motivate operator-data validation and an empirical lead-time mixture before deployment rather than unconditional savings claims.

Funding, Competing Interests, and Data Availability

Funding. This research received no funding.

Competing interests. The authors declare that they have no competing interests.

Data and code availability. The accompanying reproducibility archive contains the complete synthetic panels for each of the six runs, the generator and its parameter settings, the calibration code, the TensorFlow/Keras LSTM implementation, generated JSON and CSV result tables, saved forecast arrays, the figure-generation script

and the manuscript source. The dataset is synthetic and is not represented as confidential airline transaction data. No public archival DOI or permanent repository has been assigned.

Previous dissemination. A closely related conference paper by overlapping authors is cited as [31].

A. Experimental configuration

Table 8. Six-panel end-to-end training and simulation settings.

Component	Configuration
XGBoost	Tweedie objective; 14 seeded random-search configurations on 2023 validation, per-horizon early stopping, refit on 2019–2023; actual best configurations and rounds saved separately for every run
Random forest	120 trees; three configurations (depth/minimum leaf/feature fraction 8/10/0.4, 12/5/0.4, 16/5/0.6) compared on validation, best one retained per panel
LSTM	TensorFlow/Keras: 26-week input window, 32/64 hidden units by validation search, eight-dimensional part embedding, 64-wide dense layer, three softplus outputs, Tweedie deviance $\rho = 1.3$, dropout 0.2/0.1, Adam learning rates 0.001/0.003, batch size 256, up to 30 epochs, three training seeds per run; validation-based early stopping
Ensemble	Class-wise dynamic softmax weighting; $\eta \in \{0, 2, 5, 10, 20, 40\}$, $W \in \{13, 26, 52\}$, selected separately for every panel; updates use matured outcomes only
Probabilistic fit	Class-wise NB2 dispersion MLE using only origins t with $t + H \leq t_{\text{val,end}}$; one-, four-, thirteen-week horizons
Inventory	Six fixed named policy variants; annual holding 20%; shortage costs by tier 20 000/2 500/300 USD per backordered unit-week; 20 replications per policy, shared random streams, 26-week warmup; holding and penalty sensitivity grid as specified in methodology
Generator	150 parts; independent seeds 2026, 2027 and 2028 at coupling 1.00 and 0.25; 2019–2024 weekly chronology

B. Reproducibility checklist

- Each of six panels uses a separate data-generation seed; the same 150-part scale, calendar and procedure are maintained, with coupling strengths 1 and 0.25.
- For horizon H, NB2 validation-origin masks require $t + H \leq t_{\text{val,end}}$; a unit test fails if a fit target crosses into 2024.
- Three forecasting families and all six inventory policies were run on every panel, with twenty inventory replications and common random streams for policy comparison.
- The LSTM uses TensorFlow/Keras with a three-configuration search and three seeds per panel.
- The latent oracle is used only for a retrospective reference, never as a fitted predictor.
- Figures are generated from the seed-2026 outputs, while the robustness tables use all panel-specific JSON files.

References

[1] A. Regattieri, M. Gamberi, R. Gamberini, R. Manzini, Managing lumpy demand for aircraft spare parts, Journal of Air Transport Management 11 (6) (2005) 426–431.

[2] A. A. Ghobbar, C. H. Friend, Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model, Computers & Operations Research 30 (14) (2003) 2097–2114.

[3] D. Prak, R. Teunter, A. Syntetos, On the calculation of safety stocks when demand is forecasted, European Journal of Operational Research 256 (2) (2017) 454–461.

[4] J. D. Croston, Forecasting and stock control for intermittent demands, Operational Research Quarterly 23 (3) (1972) 289–303.

[5] A. A. Syntetos, J. E. Boylan, The accuracy of intermittent demand estimates, International Journal of Forecasting 21 (2) (2005) 303–314.

[6] R. H. Teunter, A. A. Syntetos, M. Z. Babai, Intermittent demand: Linking forecasting to inventory obsolescence, European Journal of Operational Research 214 (3) (2011) 606–615.

[7] G. Claeskens, J. R. Magnus, A. L. Vasnev, W. Wang, The forecast combination puzzle: A simple theoretical explanation, International Journal of Forecasting 32 (3) (2016) 754–762.

[8] A. A. Syntetos, J. E. Boylan, J. D. Croston, On the categorization of demand patterns, Journal of the Operational Research Society 56 (5) (2005) 495–503.

[9] K. Nikolopoulos, A. A. Syntetos, J. E. Boylan, F. Petropoulos, V. Assimakopoulos, An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis, Journal of the Operational Research Society 62 (3) (2011) 544–554.

[10] A. A. Ghobbar, C. H. Friend, Sources of intermittent demand for aircraft spare parts within airline operations, Journal of Air Transport Management 8 (4) (2002) 221–231.

[11] W. Wang, A. A. Syntetos, Spare parts demand: Linking forecasting to equipment maintenance, Transportation Research Part E 47 (6) (2011) 1194–1209.

[12] R. D. Snyder, J. K. Ord, A. Beaumont, Forecasting the intermittent demand for slow-moving inventories: A modelling approach, International Journal of Forecasting 28 (2) (2012) 485–496.

[13] M. W. Seeger, D. Salinas, V. Flunkert, Bayesian intermittent demand forecasting for large inventories, in: Advances in Neural Information Processing Systems 29, 2016.

[14] S. Kolassa, Evaluating predictive count data distributions in retail sales forecasting, International Journal of Forecasting 32 (3) (2016) 788–803.

[15] D. Salinas, V. Flunkert, J. Gasthaus, T. Januschowski, DeepAR: Probabilistic forecasting with autoregressive recurrent networks, International Journal of Forecasting 36 (3) (2020) 1181–1191.

[16] B. Lim, S. O. Arik, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, International Journal of Forecasting 37 (4) (2021) 1748–1764. doi:10.1016/j.ijforecast.2021.03.012.

[17] B. N. Oreshkin, D. Carpov, N. Chapados, Y. Bengio, N-BEATS: Neural basis expansion analysis for interpretable time series forecasting, in: International Conference on Learning Representations (ICLR), 2020.
URL https://openreview.net/forum?id=r1ecqn4YwB

[18] K. Stankeviciute, A. M. Alaa, M. van der Schaar, Conformal time-series forecasting, in: Advances in Neural Information Processing Systems 34 (NeurIPS), 2021.
URL https://proceedings.neurips.cc/paper/2021/hash/312f1ba2a72318edaaa995a67835fad5-Abstract.html

[19] R. S. Gutierrez, A. O. Solis, S. Mukhopadhyay, Lumpy demand forecasting using neural networks, International Journal of Production Economics 111 (2) (2008) 409–420.

[20] N. Kourentzes, Intermittent demand forecasts with neural networks, International Journal of Production Economics 143 (1) (2013) 198–206.

[21] F. Lolli, R. Gamberini, A. Regattieri, E. Balugani, T. Gatos, S. Gucci, Single-hidden layer neural networks for forecasting intermittent demand, International Journal of Production Economics 183 (2017) 116–128.

[22] H. Hewamalage, C. Bergmeir, K. Bandara, Recurrent neural networks for time series forecasting: Current status and future directions, International Journal of Forecasting 37 (1) (2021) 388–427.

[23] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, in: Advances in Neural Information Processing Systems 35 (NeurIPS 2022 Datasets and Benchmarks Track), 2022.

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

[25] J. M. Bates, C. W. J. Granger, The combination of forecasts, Operational Research Quarterly 20 (4) (1969) 451–468.

[26] A. Timmermann, Forecast combinations, in: G. Elliott, C. W. J. Granger, A. Timmermann (Eds.), Handbook of Economic Forecasting, Vol. 1, Elsevier, Amsterdam, 2006, pp. 135–196.

[27] M. del Olmo, R. Domingo, Intermittent demand forecasting of aerospace rotable parts. a framework for unpredictable flight patterns, Computers & Industrial Engineering 214 (2026) 111917. doi:10.1016/j.cie.2026.111917.

[28] G. Chen, et al., Demand forecasting and inventory optimization of aviation rotable parts: A Markov queuing approach, Aerospace 13 (7) (2026) 620. doi:10.3390/aerospace13070620.

[29] M. Brito, D. Dinis, A. Barroso, Spare parts management in the aviation maintenance industry: A literature review, Procedia Computer Science 277 (2026) 2961–2970. doi:10.1016/j.procs.2026.02.332.

[30] M. Brito, D. Dinis, A. Barroso, Sustainable spare parts management in the aviation maintenance industry: A systematic literature review on forecasting approaches, Sustainability 18 (17) (2026) 8709. doi:10.3390/su18178709.

[31] B.-E. Badi, O. Bouazaoui, I. Lagrat, M. Ouhdan, D. Serrou, Proposition of a hybrid machine learning framework for aircraft spare parts forecasting and inventory optimization, in: Proceedings of the 8th International Conference on Logistics Operations Management, no. 2035 in Lecture Notes in Networks and Systems, Springer, 2027, pp. 75–87, first online 28 July 2026. doi:10.1007/978-3-032-29680-1_7.

[32] D. Prak, R. Teunter, A general method for addressing forecasting uncertainty in inventory models, International Journal of Forecasting 35 (1) (2019) 224–238.

[33] A. Bacchetti, N. Saccani, Spare parts classification and demand forecasting for stock control: Investigating the gap between research and practice, Omega 40 (6) (2012) 722–737.

[34] J. E. Boylan, A. A. Syntetos, G. C. Karakostas, Classification for forecasting and stock control: a case study, Journal of the Operational Research Society 59 (4) (2008) 473–481.

[35] T. L. Saaty, The Analytic Hierarchy Process, McGraw-Hill, New York, 1980.

[36] E. A. Silver, D. F. Pyke, R. Peterson, Inventory Management and Production Planning and Scheduling, 3rd Edition, Wiley, New York, 1998.

[37] E. L. Porteus, Foundations of Stochastic Inventory Theory, Stanford University Press, Stanford, 2002.

[38] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (2012) 281–305.

[39] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[40] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[41] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations (ICLR 2015), 2015.

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

[43] S. Prestwich, R. Rossi, S. A. Tarim, B. Hnich, Mean-based error measures for intermittent demand forecasting, International Journal of Production Research 52 (22) (2014) 6782–6791.

[44] R. J. Hyndman, A. B. Koehler, Another look at measures of forecast accuracy, International Journal of Forecasting 22 (4) (2006) 679–688.

[45] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (6) (1945) 80–83.

[46] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[47] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence 2 (2020) 56–67.