TSB-HB: A HIERARCHICAL BAYESIAN EXTENSION OF THE TSB MODEL FOR INTERMITTENT DEMAND FORECASTING

A PREPRINT

Zong-Han Bai*
hu110048138@gapp.nthu.edu.tw

Po-Yen Chu*
b10704031@ntu.edu.tw

ABSTRACT

Intermittent demand forecasting poses unique challenges due to sparse observations, cold-start items, and obsolescence. Classical models such as Croston, SBA, and the Teunter–Syntetos–Babai (TSB) method provide simple heuristics but lack a principled generative foundation. Deep learning models address these limitations but often require large datasets and sacrifice interpretability.

We introduce TSB-HB, a hierarchical Bayesian extension of TSB. Demand occurrence is modeled with a Beta–Binomial distribution, while nonzero demand sizes follow a Log-Normal distribution. Crucially, hierarchical priors enable partial pooling across items, stabilizing estimates for sparse or cold-start series while preserving heterogeneity. This framework yields a fully generative and interpretable model that generalizes classical exponential smoothing.

On the UCI Online Retail dataset, TSB-HB achieves lower RMSE and RMSSE than Croston, SBA, TSB, ADIDA, IMAPA, ARIMA and Theta, and on a subset of the M5 dataset it outperforms all classical baselines we evaluate. The model provides calibrated probabilistic forecasts and improved accuracy on intermittent and lumpy items by combining a generative formulation with hierarchical shrinkage, while remaining interpretable and scalable.

1 Introduction

Intermittent demand—long runs of zeros punctuated by occasional positives—arises in spare parts, long-tail retail, and hospital supplies. Planners face sparse histories, many cold-start items, and possible obsolescence; miscalibrated forecasts propagate to stockouts or excess inventory, making probabilistic accuracy operationally consequential.

Two broad modeling routes dominate practice. The first comprises exponential-smoothing (ES) decompositions that separate occurrence (whether demand happens) from positive size (how much, conditional on occurrence). Croston's method and its bias corrections (e.g., the Syntetos–Boylan Approximation) apply simple smoothers to interarrival times and positive sizes and combine them to obtain a mean-per-period forecast (Croston, 1972; Syntetos and Boylan, 2005). To address obsolescence, the Teunter–Syntetos–Babai (TSB) approach replaces interarrival smoothing by per-period smoothing of the occurrence probability, which updates on zeros and tends to decay forecasts more quickly under extended zero runs (Teunter et al., 2011). These heuristics remain attractive because they are simple, scalable, and interpretable; yet they lack a principled generative foundation, treat items in isolation (foregoing information sharing that could stabilize estimates in sparse settings), and are sensitive to smoothing constants.

The second route emphasizes generative, distributional forecasting with explicit zero handling (e.g., hurdle or zero-inflated constructions), often combined with hierarchical Bayesian (HB) priors to partially pool information across items. Such models produce calibrated predictive distributions and admit covariates and mixed-effects structures, improving stability for sparse series and cold-start items (e.g., Pitkin et al., 2024). Deep learning variants extend this direction further but may require large datasets and can be harder to deploy or interpret in operations.

This paper proposes a minimalist bridge between these routes. We retain the TSB multiplicative structure—occurrence probability times positive size—but cast each factor in a hierarchical, fully generative form with empirical-Bayes (EB) estimation (Robbins, 1992). Occurrence is modeled with a Beta–Binomial layer; positive sizes use a right-skew dis-

*Equal contribution

tribution on the log scale (Log-Normal by default, with a Gamma variant for robustness). Hierarchical priors enable partial pooling across items, yielding shrinkage toward panel-level ranges when evidence is weak while preserving heterogeneity as observations accumulate. The result is an interpretable model that provides coherent cold-start behavior through pooling; modeling time-varying obsolescence is outside our present scope and left for future work.

Our contributions are threefold. First, we formulate a hierarchical, generative analogue of TSB that preserves the operational intuition of its multiplicative decomposition while enabling panel-wise information sharing. Second, we derive closed-form EB estimators for both occurrence and size, leading to one-shot hyperparameter fitting and $O(N)$ per-epoch forecasting suitable for large catalogs. Third, in experiments on the Online Retail and M5 datasets (Chen, 2015; Howard et al., 2020), we observe lower RMSE and RMSSE than Croston, SBA, TSB, ADIDA (Nikolopoulos et al., 2011), IMAPA (Petropoulos et al., 2019), ARIMA (Box et al., 2015), and Theta (Assimakopoulos and Nikolopoulos, 2000) (implementations via `statsforecast` (Garza et al., 2022)), with competitive accuracy relative to selected deep-learning baselines, alongside improved calibration and service-level alignment for lumpy and intermittent items.

The remainder of the paper develops the model and estimation procedure, analyzes its shrinkage and computational profile, and evaluates performance and calibration across intermittency regimes, followed by limitations and directions for incorporating covariates and cross-item dependence.

2 Related Work

Classical intermittent-demand forecasting began with Croston's decomposition, which applies independent exponential smoothers to the inter-demand interval and the positive demand size, then forms a ratio to obtain a per-period mean forecast (Croston, 1972). Subsequent studies documented bias in the ratio estimator and proposed corrections such as the Syntetos–Boylan Approximation (SBA) and related variants that trade bias for variance under different assumptions. Overall accuracy in this family is highly sensitive to the choice of smoothing constants, and performance rankings can shift across error metrics and data regimes.

To better handle obsolescence, the Teunter–Syntetos–Babai (TSB) method smooths the period-by-period occurrence probability instead of interarrival times, ensuring updates on zeros and faster decay when demand ceases (Teunter et al., 2011). Follow-on refinements (e.g., mTSB) adjust the probability update while remaining within the ES heuristic paradigm and inheriting its dependence on smoothing hyperparameters (Yang et al., 2021). These approaches retain strong practical appeal due to their simplicity and interpretability but do not arise from a coherent generative model and do not share information across items.

An alternative line of work adopts generative, distributional models that explicitly account for zero inflation via hurdle or zero-inflated mechanisms, often coupled with hierarchical priors to enable partial pooling across a panel. Such models improve calibration, facilitate covariate inclusion, and stabilize estimates for sparse series, with recent applications to retail demand and related count processes (e.g., Pitkin et al., 2024). Deep learning methods extend this distributional perspective and can offer strong accuracy on large-scale datasets, though they may trade off interpretability and impose higher data or operational costs.

Intermittent-demand studies do not commit to a single law for positive sizes. In TSB's numerical study, the logarithmic distribution was chosen for sizes because it is discrete, offers a tunable variance-to-mean ratio, and—combined with Poisson arrivals—yields negative-binomial totals; however, the authors explicitly note that other distributions such as Lognormal and Gamma could have been used (Teunter et al., 2011).

Taken together, prior work has not, to our knowledge, simultaneously provided (i) a coherent generative model that preserves the operational intuition of TSB's multiplicative occurrence–size structure and (ii) panel-wise partial pooling with closed-form, scalable (e.g., EB) estimators suited to large catalogs. Approaches that achieve one of these aspects often either treat items in isolation or rely on heavier inference machinery, limiting scalability on intermittent retail panels.

3 Methodology

3.1 Setup, Notation, and Assumptions

We consider a panel of items $i \in \{1, \dots, N\}$ with item-specific horizons $t \in \{1, \dots, T_i\}$ and a fixed forecast origin $t_{0,i} \in \{1, \dots, T_i\}$. The in-sample (initialization) window is $\mathcal{T}_i^{\text{in}} = \{1, \dots, t_{0,i}\}$ and the evaluation window is $\mathcal{T}_i^{\text{oos}} = \{t_{0,i} + 1, \dots, T_i\}$. When a common origin is used, we write t_0, $\mathcal{T}^{\text{in}}$, and $\mathcal{T}^{\text{oos}}$ without item subscripts.

For each (i,t) we observe nonnegative demand $Y_{i,t} \geq 0$ and define the occurrence indicator $p_{i,t} = [1]\{Y_{i,t} > 0\}$. When $p_{i,t} = 1$ a conditional positive-demand size is observed, denoted $s_{i,t} = Y_{i,t}$ and $\ell_{i,t} = \log s_{i,t}$; otherwise $s_{i,t}$ is unobserved. Let $n_i = |\mathcal{T}_i^{\text{in}}|$ and $m_i = \sum_{t \in \mathcal{T}_i^{\text{in}}} p_{i,t}$. When $m_i > 0$, define $\bar{\ell}_i = \frac{1}{m_i} \sum_{t \in \mathcal{T}_i^{\text{in}}: p_{i,t}=1} \ell_{i,t}$ and $s_{\ell,i}^2 = \frac{1}{m_i - 1} \sum_{t:p_{i,t}=1} (\ell_{i,t} - \bar{\ell}_i)^2$ (if $m_i > 1$).

Assumptions. (i) Fixed-origin, time-invariant per-period forecasting throughout $\mathcal{T}_i^{\text{oos}}$; (ii) conditional on $p_{i,t} = 1$, positive sizes follow a Log-Normal distribution ($\log S_{i,t} \sim \mathcal{N}(\mu_i, \sigma^2)$). For the ablation study, we also test a Gamma-Gamma conjugate model as an alternative for the size distribution (Section 4.6); (iii) across items, (π_i, μ_i) are independent with $\mu_i \sim \mathcal{N}(\mu_0, \tau^2)$; (iv) the initialization window is used to estimate hyper-parameters and item-level posteriors; (v) no covariates are used in the main model.[2] We emphasize that the log-size distribution is right-skew accommodating via a log-normal mechanism; we do not make Pareto-type heavy-tail claims.

3.2 TSB-HB: A multiplicative forecast with hierarchical shrinkage

We retain TSB's multiplicative forecast form and replace per-series EWMAs by hierarchical Bayesian (HB) shrinkage. For each item i:

$$\widehat{Y}_i \;=\; \widehat{\pi}_i \cdot \widehat{S}_i. \tag{1}$$

Occurrence (Beta–Binomial, empirical Bayes). Let $\pi_i \sim \text{Beta}(\alpha, \beta)$ and $m_i \mid \pi_i \sim \text{Binomial}(n_i, \pi_i)$ on $\mathcal{T}_i^{\text{in}}$. Write the prior in mean–precision form (μ_π, ϕ) with $\mu_\pi = \alpha/(\alpha+\beta)$ and $\phi = \alpha + \beta$. The posterior mean is

$$\begin{aligned} \widehat{\pi}_i = \frac{\alpha + m_i}{\alpha + \beta + n_i} &= \lambda_i \cdot \frac{m_i}{n_i} + (1 - \lambda_i) \cdot \mu_\pi, \\ \lambda_i &= \frac{n_i}{n_i + \phi} \in [0,1]. \end{aligned} \tag{2}$$

We estimate (μ_π, ϕ) from the panel via maximum marginal likelihood on $\{(m_i, n_i)\}_{i=1}^N$. The corresponding Beta–Binomial marginal likelihood in (α, β) is

$$L(\alpha, \beta) \propto \prod_{i=1}^{N} \frac{\text{B}(m_i + \alpha, n_i - m_i + \beta)}{\text{B}(\alpha, \beta)},$$

with $(\alpha, \beta) = (\mu_\pi \phi, (1 - \mu_\pi)\phi)$.

Conditional size on the log scale (Normal–Normal EB). For $p_{i,t} = 1$, $\ell_{i,t} = \log s_{i,t}$ follows $\ell_{i,t} = \mu_i + \varepsilon_{i,t}$, $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma^2)$ with $\mu_i \sim \mathcal{N}(\mu_0, \tau^2)$. Let $(\widehat{\mu}_0, \widehat{\tau}^2, \widehat{\sigma}^2)$ be REML estimates from $\{\ell_{i,t} : p_{i,t} = 1,\, t \in \mathcal{T}_i^{\text{in}},\, i \leq N\}$. Given m_i positives and $\bar{\ell}_i$, the empirical-Bayes weight and posterior are

$$w_i = \frac{m_i \, \widehat{\tau}^2}{m_i \, \widehat{\tau}^2 + \widehat{\sigma}^2}, \tag{3}$$

$$\widehat{\mu}_i = w_i \, \bar{\ell}_i + (1 - w_i) \, \widehat{\mu}_0, \qquad \widehat{v}_{\mu,i} = \frac{\widehat{\sigma}^2 \, \widehat{\tau}^2}{m_i \, \widehat{\tau}^2 + \widehat{\sigma}^2}. \tag{4}$$

Integrating over μ_i and a next-period shock yields the log-normal predictive mean

$$\widehat{S}_i = \exp\Big(\widehat{\mu}_i + \tfrac{1}{2}(\widehat{\sigma}^2 + \widehat{v}_{\mu,i})\Big). \tag{5}$$

If $m_i = 0$, (3) gives $w_i = 0$, hence $(\widehat{\mu}_i, \widehat{v}_{\mu,i}) = (\widehat{\mu}_0, \widehat{\tau}^2)$ and (5) defines a coherent cold-start predictor.

Relation to TSB. Equation (1) is the same multiplicative form as TSB; the difference is that (2) and (5) derive from panel-pooled posteriors rather than per-period EWMAs. This yields interpretable shrinkage and $O(N)$ per-epoch complexity after a one-shot hyper-parameter fit.

3.3 Deterministic shrinkage properties

Conditioning on $(\widehat{\mu}_0, \widehat{\tau}^2, \widehat{\sigma}^2, \mu_\pi, \phi)$, the estimators are shrinkage rules—convex combinations of item and panel statistics with sample-size–dependent weights, paralleling classical Stein-type shrinkage behavior (James et al., 1961).

[2]Extensions with covariates can be obtained by (a) a logistic link on π_i and (b) a linear mixed model on $\ell_{i,t}$; we focus on the intercept-only panel pooling to isolate the contribution of hierarchical shrinkage.

Monotonicity and limits. From (3), w_i increases strictly in m_i with $w_i(0) = 0$ and $w_i \to 1$ as $m_i \to \infty$; thus $\widehat{\mu}_i$ in (4) moves monotonically from $\widehat{\mu}_0$ to $\bar{\ell}_i$ and $\widehat{v}_{\mu,i}$ decreases to 0. For occurrence, $\lambda_i = n_i/(n_i + \phi)$ in (2) increases in n_i, so $\widehat{\pi}_i$ shrinks from μ_π to m_i/n_i. Boundary cases: as $\widehat{\tau}^2 \to 0$ (no between-item heterogeneity), $w_i \to 0$ (complete pooling); as $\widehat{\sigma}^2 \to 0$ (no within-item noise), $w_i \to 1$ (no pooling).

Bracketing. Since $\widehat{\mu}_i \in [\min\{\bar{\ell}_i, \widehat{\mu}_0\}, \max\{\bar{\ell}_i, \widehat{\mu}_0\}]$ and $\widehat{v}_{\mu,i} \in [0, \widehat{\tau}^2]$, it follows from (5) that

$$L_i^- := \exp\big(\min\{\bar{\ell}_i, \widehat{\mu}_0\} + \tfrac{1}{2}\widehat{\sigma}^2\big), \tag{6}$$

$$L_i^+ := \exp\big(\max\{\bar{\ell}_i, \widehat{\mu}_0\} + \tfrac{1}{2}(\widehat{\sigma}^2 + \widehat{\tau}^2)\big), \tag{7}$$

$$\widehat{S}_i \in [\, L_i^-,\, L_i^+ \,]. \tag{8}$$

and $\widehat{S}_i$ is nondecreasing in w_i holding $(\bar{\ell}_i, \widehat{\mu}_0, \widehat{\sigma}^2, \widehat{\tau}^2)$ fixed.

Bias–variance decomposition on the log scale. Fix $(\widehat{\mu}_0, \widehat{\tau}^2, \widehat{\sigma}^2)$ and assume $\bar{\ell}_i \mid \mu_i \sim \mathcal{N}(\mu_i, \widehat{\sigma}^2/m_i)$. With $\widehat{\mu}_i = (1 - w_i)\widehat{\mu}_0 + w_i\,\bar{\ell}_i$ and $w_i = \frac{m_i\widehat{\tau}^2}{m_i\widehat{\tau}^2 + \widehat{\sigma}^2}$,

$$\widehat{\mu}_i - \bar{\ell}_i = (1 - w_i)(\widehat{\mu}_0 - \bar{\ell}_i).$$

Moreover, conditioning on the (unknown) μ_i rather than on $\bar{\ell}_i$ yields

$$\mathrm{Bias}(\widehat{\mu}_i \mid \mu_i) = (1 - w_i)(\widehat{\mu}_0 - \mu_i),\ \mathrm{Var}(\widehat{\mu}_i \mid \mu_i) = w_i^2\,\frac{\widehat{\sigma}^2}{m_i},$$

and hence

$$\mathbb{E}_{\bar{\ell}_i}\big[(\widehat{\mu}_i - \mu_i)^2 \mid \mu_i\big] = (1 - w_i)^2(\widehat{\mu}_0 - \mu_i)^2 + w_i^2\,\frac{\widehat{\sigma}^2}{m_i}.$$

Thus the mean squared error trades off a squared shrinkage bias term, which is large for small w_i (strong shrinkage), and a sampling variance term, which is large for large w_i (weak shrinkage).

3.4 Identifiability and large-sample guarantees

Let the true hyper-parameters be $\theta^\star = (\mu_0^\star, \tau^{2\star}, \sigma^{2\star}, \mu_\pi^\star, \phi^\star)$. (i) Conditional consistency given $\theta^\star$. By the law of large numbers, $\widehat{\pi}_i = (\alpha^\star + m_i)/(\alpha^\star + \beta^\star + n_i) \xrightarrow{p} \pi_i$ as $n_i \to \infty$. For sizes, $\bar{\ell}_i \xrightarrow{p} \mu_i$ as $m_i \to \infty$ and $w_i \to 1$, hence $\widehat{\mu}_i \xrightarrow{p} \mu_i$, $\widehat{v}_{\mu,i} \to 0$, and $\widehat{S}_i \xrightarrow{p} \exp(\mu_i + \frac{1}{2}\sigma^{2\star}) = \mathbb{E}[S_i \mid \mu_i]$. Therefore $\widehat{Y}_i = \widehat{\pi}_i\,\widehat{S}_i$ is consistent for $\pi_i\,\mathbb{E}[S_i \mid \mu_i]$ when $(n_i, m_i) \to \infty$. (ii) Identifiability and EB consistency. Under the random-intercepts model with independent errors, $(\mu_0, \tau^2, \sigma^2)$ are identifiable from the joint distribution of $\{\bar{\ell}_i, m_i\}$ (the case $\tau^{2\star} = 0$ is a boundary but identifiable complete-pooling regime). Standard LMM regularity conditions imply REML consistency for $(\widehat{\mu}_0, \widehat{\tau}^2, \widehat{\sigma}^2)$ as $N \to \infty$ with $\{m_i\}$ bounded or growing. For occurrence, the Beta–Binomial marginal likelihood is identifiable in (μ_π, ϕ); MLEs are consistent as $N \to \infty$ under mild moment conditions on $\{(m_i, n_i)\}$. (iii) Hyper-parameter uncertainty. Plug-in EB treats $\widehat{\theta}$ as fixed and thus underestimates predictive variance. A first-order delta correction and a parametric bootstrap that propagate uncertainty in $\widehat{\theta}$ are natural extensions but are not pursued in our experiments.

3.5 Computation and Complexity

One pass over the initialization window computes sufficient statistics in $O(M)$ time, $M = \sum_i n_i$, and $O(N)$ memory. Beta–Binomial hyper-parameters are obtained by maximizing a two-parameter marginal likelihood with per-iteration $O(N)$ cost (typical L-BFGS-B iterations $I \ll 100$). The LMM block uses standard REML solvers on aggregated per-item moments. Posterior means $(\widehat{\pi}_i, \widehat{S}_i)$ and forecasts (1) are $O(1)$ per item. Overall, training is $O(M + IN)$ and prediction is $O(N)$ per epoch.

Variant (for ablation). A Gamma alternative for sizes (TSB-HB-Gamma) replaces the log-normal block by a conjugate Gamma–Gamma model: conditional on an item-specific rate $\beta_{s,i}$, positive sizes follow $S_{i,t} \mid \beta_{s,i} \sim \mathrm{Gamma}(\alpha_s, \beta_{s,i})$, and the rates have a Gamma prior $\beta_{s,i} \sim \mathrm{Gamma}(a, b)$. Empirical-Bayes estimates of (α_s, a, b) yield closed-form posterior means for the item-level size, which again act as shrinkage estimators between the item sample mean and a panel-level prior mean. We use this variant only in the ablation study (Section 4.6) to probe sensitivity to the choice of size law.

4 Experiments

We design a comprehensive set of experiments to systematically evaluate our proposed TSB-HB model and quantify the benefits of its hierarchical design. Specifically, our evaluation is structured to answer the following key questions:

1. Does TSB-HB outperform established intermittent demand forecasting baselines in terms of predictive accuracy?
2. Do both the hierarchical framework and the *choice of size law* matter? Following the literature that treats size distributions as interchangeable candidates (Teunter et al., 2011), we run Log-Normal (default) and Gamma (variant) and compare accuracy and calibration diagnostics.
3. How does TSB-HB's performance vary across different types of demand intermittency, from smooth to lumpy patterns?

4.1 Experimental Setup

4.1.1 Dataset and Preprocessing

Our empirical evaluation is conducted on the widely-used Online Retail dataset (Chen, 2015), which serves as the basis for all our in-depth analyses, including performance segmentation, ablation studies, and visualizations. We perform standard preprocessing on this dataset: removing returns and zero-price transactions, capping outlier quantities at the 99.5th percentile, aggregating sales to a daily frequency, and filling non-sale days with zero demand. This process results in a final dataset of 3,649 distinct SKUs. To further validate the generalizability of our findings, we also run an additional experiment on the M5 Forecasting Competition dataset (Howard et al., 2020); those results are summarized in Section 4.3.

4.1.2 Evaluation Protocol

We adopt a fixed-origin evaluation protocol (Yang et al., 2021). For each time series, the initial one-third of its historical data is used as the initialization set for model training, and the subsequent two-thirds constitute the out-of-sample evaluation set.

4.1.3 Baseline Models

We compare TSB-HB against a suite of established intermittent demand forecasting models, implemented via the `statsforecast` library (Garza et al., 2022), version `2.0.2`. These baselines include classical heuristics—Croston's method (Croston, 1972), the Syntetos-Boylan Approximation (SBA) (Syntetos and Boylan, 2005), ADIDA (Nikolopoulos et al., 2011), and IMAPA (Petropoulos et al., 2019)—which are parameter-free and evaluated with their standard implementations.

For parameterized models, we ensured optimal or near-optimal configurations. For ARIMA and Theta, we chose AutoARIMA and AutoTheta from `statsforecast`, performing the internal model selection based on information criteria. For TSB, whose performance is sensitive to the smoothing constants (α_d, α_p), we perform a grid search over $\alpha_d, \alpha_p \in \{0.05, 0.10, \ldots, 0.50\}$ and select the pair minimizing validation MAE on the Online Retail panel; this yields a best configuration at $(\alpha_d{=}0.50, \alpha_p{=}0.45)$.

Reproducibility. All code, scripts, and configuration files needed to reproduce our experiments are available at `https://github.com/brianCHUCHU/tsb-hb`.

4.1.4 Evaluation Metrics

We assess forecast performance using both point and probabilistic metrics. For point forecasts, we report Root Mean Squared Scaled Error (RMSSE) (Makridakis et al., 2022), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE)(Wang and Lu, 2018), and Mean Error (ME) (Hyndman and Athanasopoulos, 2018). RMSSE serves as our primary metric given varying scales across series, while MAE reflects overall accuracy, RMSE highlights sensitivity to large errors, and ME provides a diagnostic of systematic bias. Point forecasts correspond to the conditional mean of demand: for heuristic methods such as Croston's variants this is the direct output, whereas for generative models (TSB-HB, AutoARIMA, AutoTheta) the mean is derived from their predictive distributions.

Beyond point accuracy, we evaluate probabilistic forecasts. First, we use the *pinball loss* at quantiles $q \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ (Koenker and Bassett Jr, 1978), a standard proper scoring rule for quantile predictions (Koenker and Bassett Jr, 1978; Koenker and Machado, 1999); lower values indicate sharper and more accurate quantile

Table 1: Overall performance comparison on the Online Retail dataset. Our proposed TSB-HB model wins on two primary metrics (RMSSE and RMSE), establishing its superior accuracy and robustness. The best result in each column is in **bold**.

Model	ME	MAE	RMSE	RMSSE
TSB-HB	-0.7972	5.6575	**17.7773**	**1.1795**
TSB	-1.5751	**5.5736**	18.7272	1.2178
ADIDA	-0.7603	5.6860	17.9617	1.1896
IMAPA	-0.7599	5.7000	17.9963	1.1920
AutoARIMA	-0.9821	6.5750	20.1736	1.2941
AutoTheta	-1.8088	8.2862	22.4711	1.3950
SBA	**-0.0468**	6.1953	18.1633	1.1991
Croston's Method	0.2010	6.3294	18.2320	1.2020

Table 2: Overall performance on 5,000 intermittent series from the M5 dataset. The best result in each column is in **bold**.

Model	ME	MAE	RMSE	RMSSE
TSB-HB	-0.2193	**1.1571**	**2.6783**	**1.1038**
TSB	-0.1009	1.2317	2.9927	1.2274
ADIDA	-0.0369	1.1930	2.9056	1.1736
IMAPA	-0.0329	1.1955	2.9038	1.1805
Croston SBA	0.0278	1.1923	2.8086	1.1128
Croston Classic	0.0952	1.2171	2.8740	1.1393

forecasts. We also report the *Mean Pinball Loss*, averaged across the five quantiles. Second, we assess calibration and sharpness of prediction intervals through empirical coverage and average interval width (AIW) at the 80% nominal level (Gneiting and Raftery, 2007). Well-calibrated models should attain coverage close to the nominal value, while sharper forecasts correspond to narrower AIW for a given coverage level (Gneiting and Raftery, 2007).

4.2 Main Results on the Online Retail Dataset

Table 1 presents the overall performance of all evaluated models. Our proposed TSB-HB model demonstrates superior performance by achieving the best results on two of the three core accuracy metrics. It secures the lowest RMSSE (4.7887), our primary scale-independent metric, and also attains the lowest RMSE (17.7773), indicating its robustness to large errors.

A direct comparison with the TSB baseline, its conceptual predecessor, remains informative. While the TSB model attains the lowest MAE, TSB-HB shows substantial improvements over this base model on all other fronts: it is less biased (ME), more robust to large errors (RMSE), and achieves a better scaled error (RMSSE).

The key takeaway is the model's consistent, top-tier performance. By winning two of the three core accuracy metrics (RMSSE and RMSE) while remaining highly competitive on the third (MAE), TSB-HB establishes itself as arguably the most reliable all-around model in this evaluation.

To validate our model's generalizability, we also evaluate TSB-HB and the baselines on 5,000 intermittent series from the M5 dataset; we report these results in Section 4.3.

4.3 Additional Results on the M5 Dataset

We further validate our model on the M5 Forecasting Competition dataset (Howard et al., 2020) by selecting 5,000 intermittent series and applying the same fixed-origin protocol as in Online Retail. Table 2 summarizes the point-forecast accuracy. TSB-HB achieves the lowest MAE and RMSE among all evaluated methods. In particular, it improves over its conceptual predecessor TSB by 6.1% in MAE (1.1571 vs. 1.2317) and 10.5% in RMSE (2.6783 vs. 2.9927), confirming that the hierarchical Bayesian extension yields consistent gains on a second large-scale benchmark. Detailed per-series diagnostics and additional breakdowns for M5 are omitted for brevity.

Table 3: Pinball loss on Online Retail. Each entry averages the quantile loss across all series and out-of-sample time points; the Mean Pinball Loss is the simple average over the five reported quantiles.

Model	q=0.10	q=0.25	q=0.50	q=0.75	q=0.90	**Mean**
TSB-HB	**0.4755**	**1.1884**	**2.2912**	**2.9753**	**2.6704**	**1.9202**
AutoARIMA	1.7090	2.5244	3.2419	3.8980	2.9148	2.8576
AutoTheta	2.3564	3.6349	4.1256	4.1979	3.0963	3.4822

Table 4: Coverage and average interval width (AIW) at 80% nominal level on Online Retail. TSB-HB achieves the narrowest intervals while maintaining adequate coverage.

Model	**Coverage@80**	**AIW@80**
TSB-HB	**0.900**	**10.184**
AutoARIMA	0.905	28.400
AutoTheta	0.940	42.088

4.4 Probabilistic Results

Table 3 reports the probabilistic forecast evaluation on Online Retail. Across all quantiles ($q = 0.10, 0.25, 0.50, 0.75, 0.90$), TSB-HB achieves the lowest pinball loss, yielding the best Mean Pinball Loss.

Complementary interval metrics in Table 4 further highlight its strengths. At the 80% nominal level, TSB-HB's empirical coverage is 90.0%. While this indicates a degree of over-coverage (Gneiting and Raftery, 2007), producing conservative (i.e., safer) intervals, the key result is that it achieves this coverage with a substantially narrower average interval width (AIW = 10.18). In contrast, AutoARIMA and AutoTheta produce much wider intervals (28.40 and 42.09 respectively) for similar coverage levels. This demonstrates that TSB-HB generates significantly sharper predictive distributions without sacrificing empirical coverage.

Since classical intermittent demand methods (Croston, SBA, TSB, ADIDA, IMAPA) produce only point forecasts and lack distributional outputs (Croston, 1972; Syntetos and Boylan, 2005; Teunter et al., 2011; Nikolopoulos et al., 2011; Petropoulos et al., 2019), we restrict probabilistic comparisons to TSB-HB and the two methods in `statsforecast` that support probabilistic forecasting: AutoARIMA and AutoTheta (Garza et al., 2022).

4.5 Performance Segmentation Analysis

To understand the model's performance across different demand patterns, we segment the time series into four categories based on their Average Demand Interval (ADI) and squared Coefficient of Variation (CV^2). Following the standard framework (Syntetos et al., 2005), we use the thresholds of ADI = 1.32 and CV^2 = 0.49. The resulting classification matrix is defined in Table 5. The results of this segmentation, presented in Table 6, show that TSB-HB achieves the lowest RMSSE in both the Intermittent and Lumpy categories. This outcome aligns with the model's primary design goal, as the TSB-HB framework was developed specifically to handle such demand patterns. Together, these two categories represent over 97% of the series in the dataset. Compared to TSB, TSB-HB shows an RMSSE improvement of 1.1% in the Lumpy category and 0.15% in the Intermittent category.

4.6 Ablation Study

Finally, we conduct an ablation study to isolate the contributions of the model's key components, with results presented in Table 7.

4.6.1 Contribution of the Hierarchical Bayesian Framework

The importance of the HB framework is clearly demonstrated by comparing the full model to its non-hierarchical MLE counterpart. Removing the HB framework (*TSB-MLE-LogNormal*) results in a significant degradation of performance across all metrics, including a 5.9% increase in MAE. This confirms the critical role of hierarchical shrinkage in improving forecast accuracy.

Table 5: Demand categorization based on ADI and CV^2 thresholds.

	$CV^2 < 0.49$	$CV^2 \geq 0.49$
$ADI < 1.32$	Smooth	Erratic
$ADI \geq 1.32$	Intermittent	Lumpy

Table 6: Performance Segmentation Results (RMSSE). TSB-HB demonstrates the best relative performance in the *Lumpy* and *Intermittent* categories, which dominate the dataset.

Model	**Smooth** (9)	**Erratic** (75)	**Intermittent** (1041)	**Lumpy** (2524)
TSB-HB	1.0941	0.8909	**2.5597**	**1.1092**
TSB	1.1150	0.9553	2.5867	1.1477
AutoARIMA	1.8342	0.9846	2.5754	1.2360
AutoTheta	1.4628	1.0644	2.6541	1.3458
ADIDA	1.0706	0.8917	2.5660	1.1211
IMAPA	1.0706	0.8917	2.5671	1.1227
SBA	**1.0472**	0.8818	2.7319	1.1164
Croston's Method	1.0555	**0.8813**	2.7525	1.1178

4.6.2 Choice of Demand Size Distribution

The choice of demand size distribution presents a nuanced trade-off. For the TSB-HB-Gamma variant, we use a standard Gamma-Gamma conjugate model, where the item-level mean size is estimated via Bayesian updating. This also provides shrinkage but through a different mechanism than the REML-based approach used for the Log-Normal model.

As shown in Table 7, the Gamma variant achieves a marginally better result on our primary metric, RMSSE, with a difference of only 0.0006. However, our proposed model with the Log-Normal distribution provides a notable 1.4% improvement in MAE. Given this gain in MAE and the prevalence of the Log-Normal distribution for intermittent data (Teunter et al., 2011), we selected it as our primary model.

This analysis not only validates our model's design choices but also highlights the robustness of the TSB-HB framework itself. The fact that it achieves near state-of-the-art performance under two different distributional assumptions underscores the fundamental importance of the hierarchical Bayesian approach to this problem.

4.7 Comparison with a Deep Learning Baseline

To position TSB-HB against modern deep learning approaches, we compare it with AutoDeepAR from the `neuralforecast` library, a recurrent neural network (RNN) model widely used as a strong benchmark for intermittent demand forecasting (Jeon and Seong, 2022). We follow the same fixed-origin protocol and evaluate both models on a random sample of 100 series from the Online Retail dataset with a 10-day forecast horizon.

Table 8 reports the results. TSB-HB achieves lower MAE and RMSE than AutoDeepAR, indicating more accurate and robust point forecasts, while AutoDeepAR attains a slightly lower RMSSE (an absolute difference of 0.003). Overall, the two models are very close in accuracy, which is notable given that TSB-HB is simpler, fully interpretable, and does not require heavy neural training.

4.8 Visualization of Shrinkage

To illustrate and quantify the mechanism behind the model's gains, Figure 1 visualizes the hierarchical shrinkage effect. We supplement this visualization with quantitative metrics measuring the correlation and variance reduction between the per-item Maximum Likelihood Estimates (MLE) and our model's posterior estimates.

For the demand probability p, where evidence is ample for each item, shrinkage is minimal. This is confirmed quantitatively: the posterior estimates are almost perfectly correlated with the MLEs (Pearson's $r = 0.9987$), and the variance of the estimates is reduced by a modest 11.44%. Conversely, for the conditional demand size, which is observed only on non-zero demand days and is thus much sparser, the shrinkage effect is more pronounced. The correlation between posterior and MLE estimates is lower (Pearson's $r = 0.9194$), and the variance is reduced more substantially, by 18.55%.

Table 7: Ablation study on the Online Retail dataset. Both HB pooling and the Log-Normal assumption contribute to performance gains.

Model Variant	ME	MAE	RMSE	RMSSE
TSB-HB-LogNormal (Full)	-0.7972	**5.6575**	17.7773	1.1795
Ablation: TSB-HB-Gamma	-0.6749	5.7380	**17.7035**	**1.1773**
Ablation: TSB-MLE-LogNormal	-0.1145	5.9931	17.7577	1.1784

Table 8: Comparison with AutoDeepAR on 100 sampled series (10-day horizon) from Online Retail. The best result in each column is in **bold**.

Model	ME	MAE	RMSE	RMSSE
TSB-HB	-0.3725	**5.0648**	**14.7106**	0.9248
AutoDeepAR	0.8306	6.0029	15.0428	**0.9221**

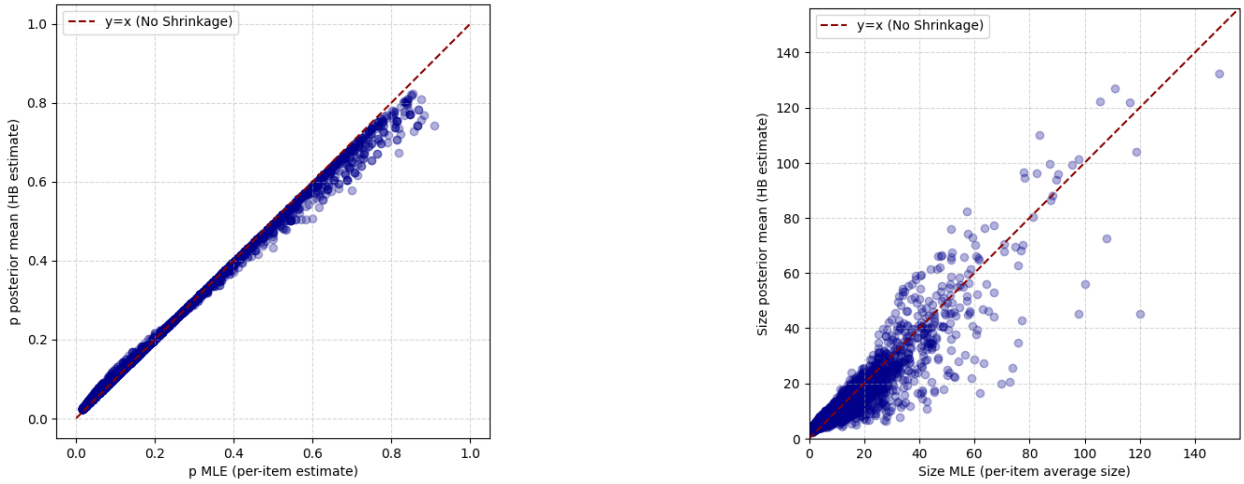
This demonstrates how TSB-HB adaptively applies stronger regularization where the data is less reliable (demand size) while preserving high-quality information where data is rich (demand probability). This mechanism is key to explaining the model's robust performance.

5 Conclusion

This work proposed TSB-HB, a hierarchical Bayesian extension of the TSB paradigm that keeps the operational intuition of an occurrence–size factorization while supplying a principled, fully generative backbone. A Beta–Binomial layer regularizes demand occurrence and a pooled log-normal EB model stabilizes conditional sizes, yielding closed-form updates, scalable computation, and forecasts that remain easy to explain to planners.

Methodologically, TSB-HB (i) generalizes TSB into a coherent probabilistic model, (ii) introduces panel-wise partial pooling that is adaptive to data sparsity and cold starts, and (iii) delivers calibrated distributional forecasts without resorting to heavy inference machinery. Empirical studies indicate that the benefits concentrate where they matter operationally—intermittent and lumpy series—while maintaining the transparency valued in inventory settings.

The framework is intentionally minimalist. It does not yet model time-varying obsolescence or propagate hyperparameter uncertainty beyond plug-in EB. Future work may add state-space dynamics for both occurrence and size, integrate covariates and cross-item dependence, and target cost-aware policies aligned with service levels and capacity constraints.



(a) Shrinkage on Demand Probability (p)

(b) Shrinkage on Demand Size

Figure 1: Visualization of the hierarchical shrinkage effect. Each point corresponds to a single item (SKU), plotting its per-item Maximum Likelihood Estimate (MLE) on the x-axis against the model's hierarchical Bayesian (HB) posterior estimate on the y-axis. (a) For demand probability p, where data is ample, the estimates are highly correlated, indicating weak shrinkage. (b) For conditional demand size, where data is sparser, the shrinkage is more pronounced, pulling unreliable MLEs towards a robust global mean.

References

Vassilis Assimakopoulos and Konstantinos Nikolopoulos. The theta model: a decomposition approach to forecasting. *International journal of forecasting*, 16(4):521–530, 2000.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Daqing Chen. Online Retail. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5BW33.

JD Croston. Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, 23 (3):289–303, 1972.

Azul Garza, Max Mergenthaler Canseco, Cristian Challú, and Kin G. Olivares. StatsForecast: Lightning fast forecasting with statistical and econometric models. PyCon Salt Lake City, Utah, US 2022, 2022. URL `https://github.com/Nixtla/statsforecast`.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL `https://doi.org/10.1198/016214506000001437`.

Addison Howard, inversion, Spyros Makridakis, and vangelis. M5 forecasting - accuracy. `https://kaggle.com/competitions/m5-forecasting-accuracy`, 2020. Kaggle.

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

William James, Charles Stein, et al. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379. University of California Press, 1961.

Yunho Jeon and Sihyeon Seong. Robust recurrent network model for intermittent time-series forecasting. *Applied Soft Computing*, 123:108933, 2022.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker and Jose AF Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448):1296–1310, 1999.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2021.11.013. URL `https://www.sciencedirect.com/science/article/pii/S0169207021001874`. Special Issue: M5 competition.

Konstantinos Nikolopoulos, Aris A Syntetos, John E Boylan, Fotios Petropoulos, and Vassilis Assimakopoulos. An aggregate–disaggregate intermittent demand approach (adida) to forecasting: an empirical proposition and analysis. *Journal of the Operational Research Society*, 62(3):544–554, 2011.

Fotios Petropoulos, Xun Wang, and Stephen M. Disney. The inventory performance of forecasting methods: Evidence from the m3 competition data. *International Journal of Forecasting*, 35(1):251–265, 2019. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2018.01.004. URL `https://www.sciencedirect.com/science/article/pii/S0169207018300232`. Special Section: Supply Chain Forecasting.

James Pitkin, Ioanna Manolopoulou, and Gordon Ross. Bayesian hierarchical modelling of sparse count processes in retail analytics. *The Annals of Applied Statistics*, 18(2):946–965, 2024.

Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 388–394. Springer, 1992.

Aris A Syntetos and John E Boylan. The accuracy of intermittent demand estimates. *International Journal of forecasting*, 21(2):303–314, 2005.

Aris A Syntetos, John E Boylan, and JD Croston. On the categorization of demand patterns. *Journal of the operational research society*, 56(5):495–503, 2005.

Ruud H Teunter, Aris A Syntetos, and M Zied Babai. Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214(3):606–615, 2011.

Weijie Wang and Yanmin Lu. Analysis of the mean absolute error (mae) and the root mean square error (rmse) in assessing rounding model. *IOP Conference Series: Materials Science and Engineering*, 324(1):012049, 2018. doi: 10.1088/1757-899X/324/1/012049.

Youpeng Yang, Chenyi Ding, Sanghyuk Lee, Limin Yu, and Fei Ma. A modified teunter-syntetos-babai method for intermittent demand forecasting. *Journal of Management Science and Engineering*, 6(1):53–63, 2021.