



DEMS WORKING PAPER SERIES

The cost of ensembling: is it always worth combining?

Marco Zanotti

No. 554 – June 2025

**Department of Economics, Management and Statistics
University of Milano – Bicocca
Piazza Ateneo Nuovo 1 – 2016 Milan, Italy
<http://dems.unimib.it/>**

The cost of ensembling: is it always worth combining?

Marco Zanotti^{a,*}

^a*Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy*

Abstract

Given the continuous increase in dataset sizes and the complexity of forecasting models, the trade-off between forecast accuracy and computational cost is emerging as an extremely relevant topic, especially in the context of ensemble learning for time series forecasting. To assess it, we evaluated ten base models and eight ensemble configurations across two large-scale retail datasets (M5 and VN1), considering both point and probabilistic accuracy under varying retraining frequencies. We showed that ensembles consistently improve forecasting performance, particularly in probabilistic settings. However, these gains come at a substantial computational cost, especially for larger, accuracy-driven ensembles. We found that reducing retraining frequency significantly lowers costs, with minimal impact on accuracy, particularly for point forecasts. Moreover, efficiency-driven ensembles offer a strong balance, achieving competitive accuracy with considerably lower costs compared to accuracy-optimized combinations. Most importantly, small ensembles of two or three models are often sufficient to achieve near-optimal results. These findings provide practical guidelines for deploying scalable and cost-efficient forecasting systems, supporting the broader goals of sustainable AI in forecasting. Overall, this work shows that careful ensemble design and retraining strategy selection can yield accurate, robust, and cost-effective forecasts suitable for real-world applications.

Keywords: Time series, Demand forecasting, Forecasting competitions, Cross-learning, Global models, Forecast combinations, Ensemble learning, Machine learning, Deep learning, Conformal predictions, Green AI

JEL: C53, C52, C55

*Corresponding author

Email address: zanottimarco17@gmail.com (Marco Zanotti)

1. Introduction

Ensemble learning has become a powerful and widely adopted approach in time series forecasting due to its ability to improve predictive accuracy and robustness by aggregating the outputs of multiple models. The fundamental idea behind ensemble methods is that combining forecasts from diverse models can help balance their individual strengths and weaknesses, thereby reducing the risk of overfitting and mitigating model-specific biases or errors. This approach leverages the principle that while individual models might make different types of errors, their combination can average out these errors. One of the most intriguing and empirically validated findings in the literature is that simple ensemble strategies, such as taking the mean or median of forecasts, often outperform more sophisticated combination methods like stacking (also known as meta-learning) or weighted ensembles. This phenomenon, known as the forecast combination puzzle (Claeskens et al., 2016), suggests that in many cases the added complexity of estimating optimal weights or training a meta-learner does not yield meaningful gains in accuracy and can even lead to overfitting, especially when the number of forecast points is limited. Indeed, the success of simple combinations is partly due to their ability to remain robust in the presence of model misspecification, parameter uncertainty, and changing data dynamics, issues that more complex ensemble methods may struggle to accommodate effectively.

In general, the benefits of ensemble learning are numerous: it improves forecast accuracy, enhances model stability across time and datasets, and increases resilience to structural breaks or shifts in the data. Ensembles also provide a practical hedge against model uncertainty, allowing practitioners to rely less on the assumption that any single model is correctly specified. In the context of global forecasting models, ensemble methods become especially valuable. Global models are trained across many time series simultaneously, and while they excel at identifying cross-series patterns, they can also suffer from model instability or biases if the chosen architecture fails to generalize across heterogeneous series. Combining multiple global models, each with different biases, architectures, or feature representations, can help counteract these issues by smoothing out individual model errors and reducing the risk of overfitting to shared but spurious signals across series (Wang et al., 2023). Additionally, ensembles serve as a hedge against model selection uncertainty, which is particularly relevant in global settings where the optimal model may vary substantially across groups of series. By pooling predictions, ensembles offer a form of model

diversification that improves generalization, making them a powerful and practical tool in modern forecasting systems.

However, ensemble methods are not without drawbacks. One significant limitation is the increased computational cost, especially when combining large numbers of complex models like deep learning architectures or global forecasting models. This cost is magnified when retraining or model tuning is required for each component model at multiple forecast origins. Additionally, from a business perspective, managing and maintaining multiple models adds complexity to the production of forecasts. Moreover, in cases where the component models are very similar or highly correlated, the marginal benefit of ensembling diminishes, potentially offering little improvement over the best individual model (Kolassa, 2011).

Therefore, one of the major advantages of adopting a global modeling approach can be offset when ensemble learning is introduced. Combining several models, even with simple techniques like averaging, inherently increases the computational time, as multiple models must be trained, validated, and maintained, producing substantial computational overhead. In this sense, ensemble learning may weaken one of the core benefits of the global modeling paradigm: reduced training time and resource usage. Moreover, since forecasts are typically generated using cloud computing services with pay-as-you-go pricing models (Fotios Petropoulos & Spiliotis, 2024), increased computational time and resource consumption directly leads to higher forecasting costs for organizations. These costs may also rise exponentially when frequent retraining strategies are adopted. Indeed, it is a widespread practice to update forecasting models whenever new data becomes available, often driven by the belief that frequent updates enhance the model’s ability to adapt to evolving patterns and improve predictive accuracy. However, when ensembles are used, all the models within the ensembles must be retrained to obtain the final predictions, abruptly increasing the costs of forecasting.

This trade-off between performance and computational time in the context of ensemble learning raises important questions about the balance between accuracy gains and cost efficiency, especially in large-scale forecasting applications. However, despite these challenges, ensemble learning remains one of the most widely used strategies in the forecasting industry, which further amplifies concerns about the long-term sustainability and scalability of forecasting systems. Indeed, from a practical perspective, forecast combinations, coupled with frequent retraining, has also significant environmental costs, since the energy consumption of model training extends beyond direct computational

costs, contributing to energy consumption and carbon emissions (Schwartz et al., 2020). Therefore, understanding the cost of ensemble forecasting models and the effects of retraining on their performance is of paramount importance for advancing more sustainable forecasting practices.

1.1. Research Question

We aim to address the question "*Are ensembles always the solution?*". Specifically, we focus on the trade-off between accuracy and sustainability, in terms of the cost of producing the forecasts, among different types of ensembles of global forecasting models. This cost is particularly relevant for ensemble models since it becomes exponentially larger with the number of base models used to produce the combined forecasts. To this end, we use ten global models as base learners, ranging from traditional machine learning methods to advanced deep learning architectures. In addition, we evaluate eight ensemble approaches, all trained and tested on the two most recent and comprehensive retail forecasting datasets: the M5 and VN1 competition datasets.

We also investigated the effects of retraining on the forecasting performance of ensembles, compared to that of the base models. In particular, we assess whether reducing the frequency of retraining, by avoiding re-estimation of base models as new data becomes available, can effectively enhance the cost-accuracy trade-off in ensemble forecasting. Hence, we explored a wide range of retraining strategies, from continuous retraining to a no-retraining approach, including various intermediate periodic retraining schemes to encompass the most practical and effective scenarios.

1.2. Contributions

Our contribution is fourfold:

- We provide the first comprehensive study of the cost-accuracy trade-off of ensembles of global models, using 10 distinct methods, some of the most relevant real-world datasets, and evaluating both point and probabilistic predictions.
- We analyze different ensemble strategies to assess the effects of the ensemble size on forecast accuracy and costs.
- We compare various retraining solutions, such as continuous, periodic, and no retraining, across multiple datasets, to quantify the impact of the retraining frequency on the computational cost of ensemble forecasting.

- We provide practical guidelines for organizations and practitioners on assessing when and how frequently to retrain ensemble forecasting models to obtain an effective balance between predictive accuracy and computational cost.

By tackling these issues, this paper contributes to both forecasting and machine learning communities, offering valuable insights into the trade-offs among accuracy, computational efficiency, and sustainability in the use of an ensemble of global forecasting models.

1.3. Overview

The rest of this paper is organized as follows. After a brief review of related works (Section 2), in Section 3 we describe the design of the experiment used in our study. The datasets and their characteristics are presented in 3.1, while the global forecasting models and the ensemble strategies are shown in 3.2. The concepts related to rolling origin evaluation and retrain scenario are explained in 3.3, while the metrics used to assess the accuracy and cost of models are discussed in 3.4. In Section 4 we show the empirical findings of our study, including forecast accuracy, computing time, and cost analysis of the different scenarios and across the different ensembles. Finally, Section 5 contains our summary and conclusions.

2. Related works

The cross-learning approach in time series forecasting has witnessed substantial advancements in recent years, becoming a central theme in contemporary research. Today, global models are commonly used as benchmarks in empirical studies, underscoring their growing relevance in the field Semenov et al. (2021). Additionally, theoretical contributions from Montero-Manso & Hyndman (2021) and Montero-Manso (2023) have shown that global models can achieve accuracy comparable to local models with reduced complexity and without relying on assumptions about data similarity. Indeed, global models have proven particularly effective across a variety of forecasting domains, especially in retail forecasting domain (Makridakis et al., 2022a), and several methods have been proposed to further boost their performance (Godaheva et al., 2021a; Bandara et al., 2020, 2021). In particular, ensemble learning is a powerful tool in the forecasters’ hands offering a reliable strategy to enhance accuracy and robustness by combining the strengths of multiple models, not only in the cross-learning context (Wang et al., 2023). Relative to point forecasting, many different

combination methods have been proposed: linear combinations with optimal weights (Newbold & Granger, 1974), performance based weighting (Pawlikowski & Chorowska, 2020), criteria-based weighting (Kolassa, 2011), and many different approaches based on stacking, or meta-learning (Wang et al., 2023). However, time series forecasting competitions have shown that simple ensemble strategies, like simple mean or median, are extremely accurate, yet efficient (Makridakis et al., 2022b). Indeed, simple combination schemes are hard to beat, and the simple arithmetic average of predictions with equal weights remains the most widely used and surprisingly effective combination rule (Claeskens et al., 2016).

Most evaluations of global models emphasize point forecast accuracy, likely due to the fact that many machine learning and deep learning algorithms do not natively produce probabilistic outputs (Makridakis et al., 2022c). However, in applications such as supply chain management, quantifying uncertainty is essential, whether through prediction intervals, quantiles, or full predictive distributions (Fildes et al., 2022). In this regard, the Conformal Inference framework introduced by Vovk et al. (2005) offers a flexible tool for uncertainty quantification, and has been successfully applied to time series forecasting problems (Stankeviciute et al., 2021). Forecast combinations in the context of probabilistic predictions are an active area of research. Different methods have been proposed, from linear pooling, Bayesian model averages, and integral combinations (Wang et al., 2023). However, quantile aggregation via simple average has proven to be effective and accurate (Smyl & Hua, 2019; Buseti, 2017).

From the perspective of model retraining, the most comprehensive investigation in the context of global models has been conducted by Zanotti (2025), who found that reducing the retraining frequency of global models can lower forecasting costs without harming accuracy. Other works, such as Spiliotis & Petropoulos (2024), explored retraining strategies and parameter updates for local exponential smoothing models, while Huber & Stuckenschmidt (2020) examined retraining in retail demand forecasting using a limited range of models and datasets. Although these studies offer useful insights, direct analysis on the effects of retraining ensemble models remains unexplored.

The question of whether ensemble methods are consistently worthwhile is relevant both for the forecasting and the broader machine learning community. Indeed, the growing advocacy for sustainable AI (Schwartz et al., 2020; Getzner et al., 2023) highlights the importance of evaluating the environmental and computational costs of model development and maintenance. Our study

directly addresses this gap by systematically evaluating different ensemble strategies’ cost-accuracy trade-off. We also assess how different updating frequencies affect the forecasting accuracy across the ensemble models, with the aim of testing whether less frequent retraining can be an effective tool to manage this trade-off. Answering these questions, this research provides both theoretical clarity and practical guidelines towards more sustainable forecasting practices.

3. Experimental design

This section presents the empirical analysis performed to assess the performance of ensemble models and to determine whether reducing the frequency of retraining can yield forecasting accuracy comparable to that of a baseline scenario, involving continuous retraining, while systematically decreasing the forecasting costs. We begin by introducing the datasets used in the experiments, followed by a description of the ensemble learning models employed. Finally, we outline the evaluation strategy, including the performance metrics, the different retraining scenarios considered, and the approach used to assess forecast performance and costs.

3.1. Datasets

For our experiments, we utilized two prominent retail forecasting datasets: the M5 and VN1 competition datasets. The M5 competition, part of the M-competitions series led by Spyros Makridakis and collaborators, was designed to benchmark forecasting methods in a retail demand setting (Makridakis et al., 2022b). The M5 dataset (Howard & Makridakis, 2020) is widely recognized and extensively studied, comprising 3,049 daily time series representing unit sales of Walmart products. These sales span three main product categories, Food, Hobbies, and Household, across ten stores in three U.S. states: California, Texas, and Wisconsin. Covering the period from 2011 to 2016, the dataset features highly intermittent time series organized hierarchically, enabling multi-level forecasting (e.g., individual SKUs, categories, stores, and states). It also includes relevant exogenous variables such as prices, promotions, and special events like holidays. The VN1 Forecasting – Accuracy Challenge, launched in October 2024 by Flieber, Syrup Tech, and SupChains, marked the first edition of its kind (Vandeput, 2024). This dataset consists of weekly sales data for 15,053 products sold from 2020 to 2024 by various e-vendors, primarily based in the U.S. Unlike the M5 dataset, which contains sales from a single retailer (Walmart) and a limited number of physical stores, the VN1 dataset aggregates sales across 328 warehouses operated by 46

distinct retailers. To the best of our knowledge, we are among the first to benchmark forecasting models on this dataset. Together, these two datasets represent the most recent and comprehensive publicly available collections of time series related to retail demand, providing a strong foundation for generalizing our findings in the domain of demand forecasting.

Table 1: The M5 and the VN1 datasets used in the experiments.

Dataset	Frequency	N. Series	Min Obs per Series
M5	Daily (7)	28.298	730
VN1	Weekly (52)	15.053	157

In both datasets we concentrate on the most granular level (individual SKUs), because the potential gains from reduced retraining are greatest at lower aggregation levels. To ensure a consistent evaluation strategy (Section 3.3), we filtered the series: daily SKUs retained at least two years of data (more than 730 observations), while weekly SKUs required a minimum of three years (more than 157 observations).

3.2. Forecasting models

In this study, we focused exclusively on global forecasting methods, as we used only this category of models as base learners for our ensembles. Global approaches have become standard in many industries dealing with large-scale time series data, such as retail demand forecasting, where predictions must be made for thousands of SKUs (Januschowski et al., 2020). Indeed, our primary objective is to assess the performance of ensemble models obtained by combining the predictions of different global models trained on large datasets.

A global forecasting model can be defined as:

$$Y_i^h = F(\mathcal{Y}, \Theta), \quad (1)$$

where forecasts for a horizon h for each individual time series Y_i are generated using a single model F trained on the entire set of time series in a dataset \mathcal{Y} . Hence, using the global modeling paradigm, it is possible to leverage cross-learning to allow the model to learn shared patterns across all time series. Indeed, the model parameters Θ are shared across all series.

Table 2: Global forecasting models used in the experiment.

Machine Learning	Deep Learning
Linear Regression (LR) (Godahewa et al., 2021b)	MLP (Rosenblatt, 1958)
Random Forest (RF) (Breiman, 2001)	LSTM (Hochreiter & Schmidhuber, 1997)
XGBoost (Chen & Guestrin, 2016)	TCN (Van den Oord et al., 2016)
LGBM (Ke et al., 2017)	NBEATS (Oreshkin et al., 2020)
CatBoost (Prokhorenkova et al., 2018)	NHITS (Challu et al., 2022)

To ensure a comprehensive evaluation of the ensemble models, we included both traditional machine learning models and state-of-the-art deep learning techniques as base learners, chosen for their proven effectiveness in time series forecasting and their methodological diversity. In particular, Table 2 shows the ten different global forecasting models we trained in our experiments.

Machine learning models are generally easier to train than deep learning approaches but often rely heavily on extensive and careful feature engineering to achieve strong forecasting performance (Januschowski et al., 2022). In contrast, deep learning models can automatically learn relevant features, such as lags or rolling statistics, within their architecture, reducing the need for manual preprocessing. However, they are typically more challenging to train due to their larger number of hyperparameters, which can significantly influence forecast accuracy (Smyl, 2020). As in Zanotti (2025), we implemented simplified feature engineering pipelines inspired by the top-performing solutions in the M5 and VN1 competitions. Our feature set included standard time series features such as lags, rolling and expanding means, as well as calendar-related variables like year, month, week, and day of the week. We also integrated static metadata, such as store, product, category, and location identifiers, based on the dataset’s characteristics. For the M5 dataset, we further included external variables like special events. Hyperparameters were chosen based on configurations from leading competition entries when available; otherwise, we used the default values recommended by the respective libraries.

3.2.1. Ensemble learning

Ensemble learning aggregates the predictions of multiple base models to enhance forecast accuracy and reliability, especially when different models capture diverse and complementary

patterns in the data. This approach is particularly beneficial in global forecasting, where the use of a single model can lead to instability or overfitting due to the heterogeneous nature of the time series being modeled (Wang et al., 2023). To address these challenges, forecast combinations help by balancing out the individual biases and variances of the base models. In a general formulation, an ensemble forecasting model combines the outputs of multiple models. Using the same notation as before, where \mathcal{Y} is the set of all time series and $\Theta^{(j)}$ denotes the parameters of the j -th base model $F^{(j)}$, the ensemble prediction for series Y_i at horizon h can be defined as:

$$Y_i^h = G\left(F^{(1)}(\mathcal{Y}, \Theta^{(1)}), F^{(2)}(\mathcal{Y}, \Theta^{(2)}), \dots, F^{(J)}(\mathcal{Y}, \Theta^{(J)})\right). \quad (2)$$

Here, G is the ensemble function (e.g., mean, median, weighted average) that combines the forecasts of J base models.

In our study, we adopted an ensemble approach that combines the models' predictions using a simple average. The simple average is a widely used method in forecasting to create forecast combinations due to its simplicity, proven effectiveness in improving prediction robustness, and is often more accurate than theoretically optimal combinations (Claeskens et al., 2016). Therefore, being G the simple mean, the ensemble forecast for series Y_i at horizon h becomes:

$$Y_i^h = \frac{1}{J} \sum_{j=1}^J F^{(j)}(\mathcal{Y}, \Theta^{(j)}). \quad (3)$$

Furthermore, since we are not only interested in point forecast accuracy, in the context of probabilistic forecasting, ensemble methods can also be applied to combine predictive distributions. A common and straightforward approach is to average the predicted quantiles across base models (Wang et al., 2023). That is, for each desired quantile level (e.g., 0.1, 0.5, 0.9), the ensemble forecast simply takes the mean of the corresponding quantiles predicted by the individual models. Despite its simplicity, this method is often effective and overall seems to be preferred compared to other combination techniques (Smyl & Hua, 2019). Moreover, equally weighting quantiles through a simple average yields robust and improved forecast skill because the error in estimating optimal weights usually exacerbates the ensemble predictions (Busetti, 2017). Lastly, simple combination methods have the lowest computational time possible ¹, allowing us to effectively study, without

¹Simple combinations, such as the mean or median, require only basic arithmetic operations on the forecast outputs of the base models, avoiding the need for model training or optimization, as is required in more complex approaches like meta-learning.

loss of generality, the associated costs of ensembling.

In our study, we employed simple mean ensembles to combine the forecasts of multiple global models, following two distinct selection criteria. The first strategy, which we refer to as *ENSACC*, combines the base models that demonstrated the highest individual point forecast accuracy (measured by Equation 4). This approach aims to leverage the strengths of top-performing models, under the assumption that their combined output will retain strong predictive performance while potentially offsetting individual weaknesses. The second strategy, *ENSTIME*, focuses instead on computational efficiency by combining the models with the lowest training and inference times. This ensemble reflects a pragmatic choice for real-world forecasting systems where computational cost is a critical constraint, such as in large-scale retail applications. For both strategies, we constructed ensembles of increasing size, combining the top 2, 3, 4, and 5 models according to each criterion. Table 3 shows the composition of the eight different ensembles for each dataset. This tiered design allows us to assess how forecast accuracy and computational efficiency evolve as additional models are added to the ensemble. Limiting the maximum ensemble size to five models (out of eight or ten, depending on the dataset) reflects a balance between potential accuracy gains and the diminishing returns or increased complexity often observed with larger ensembles (Wang et al., 2023). This systematic approach enables a thorough evaluation of the trade-off between accuracy and computational cost in ensemble-based global forecasting.

All models were implemented in Python using Nixtla’s framework (Nixtla, 2022). Specifically, the *mlforecast* library was used to train the machine learning models, while *neuralforecast* was employed for efficiently training the deep learning models.

3.3. Evaluation strategy

In this section, we introduce the concepts of retraining scenarios and rolling window forecasting that we used in our experiment.

3.3.1. Retrain scenario

Following Zanotti (2025), we evaluated multiple retraining scenarios, or retrain windows. A retrain scenario r denotes a positive integer indicating how often the model is updated, or retrained. Specifically, r defines the number of new observations that must be collected before retraining occurs. These scenarios are tailored to the frequency of each dataset, which in turn determines the

Table 3: Composition of ENSACC and ENSTIME ensembles for the M5 and VN1 datasets.

Dataset	Ensemble	Models Included
M5	ENSACC (Accuracy-driven combinations)	
	ENS2A	XGBoost, LGBM
	ENS3A	XGBoost, LGBM, LR
	ENS4A	XGBoost, LGBM, LR, NBEATSx
	ENS5A	XGBoost, LGBM, LR, NBEATSx, MLP
	ENSTIME (Time-efficient combinations)	
	ENS2T	CatBoost, LR
	ENS3T	CatBoost, LR, XGBoost
	ENS4T	CatBoost, LR, XGBoost, MLP
	ENS5T	CatBoost, LR, XGBoost, MLP, NBEATSx
VN1	ENSACC (Accuracy-driven combinations)	
	ENS2A	MLP, NBEATSx
	ENS3A	MLP, NBEATSx, NHITS
	ENS4A	MLP, NBEATSx, NHITS, RF
	ENS5A	MLP, NBEATSx, NHITS, RF, XGBoost
	ENSTIME (Time-efficient combinations)	
	ENS2T	LR, XGBoost
	ENS3T	LR, XGBoost, CatBoost
	ENS4T	LR, XGBoost, CatBoost, MLP
	ENS5T	LR, XGBoost, CatBoost, MLP, TCN

forecast horizon and business review cycles. Table 4 summarizes the selected retraining windows, test periods, and forecast horizons.

Table 4: The retraining scenarios, the test window, and the horizon for the M5 and the VN1 datasets.

Dataset	Frequency	Retraining Scenarios (r)	Test Window (T)	Horizon (h)
M5	Daily (7)	7, 14, 21, 30, 60, 90, 120, 150, 180, 364	364	28
VN1	Weekly (52)	1, 2, 3, 4, 6, 8, 10, 13, 26, 52	52	13

For example, $r = 7$ for daily data implies weekly retraining. Each list contains ten scenarios to ensure a broad yet computationally feasible exploration. The scenario with $r = 1$, known as continuous retraining, is the most computationally intensive and typically the most accurate, as the model always uses the most recent data. Hence, we adopt it as our benchmark for both accuracy and cost. However, for daily data, we consider $r = 7$ (weekly retraining) as the practical benchmark, since daily updates are uncommon in real-world settings. Conversely, the no-retraining scenario $r = T$, where T is the length of the test set, fits the model only once and uses it for the entire forecasting horizon, minimizing computational load but likely yielding the lowest accuracy. All intermediate values $1 < r < T$ represent periodic retraining strategies, where both forecast accuracy and computational cost are expected to decrease as r increases.

Since we use global models, the training set at each retraining step includes all time series in the dataset, each extended by r new observations. We evaluate two update strategies: (i) full retraining at each window r , or (ii) no update between retrain points, i.e., the model remains unchanged for the next r periods.

3.3.2. Rolling window evaluation

Out-of-sample evaluation is essential in time series forecasting to assess a model’s ability to generalize to unseen data, especially given the potential for structural changes or unanticipated shifts in future values (Tashman, 2000). Among evaluation strategies, rolling origin evaluation is widely recognized as the most appropriate approach (Bergmeir & Benítez, 2012). This method systematically assesses forecast accuracy over multiple iterations by simulating repeated forecasting cycles. The procedure starts by splitting the series into a training and a test set, maintaining the chronological order. At each step, the model is trained on the current training set and used to

predict the next h observations. The forecast origin is then shifted forward by a fixed step, and the process is repeated, either with a growing (expanding) or fixed-size training window. Performance metrics (see Section 3.4) are averaged over all iterations to provide a robust estimate of forecasting accuracy.

Compared to fixed origin evaluation, which offers only a single evaluation point, rolling origin evaluation provides a more robust assessment by capturing performance across varying conditions such as seasonal shifts, level changes, or trend evolutions (Bergmeir & Benítez, 2012). This is particularly beneficial in dynamic settings like retail or supply chain management, where models must adapt to frequent changes. The setup is flexible: the training window can either expand to include all available past data or remain fixed to a specified length. Most practitioners favor the expanding window, especially when time series are short, as it leverages the full data history (Petropoulos & et al., 2022). In our study, we adopt the expanding window strategy to better reflect realistic business forecasting practices and to accommodate short series, such as those in the weekly VN1 dataset. Following Zanotti (2025), Table 4 outlines the key parameters used in our experiments. Lastly, we set the step size to one in all cases to maximize the number of evaluation points across each retraining scenario.

3.4. Evaluation metrics

Evaluating the accuracy of point forecasts in time series analysis remains a debated topic. Although numerous metrics exist to assess model performance, there is no clear consensus in the literature on which metric is superior (Hewamalage et al., 2023). Given that our focus is on SKU-level demand forecasting, where data is typically intermittent, error metrics based on absolute or percentage deviations are suboptimal, as they emphasize the median rather than the full distribution (Kolassa, 2020). Furthermore, due to scale variations across series, employing a scale-independent metric is essential. To address these concerns, following Zanotti (2025), we adopted the Root Mean Squared Scaled Error (RMSSE), introduced by Hyndman & Koehler (2006), as our primary point forecast accuracy measure. RMSSE is calculated as:

$$\text{RMSSE} = \sqrt{\frac{\frac{1}{h} \sum_{t=n+1}^{n+h} (y_t - \hat{y}_t)^2}{\frac{1}{n-s} \sum_{t=s+1}^n (y_t - y_{t-s})^2}}. \quad (4)$$

This metric compares the mean squared error of a forecast to that of a seasonal naïve benchmark, providing a relative performance measure. It was the official evaluation metric used in the M5

competition (with $s = 1$) (Makridakis et al., 2022b), one of the datasets analyzed in our study. Lower RMSSE values indicate more accurate forecasts.

Beyond point accuracy, we also assessed the probabilistic quality of forecasts across different retraining scenarios. Since most of the machine learning and deep learning models we use do not natively produce probabilistic outputs, we employed Conformal Inference to generate prediction intervals. Conformal Inference is a versatile framework that quantifies predictive uncertainty using a calibration (validation) set, without relying on strong distributional assumptions (Vovk et al., 2005). Originally developed for i.i.d. data, it has recently been adapted to time series contexts (Stankeviciute et al., 2021). Its properties, distribution-free, model-agnostic, efficient, and suitable for small datasets, make it especially well-suited for forecasting benchmarks involving diverse models and datasets. As in Zanotti (2025), to evaluate the quality of the prediction intervals, we used the Quantile Loss (QL), also known as Pinball Loss, and its average form, the Multi-Quantile Loss (MQL):

$$\text{QL} = \frac{1}{h} \sum_{t=n+1}^{n+h} (q \cdot (y_t - \hat{y}_t) \cdot \mathbb{I}_{y_t \geq \hat{y}_t} + (1 - q) \cdot (\hat{y}_t - y_t) \cdot \mathbb{I}_{y_t < \hat{y}_t}), \quad (5)$$

$$\text{MQL} = \frac{1}{Q} \sum_{q \in Q} \text{QL}(q). \quad (6)$$

As a proper scoring rule, QL enables a rigorous evaluation of probabilistic forecasts (Kolassa, 2016). MQL, particularly in its weighted form, was the principal metric in the M5 Uncertainty competition (Makridakis et al., 2022c).

For our analysis, we constructed prediction intervals around the median and six central levels: 60%, 70%, 80%, 90%, 95%, and 99%, yielding a total of 13 quantiles. The lower quantiles (e.g., median, 60%, 70%) describe the forecast center, while the upper quantiles (90% and above) are crucial for assessing tail risks, key to managing safety stock in retail settings (Barrow & Kourentzes, 2016). These quantiles offer a well-rounded view of forecast uncertainty. To ensure valid estimation of predictive uncertainty, conformal prediction intervals were computed on validation sets at least twice the length of the forecast horizon. This requirement limited the number of time series we could retain in the dataset, as explained in previous sections.

In addition to accuracy, we explicitly evaluated the computational cost of generating forecasts by measuring Computing Time (CT), defined as the number of seconds needed to train a model and

produce h step-ahead forecasts (Zanotti, 2025). CT directly reflects forecasting costs in real-world deployments, especially when using cloud-based infrastructure with pay-as-you-go pricing (Spiliotis & Petropoulos, 2024). Thus, CT was recorded for each model across all retraining scenarios, with lower CT values indicating more efficient forecasting.

To facilitate consistent comparisons, results for all evaluation metrics were normalized relative to the baseline (most frequent) retraining scenario for each dataset frequency. To statistically assess performance differences across scenarios, we applied the Friedman-Nemenyi test for multiple comparisons (Demšar, 2006).

All experiments were conducted using a Microsoft Azure NC6s_v3 cloud instance running Ubuntu 24, equipped with 1 GPU, 6 CPU cores, and 112 GB of memory.

4. Results and discussion

This section presents and interprets the empirical findings of our study, highlighting the interplay between accuracy, probabilistic performance, computing time, and total cost across different retraining strategies and ensemble configurations. We draw insights from both the M5 and VN1 datasets to identify consistent patterns and dataset-specific nuances.

Table 5 summarizes the performance and computational time of all forecasting methods evaluated in this study, across both the M5 and VN1 datasets. The models are grouped into four categories: Machine Learning (ML), Deep Learning (DL), ENSACC (ensembles optimized for accuracy), and ENSTIME (ensembles optimized for computational efficiency). For each method, the table reports three key metrics per dataset: RMSSE (Root Mean Squared Scaled Error) to assess point forecast accuracy, MQL (Multi-Quantile Loss) to evaluate probabilistic forecast performance, and CT (Computing Time, in seconds) to reflect the computational time under a cloud-based setting. Overall ², the models evaluated achieved better absolute performance on the M5 dataset than on VN1. Several factors may explain this difference, including the larger dataset size, the higher frequency of the time series in M5, the availability of rich external regressors (e.g., promotions, special events), and the presence of well-established benchmark hyperparameter settings, many of which were not available for VN1 during model training.

²The overall results are derived from the baseline scenario, that is $r = 7$ for M5 and $r = 1$ for VN1, as this setting is regarded as standard in both theoretical and practical applications.

Table 5: Overall forecasting performance and computational cost for each method across datasets. RMSSE, MQL, and CT (in seconds). Minimum values per column are highlighted in bold.

Type	Method	M5			VN1		
		RMSSE	MQL	CT	RMSSE	MQL	CT
ML	LR	0.777	0.267	11373	6.549	2.896	236
	RF	–	–	–	1.868	2.590	24862
	XGBoost	0.755	0.258	15417	1.890	2.469	530
	LGBM	0.771	0.256	44429	3.542	2.625	7824
	CatBoost	0.947	0.263	10424	5.762	2.845	805
DL	MLP	0.821	0.281	17584	1.543	2.492	962
	LSTM	–	–	–	1.913	2.843	1284
	TCN	0.865	0.290	33364	1.913	2.843	1127
	NBEATSx	0.815	0.279	21226	1.698	2.626	1244
	NHITS	0.828	0.284	21969	1.699	2.632	1251
ENSACC	Ens2A	0.757	0.255	59846	1.472	2.369	2205
	Ens3A	0.757	0.256	71219	1.517	2.410	3456
	Ens4A	0.758	0.249	92445	1.524	2.386	28318
	Ens5A	0.763	0.251	110029	1.544	2.375	28849
ENSTIME	Ens2T	0.814	0.230	21797	5.719	2.545	766
	Ens3T	0.776	0.223	37214	5.536	2.588	1572
	Ens4T	0.764	0.218	54797	4.472	2.419	2533
	Ens5T	0.763	0.220	76024	3.672	2.339	3660

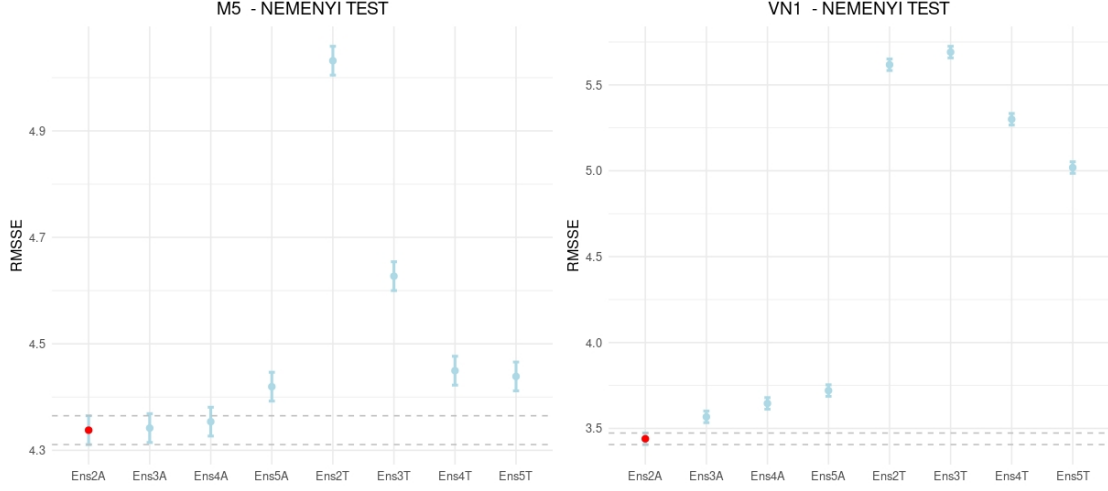


Figure 1: Friedman-Nemenyi test results of ensembles based on RMSSE.

Analyzing the point forecasting accuracy, the RMSSE confirms the effectiveness of ensemble methods in both datasets. In general, ENSACC ensembles (those optimized for accuracy) consistently outperform individual base models. The accuracy improvements, however, show diminishing returns beyond three or four models, aligning with the well-documented phenomenon in ensemble literature where the marginal gain from adding more models decreases. Notably, while ENSACC ensembles dominate in accuracy, ENSTIME ensembles (optimized for computational efficiency) still provide competitive results in the M5 dataset, demonstrating that even time-efficient combinations can retain good predictive power. Interestingly, ENSTIME ensembles show increasing returns in accuracy as more models are added: their performance consistently improves with each additional component. Indeed, in contrast to ENSACC ensembles, where combining just two models is often sufficient to achieve optimal performance (as illustrated in Figure 1), ENSTIME combinations benefit from larger ensemble sizes, with forecast accuracy continuing to improve as more low-cost models are included.

A similar pattern emerges for probabilistic accuracy evaluated using Multi-Quantile Loss (MQL). Compared to base models, ensembles achieve superior uncertainty quantification, benefiting from the diversity in quantile predictions among different models. Simple averaging across quantiles proves to be a highly effective strategy. Also, in this case, ENSACC forecast combinations exhibit diminishing returns in accuracy as more models are added. On the contrary, increasing the number of models

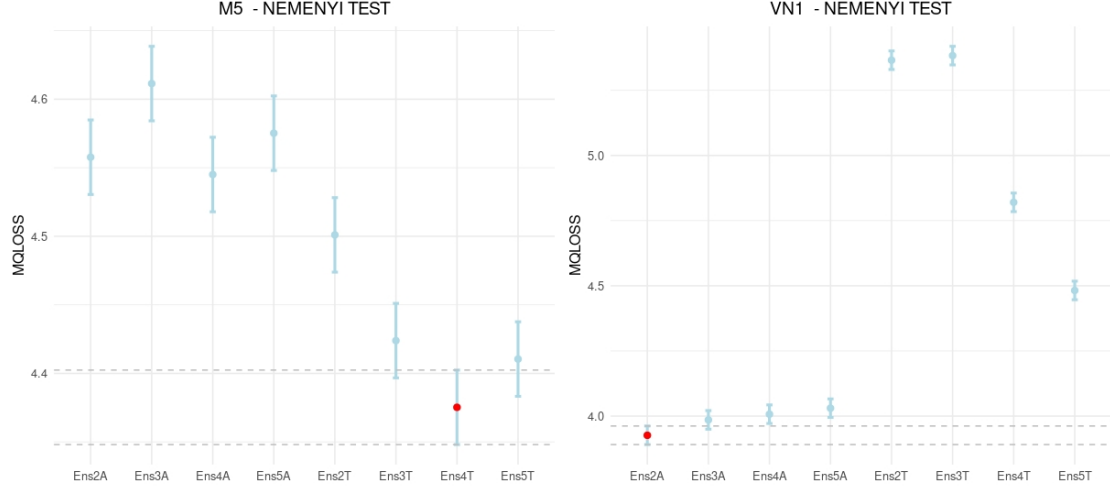


Figure 2: Friedman-Nemenyi test results of ensembles based on MQL.

in the ensemble based on computational efficiency leads to consistently improved probabilistic forecasting performance, which may even outperform the ENSACC combinations. As shown in Figure 2, ENSTIME ensembles achieve significantly better probabilistic forecasting performance on the M5 dataset compared to all accuracy-based combinations. In contrast, for ENSACC ensembles, combining more than two base models does not yield additional improvements, confirming once again that a small, high-performing subset is sufficient in that case.

Unsurprisingly, ensemble models require significantly more computing time than individual models, with algorithms like LGBM and RF contributing disproportionately to this overhead. ENSACC ensembles are especially time-intensive, as they often include the most computationally demanding models. In contrast, ENSTIME ensembles maintain a more manageable runtime profile by design, though they still incur a higher computational cost than their individual components due to the cumulative effect of combining multiple models.

Figure 3 presents the point forecast accuracy of each ensemble model across various retraining scenarios for both the M5 and VN1 datasets. To facilitate comparison, the results are expressed in relative terms with respect to the baseline retraining scenarios ($r = 7$ for M5 and $r = 1$ for VN1). The RMSSE trajectories of the ENSACC ensembles exhibit remarkable stability across retraining periods. In the M5 dataset, their accuracy remains virtually unchanged regardless of retraining frequency, while in the VN1 dataset, performance even improves in some scenarios. For most

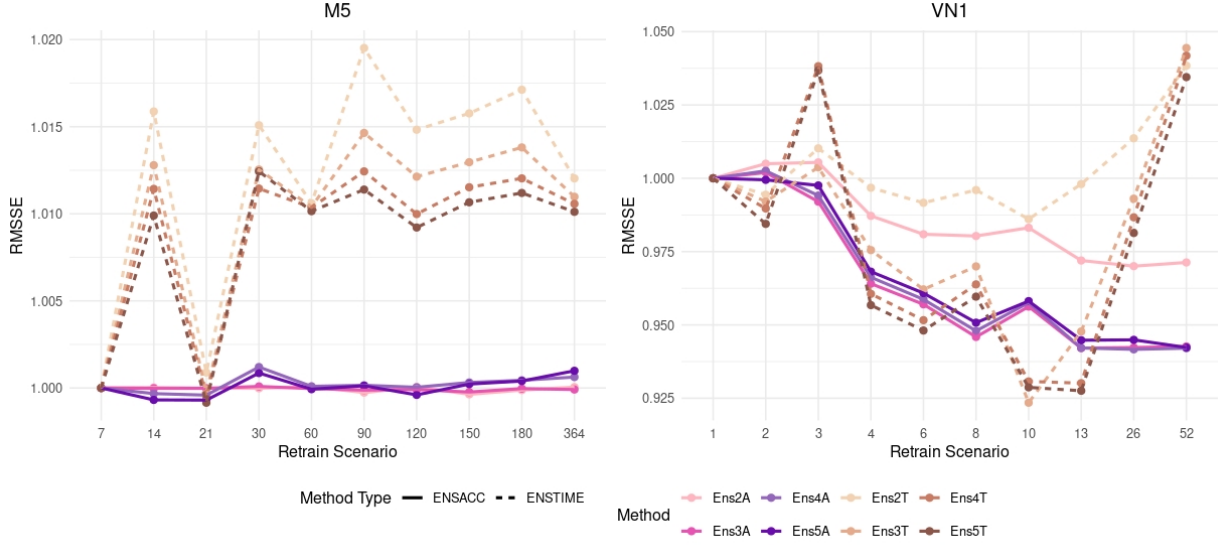


Figure 3: RMSSE results for each ensemble and retrain scenario combination in relative terms with respect to the baseline scenario, $r = 7$ for the M5 dataset and $r = 1$ for the VN1 dataset.

ensemble models, particularly under low-frequency retraining (i.e., periodic updates), performance is nearly indistinguishable from the baseline. Although slight accuracy degradation is observed for the ENSTIME ensembles under less frequent retraining, it remains modest—below 2% in M5 and under 5% in VN1, even in the no-retraining condition. These findings build upon and extend the results of Zanotti (2025), suggesting that reducing retraining frequency does not significantly harm the point forecast accuracy of global model ensembles. This robustness likely stems from the relative stability of the datasets, which do not exhibit substantial structural changes or concept drift. Under such conditions, base models continue to capture underlying demand patterns effectively over time, and ensemble combinations further enhance stability as retraining becomes less frequent.

Similarly, Figure 4 illustrates the relative accuracy of the ensemble models in a probabilistic forecasting context, as measured by the Multi-Quantile Loss (MQL). For the M5 dataset, we observe a clear trend where probabilistic accuracy decreases as retraining becomes less frequent, indicating that regular updates are beneficial for maintaining high-quality uncertainty estimates. This pattern holds across the different ensemble strategies. However, for ENSACC ensembles, the decline in accuracy is minimal for frequent retraining scenarios and becomes slightly more noticeable at higher levels of retraining, though it remains within 5 percentage points. A consistent performance gap is also evident between ENSACC and ENSTIME ensembles, with the former outperforming the

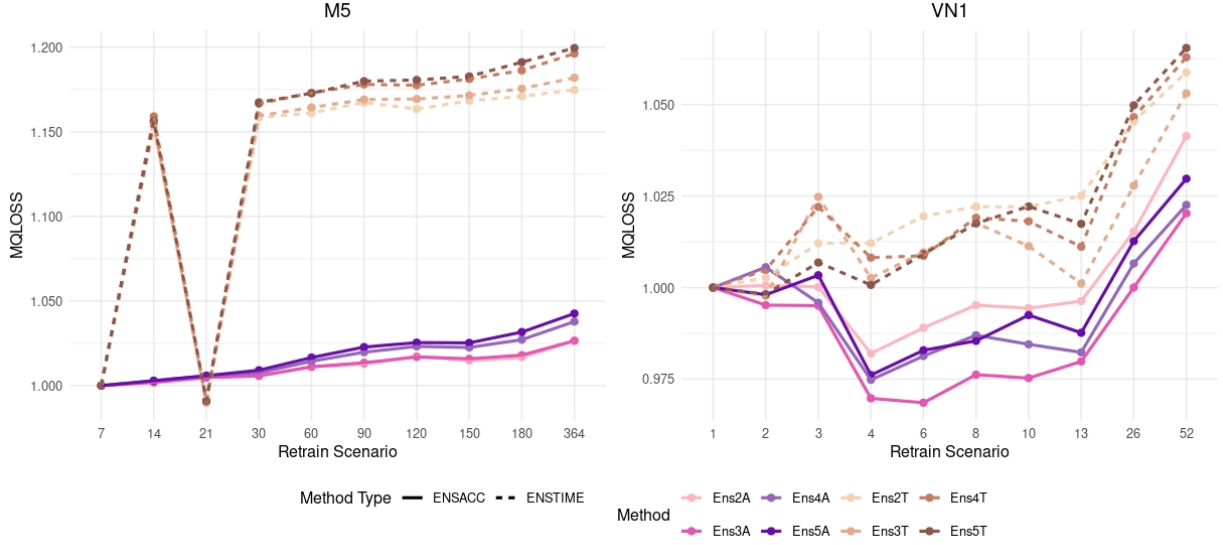


Figure 4: MQL results for each ensemble and retrain scenario combination in relative terms with respect to the baseline scenario, $r = 7$ for the M5 dataset and $r = 1$ for the VN1 dataset.

latter as the retraining scenario increases. This difference is largely attributable to the inclusion of CatBoost in the ENSTIME ensembles, a model that shows relatively weak performance in this setting. In contrast, the VN1 dataset reveals a different behavior. Here, the relationship between retraining frequency and probabilistic accuracy follows a near-convex pattern. Performance initially improves with less frequent retraining, peaking around $r = 4$, and begins to decline thereafter. This suggests that very frequent updates may not be necessary for probabilistic performance in stable, lower-frequency datasets like VN1. Nevertheless, ENSACC ensembles continue to outperform ENSTIME ones across the retraining scenarios. Interestingly, the rate of performance deterioration with increased retraining window is comparable across both ensemble strategies in VN1, indicating that the marginal impact of retraining frequency is relatively uniform across the proposed ensemble techniques. Also these results on the probabilistic performance of ensembles are comparable to those obtained by Zanotti (2025) on the base models.

The Friedman-Nemenyi tests, used to compare the accuracy produced by the same model over different retraining scenarios, confirm the previous results (see Supplementary material). Specifically, in the evaluation of both point forecast accuracy and probabilistic forecast accuracy, some level of periodic retraining is at least as good as the continuous retraining scenario.

Figure 5 illustrates the relative computational time (CT) across different retraining scenarios

for both the M5 and VN1 datasets. As anticipated, CT decreases sharply as the retraining interval increases, with the reduction following an approximately exponential pattern. This effect is more pronounced in the VN1 dataset, likely due to its smaller size compared to M5. Specifically, moving from the baseline retraining scenario to the first periodic setting ($r = 14$) results in a roughly 33% reduction in computing time for M5, while for VN1, CT is nearly halved. As dataset size increases,

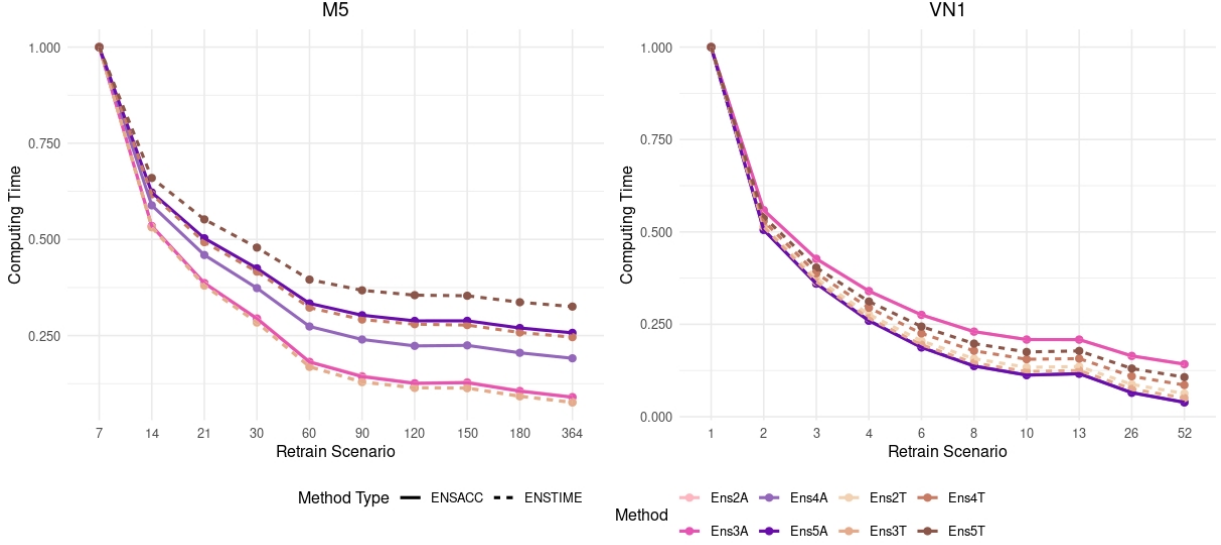


Figure 5: CT results for each ensemble and retrain scenario combination in relative terms with respect to the baseline scenario, $r = 7$ for the M5 dataset and $r = 1$ for the VN1 dataset.

the marginal gains in CT from reducing retraining frequency appear to depend more on the number of models in the ensemble (ensemble depth) than on the ensemble type (ENSACC vs. ENSTIME). Simpler ensembles composed of only 2 or 3 models benefit most from reduced retraining, exhibiting steeper declines in CT. Moreover, ENSACC and ENSTIME ensembles show very similar CT profiles when matched by depth, suggesting that computational efficiency in this context is more a function of ensemble size than selection strategy. This distinction is less pronounced in the VN1 dataset, given its smaller scale and shorter time series. These findings have direct implications for forecasting cost management, emphasizing the importance of considering retraining frequency and ensemble size when selecting models for large-scale operational use.

Overall, the evidence from Figures 3, 4, and 5, along with the statistical significance tests, indicates that reducing the retraining frequency of ensemble models does not negatively impact their predictive accuracy. At the same time, it lowers the computational time required to generate

forecasts. However, effectively leveraging this time reduction depends on carefully considering the ensemble size. As more models are included in the ensemble, the relative benefit of periodic retraining diminishes compared to using only a single model as in Zanutti (2025). Consequently, both the retraining frequency and the ensemble depth play a critical role in managing computational efficiency, and ultimately the cost, of ensemble-based forecasting systems.

As previously noted, computational time can be directly translated into monetary cost for organizations. In line with Nikolopoulos & Petropoulos (2018); Fotios Petropoulos & Spiliotis (2024); Zanutti (2025), we adopted standard pricing assumptions for cloud computing services to estimate both the overall cost of forecasting and the specific costs associated with each retraining scenario for the base global models and each ensemble ³. Figures 6, 7, 8 9, and 10 illustrate the estimated forecasting costs for a large-scale retailer, assuming a computing service rate of \$3.5 per hour, with 200,000 unique SKUs and 5,000 stores. Total forecast production cost, computed by combining training time and forecast generation under a cloud-computing pricing model, highlights the trade-off central to our study. ENSACC ensembles incur the highest costs, rapidly reaching millions of dollars under frequent retraining. In contrast, ENSTIME ensembles offer a more sustainable alternative, reducing costs by over 50% in many cases. Moreover, as dataset size increases, the trade-off between forecast accuracy and cost becomes significantly more pronounced. This is clearly illustrated in Figures 7 and 8. For the VN1 dataset, ensemble methods offer noticeable improvements in both point and probabilistic forecasting accuracy compared to most individual base models, while incurring only a modest increase in cost. In this context, there is little reason to favor ENSTIME ensembles over ENSACC ones, as the cost difference is minor and the performance gain from accuracy-based combinations is evident. However, the situation is quite different for the M5 dataset. Here, ensemble models, particularly those optimized for accuracy, are significantly more expensive than individual models. Additionally, the performance advantage of ENSACC ensembles over the most accurate base models is minimal in terms of both RMSSE and MQL. In contrast, ENSTIME ensembles can achieve comparable levels of accuracy with substantially lower additional cost, making them a more attractive option. It is also clear that increasing the number of models in an ensemble does not necessarily lead to proportional gains in forecasting accuracy, which raises questions about

³It is important to note that the costs have been normalized by the number of SKUs in each dataset, allowing for direct comparison across datasets. This normalization enables meaningful conclusions to be drawn regarding the impact of time series frequency on forecasting costs.

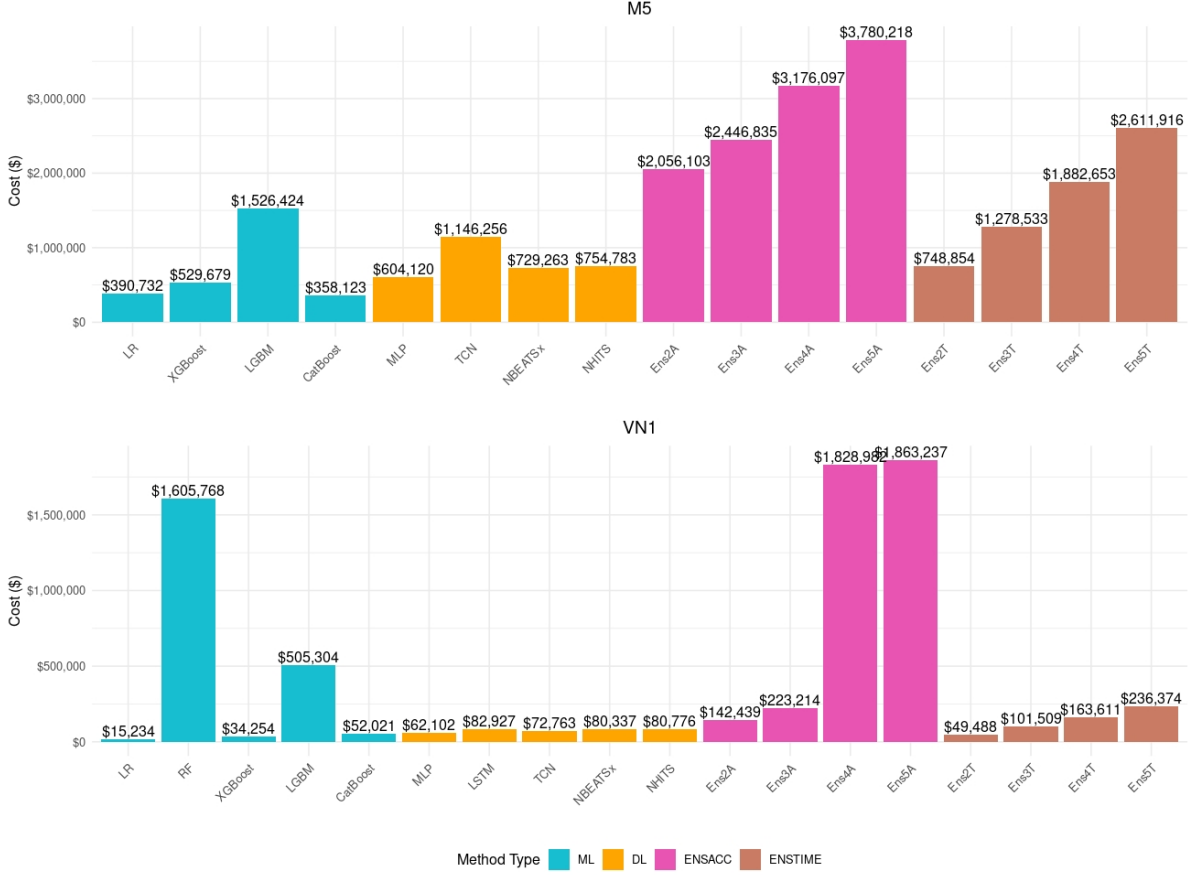


Figure 6: Cost analysis. Overall cost results for the M5 and VN1 datasets.

the cost-effectiveness of larger combinations. Notably, forecast combinations that disregard the computational efficiency of their base models can lead to excessive and unjustified costs. This is particularly evident in the case of Ens4A and Ens5A on the VN1 dataset, where the inclusion of Random Forest (a highly expensive model in this context) results in disproportionately high forecasting costs with only marginal accuracy improvements.

Finally, 9 and 10 show the comparison between base models and ensembles across all retraining scenarios. As expected, forecasting costs decrease exponentially with less frequent retraining. Across both datasets, ENSTIME ensembles consistently incur lower costs than their ENSACC counterparts. Moreover, the cost-saving benefits of reduced retraining frequency become more pronounced as dataset size increases. In the VN1 dataset, cost differences between retraining scenarios remain relatively small, even for ensemble models, suggesting limited sensitivity to retraining frequency. In

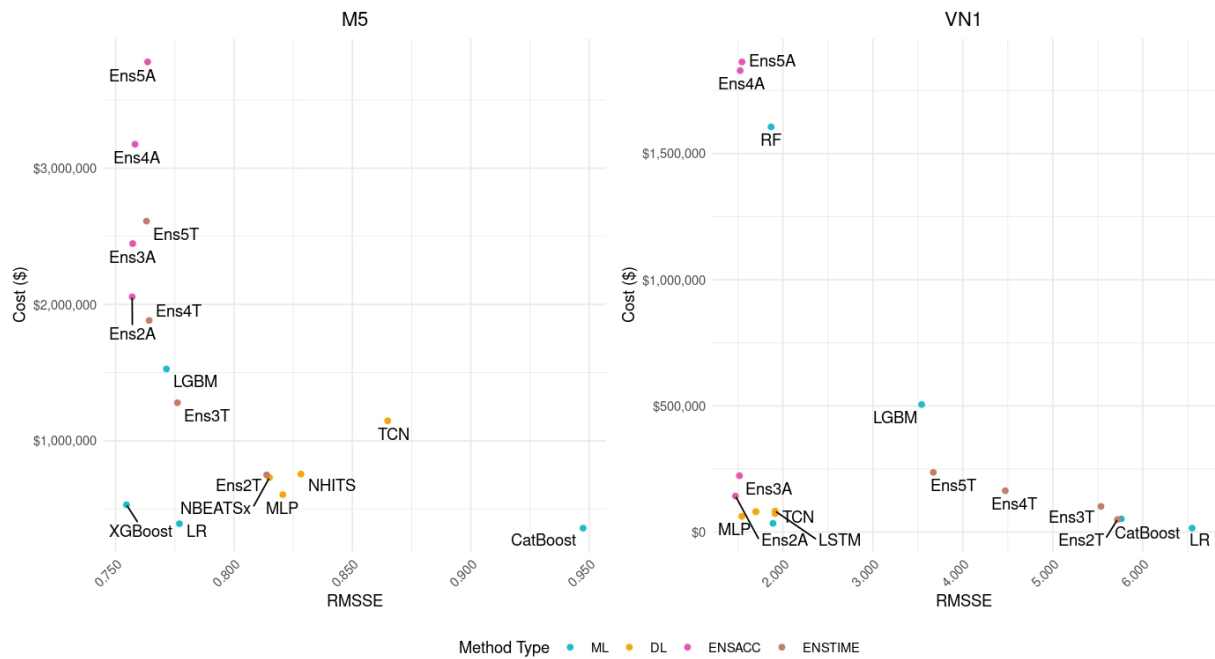


Figure 7: Cost analysis. RMSSE vs Cost (\$) for the M5 and VN1 datasets.

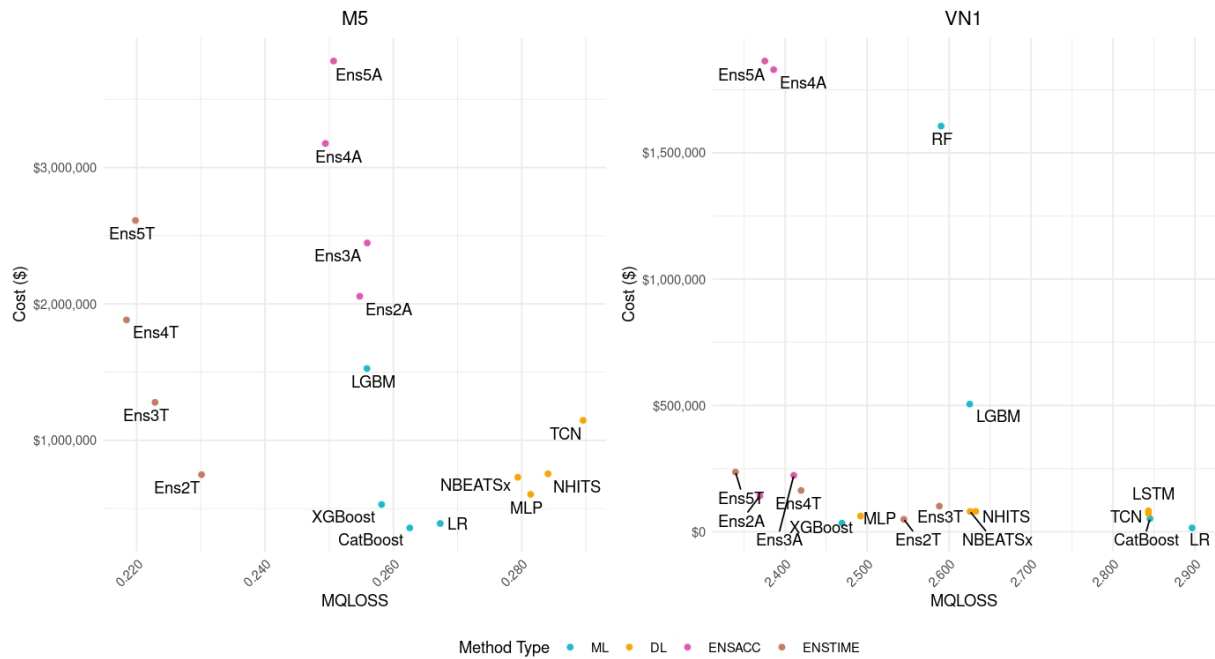


Figure 8: Cost analysis. MQL vs Cost (\$) for the M5 and VN1 datasets.

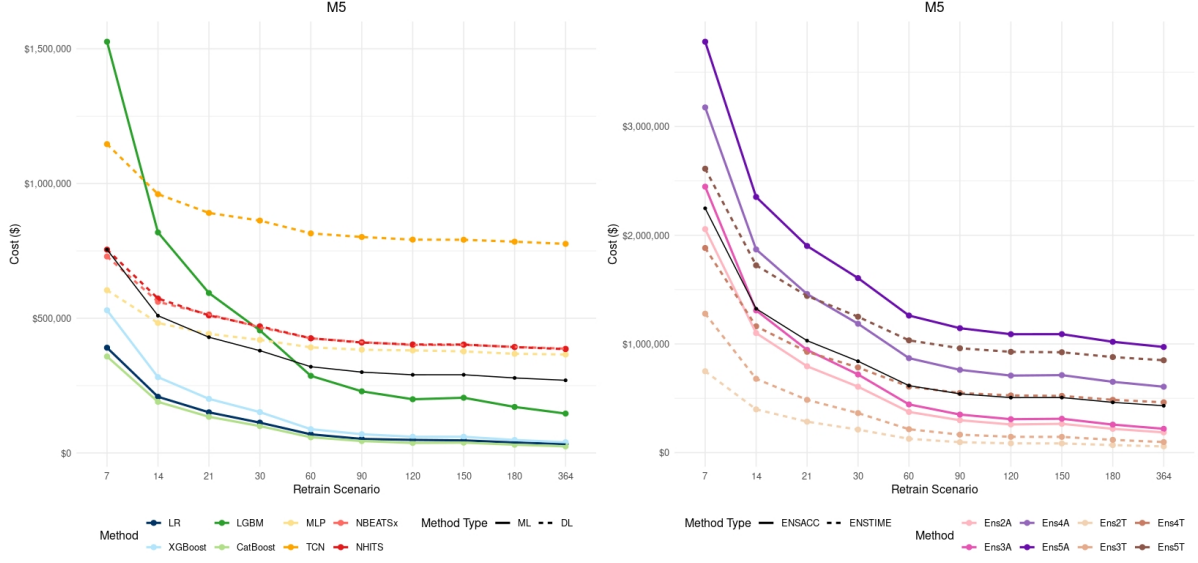


Figure 9: Cost analysis. Comparison between base models and ensembles for the M5 dataset. Each method and retrain scenario combination are shown in relative terms with respect to the baseline scenario, $r = 7$.

contrast, the M5 dataset exhibits steady, incremental reductions in cost as retraining becomes less frequent. For example, in the M5 setting, the average cost of continuous retraining for ensemble models is approximately \$2,250,000, more than triple the average cost of individual base models. When retraining is entirely eliminated, this figure drops to around \$500,000, representing a cost reduction of over 75%. However, even in the no-retraining scenario, ensemble models remain roughly twice as expensive as the average base model. Interestingly, the cost gap between ENSACC and ENSTIME ensembles narrows for smaller ensembles (e.g., those with two or three base models) as retraining frequency decreases. This indicates that with less frequent retraining, accuracy-based combinations (ENSACC) become increasingly comparable to efficiency-based ones (ENSTIME) in terms of forecasting cost. Thus, infrequent retraining can help align performance-optimized ensemble strategies with more budget-conscious forecasting objectives.

One might argue that, for a large retailer, these costs (and the potential savings) are relatively minor. However, it's important to emphasize that the cost reductions achieved through a combination of less frequent retraining, smaller ensemble sizes, and more efficient model combinations typically come with no loss in forecasting accuracy, particularly for point forecasts. This makes such strategies both economically and operationally attractive. These results further support our findings. ENSACC ensembles dominate in accuracy but at a significant computational premium. ENSTIME ensembles

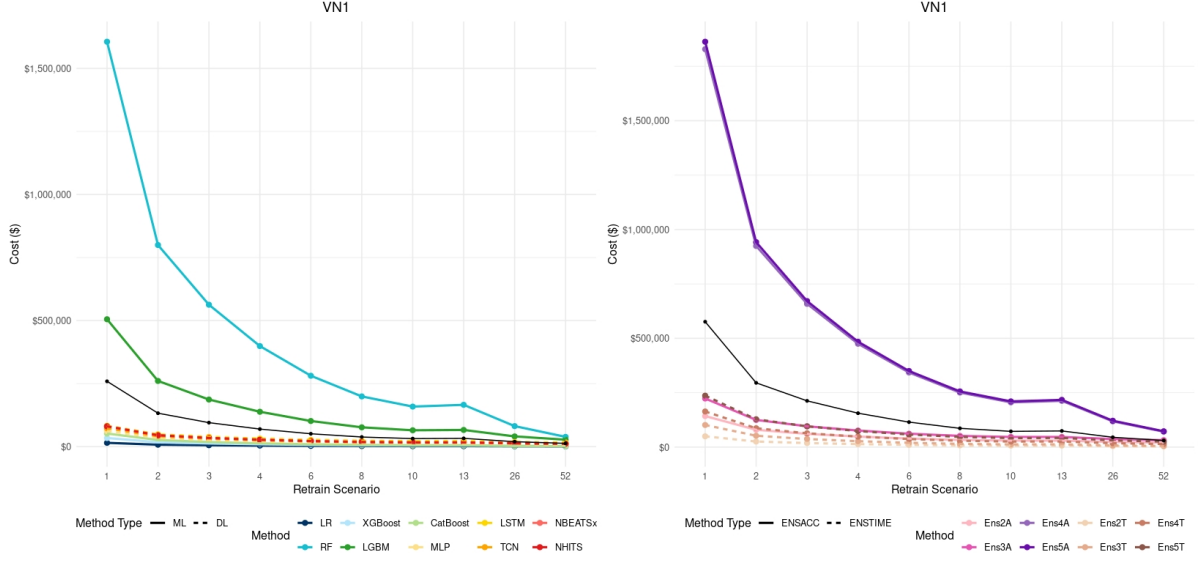


Figure 10: Cost analysis. Comparison between base models and ensembles for the VN1 dataset. Each method and retrain scenario combination are shown in relative terms with respect to the baseline scenario, $r = 1$.

represent a middle ground. Importantly, ensembles of just two or three models (e.g., ENS2A, ENS3T) often achieve near-optimal performance, providing a compelling trade-off between complexity, cost, and accuracy, and lowering the retraining frequency can be a good strategy to improve this trade-off even in the context of ensemble learning. This comprehensive analysis reinforces the core message of the article: ensemble methods offer strong accuracy benefits, but their cost can be substantial, especially under frequent retraining. Balancing accuracy and sustainability requires a deliberate model selection and retraining design, with lightweight ensembles, like ENSTIME, being a possible solution in large-scale forecasting systems.

5. Conclusions

This study investigated the cost-effectiveness of ensemble learning in global time series forecasting, focusing on the interplay between forecast accuracy, computational efficiency, and retraining strategies. We evaluated ten global forecasting models, spanning both classical machine learning and deep learning techniques, alongside eight ensemble configurations, across two large-scale and industry-relevant datasets: the M5 and VN1 retail forecasting datasets. By systematically comparing different ensemble sizes, selection criteria (accuracy-based vs. efficiency-based), and retraining

frequencies, our aim was to assess whether the accuracy gains of ensemble methods justify their additional computational cost, particularly in large-scale operational settings.

Our findings showed that ensemble models consistently improve both point and probabilistic forecasting performance over most individual base models. This improvement is also relevant for probabilistic accuracy, where ensembles provide better uncertainty quantification through the aggregation of diverse quantile predictions. The results support the notion that ensemble learning offers a robust mechanism for improving predictive reliability and mitigating individual model weaknesses, even in the context of global modeling. However, these benefits come at a cost. Ensemble methods are substantially more computationally intensive than single-model approaches, particularly when composed of complex or resource-heavy algorithms such as LightGBM or Random Forest. These computational demands translate directly into higher forecasting costs under cloud-based computing environments.

Our analysis further distinguished between two ensemble design philosophies: ENSACC (accuracy-based) and ENSTIME (efficiency-based). While ENSACC ensembles generally deliver higher accuracy, especially when composed of two or three strong base models, they are also more expensive. In contrast, ENSTIME ensembles, composed of lightweight, computationally efficient models, can achieve comparable accuracy at a much lower cost, making them an attractive compromise in budget-sensitive applications. Most importantly, we find that increasing the number of models in an ensemble does not always yield proportional accuracy gains and often results in rapidly escalating costs. Small ensembles (with two or three base models) are often sufficient to capture most of the benefit in accuracy.

Crucially, our study explored the effects of retraining frequency on ensemble performance and cost. Even if these effects are more pronounced for the base models alone, we showed that less frequent retraining can reduce computational time, without compromising point forecast accuracy, also in the context of ensembles. In many scenarios, especially for relatively stable datasets, even the no-retraining setup performed nearly as well as continuous retraining, confirming and extending the findings from previous work on global models (Zanotti, 2025). Probabilistic forecasting accuracy is more sensitive to retraining, especially for higher-frequency datasets like M5, but performance declines remain modest for moderate retraining intervals. These insights are highly relevant for real-world forecasting systems, where minimizing operational cost without sacrificing accuracy

is essential. Indeed, organizations can achieve economic benefits by optimizing their retraining strategies without compromising the forecast quality.

From a practical standpoint, our findings offer several actionable guidelines for forecasters and organizations deploying large-scale forecasting systems. First, ensembles should be kept small and strategically designed, prioritizing accuracy or efficiency depending on the business context. Second, frequent retraining is often unnecessary. Overall, monthly retraining emerges as a practical compromise for balancing probabilistic forecast accuracy with computational costs. However, if the primary forecasting objective is point prediction, even less frequent retraining intervals can be adopted without significantly affecting performance. Third, ENSTIME strategies represent a viable, low-cost pathway to ensemble forecasting, especially when forecast robustness is desired without substantial computational investment. Retailers and other large-scale forecasters can leverage these insights to build forecasting systems that are both accurate and sustainable. Indeed, these findings also carry broader implications for the sustainability of cross-learning driven forecasting systems. Reducing the frequency of retraining not only lowers operational costs but also leads to significant energy savings, thereby enhancing the environmental sustainability of forecasting processes. This aligns with the principles of "Green AI," which advocates for the responsible use of computational resources to minimize the ecological footprint of machine learning applications.

While our findings provide strong evidence on the effects of ensembling global forecasting models, some limitations remain. First, we restricted our experiments to two retail datasets, which, while comprehensive and realistic, may not capture the full diversity of time series behaviors found in other domains such as finance, health care, or energy. Additionally, this study operates under the assumption that the data-generating process remains stable, without notable trends or concept drift. However, in many real-world scenarios, this assumption may not hold, potentially affecting the reliability of less frequent retraining strategies. Moreover, our ensemble strategies used simple averaging without dynamic weighting or more sophisticated stacking techniques, which may further improve performance at the cost of added complexity and computation time. We also assumed static cost estimates for computational resources, whereas real-world costs may vary based on cloud infrastructure, parallelism, or vendor-specific pricing. We also deliberately focused on a global modeling approach only, but local or hybrid frameworks may offer different cost-accuracy trade-offs worth exploring. Finally, we encourage further studies to explore the possible trade-off between

cost and forecast stability in the context of ensemble learning. In fact, the stability of forecasts is a highly relevant yet underexplored topic, with important implications for model selection in time series forecasting.

In conclusion, our study reinforces the value of ensemble forecasting in global modeling, showing that well-designed ensemble strategies, particularly those leveraging periodic retraining and lightweight components, can deliver high accuracy at a fraction of the computational cost. Striking the right balance between accuracy, efficiency, and retraining frequency is key to deploying scalable and sustainable forecasting systems in practice.

References

- Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, 140, 112896. URL: <https://www.sciencedirect.com/science/article/pii/S0957417419306128>. doi:<https://doi.org/10.1016/j.eswa.2019.112896>.
- Bandara, K., Hewamalage, H., Liu, Y.-H., Kang, Y., & Bergmeir, C. (2021). Improving the accuracy of global forecasting models using time series data augmentation. *Pattern Recognition*, 120, 108148. URL: <https://www.sciencedirect.com/science/article/pii/S0031320321003356>. doi:<https://doi.org/10.1016/j.patcog.2021.108148>.
- Barrow, D. K., & Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: Implications for inventory management. *International Journal of Production Economics*, 177, 24–33. URL: <https://www.sciencedirect.com/science/article/pii/S0925527316300226>. doi:<https://doi.org/10.1016/j.ijpe.2016.03.017>.
- Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213. URL: <https://www.sciencedirect.com/science/article/pii/S0020025511006773>. doi:<https://doi.org/10.1016/j.ins.2011.12.028>. Data Mining for Software Trustworthiness.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. doi:10.1023/A:1010933404324.
- Busetti, F. (2017). Quantile aggregation of density forecasts. *Oxford Bulletin of Economics and Statistics*, 79, 495–512. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/obes.12163>. doi:<https://doi.org/10.1111/obes.12163>. arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/obes.12163>.
- Challu, C., Olivares, K. G., Oreshkin, B. N., Garza, F., Mergenthaler-Canseco, M., & Dubrawski, A. W. (2022). N-hits: Neural hierarchical interpolation for time series forecasting. *ArXiv*, abs/2201.12886. URL: <https://api.semanticscholar.org/CorpusID:246430557>.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16* (p. 785–794). New

- York, NY, USA: Association for Computing Machinery. URL: <https://doi.org/10.1145/2939672.2939785>. doi:10.1145/2939672.2939785.
- Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32, 754–762. URL: <https://www.sciencedirect.com/science/article/pii/S0169207016000327>. doi:<https://doi.org/10.1016/j.ijforecast.2015.12.005>.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30. URL: <http://jmlr.org/papers/v7/demsar06a.html>.
- Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, 38, 1283–1318. URL: <https://www.sciencedirect.com/science/article/pii/S016920701930192X>. doi:<https://doi.org/10.1016/j.ijforecast.2019.06.004>. Special Issue: M5 competition.
- Fotios Petropoulos, E. S., Yael Grushka-Cockayne, & Spiliotis, E. (2024). Wielding occam’s razor: Fast and frugal retail forecasting. *Journal of the Operational Research Society*, 0, 1–20. URL: <https://doi.org/10.1080/01605682.2024.2421339>. doi:10.1080/01605682.2024.2421339. arXiv:<https://doi.org/10.1080/01605682.2024.2421339>.
- Getzner, J., Charpentier, B., & Günnemann, S. (2023). Accuracy is not the only metric that matters: Estimating the energy consumption of deep learning models. URL: <https://arxiv.org/abs/2304.00897>. arXiv:2304.00897.
- Godahewa, R., Bandara, K., Webb, G. I., Smyl, S., & Bergmeir, C. (2021a). Ensembles of localised models for time series forecasting. *Knowledge-Based Systems*, 233, 107518. URL: <https://www.sciencedirect.com/science/article/pii/S0950705121007802>. doi:<https://doi.org/10.1016/j.knosys.2021.107518>.
- Godahewa, R. W., Bergmeir, C., Webb, G. I., Hyndman, R., & Montero-Manso, P. (2021b). Monash time series forecasting archive. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. URL: <https://openreview.net/forum?id=wEc1mgAjU->.
- Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37, 788–832. doi:10.1007/s10618-022-00894-5.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9, 1735–1780. URL: <https://doi.org/10.1162/neco.1997.9.8.1735>. doi:10.1162/neco.1997.9.8.1735.
- Howard, A., & Makridakis, S. (2020). M5 forecasting - accuracy. <https://kaggle.com/competitions/m5-forecasting-accuracy>. Kaggle.
- Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, 36, 1420–1438. URL: <https://www.sciencedirect.com/science/article/pii/S0169207020300224>. doi:<https://doi.org/10.1016/j.ijforecast.2020.02.005>.
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22, 679–688. URL: <https://www.sciencedirect.com/science/article/pii/S0169207006000239>. doi:<https://doi.org/10.1016/j.ijforecast.2006.03.001>.
- Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., & Callot, L. (2020). Criteria for classifying forecasting methods. *International Journal of Forecasting*, 36, 167–177. URL: <https://www.sciencedirect.com/science/article/pii/S0169207019301529>. doi:<https://doi.org/10.1016/j.ijforecast.2019.05.008>. M4 Competition.
- Januschowski, T., Wang, Y., Torkkola, K., Erkkilä, T., Hasson, H., & Gasthaus, J. (2022). Forecasting with

- trees. *International Journal of Forecasting*, 38, 1473–1481. URL: <https://www.sciencedirect.com/science/article/pii/S0169207021001679>. doi:<https://doi.org/10.1016/j.ijforecast.2021.10.004>. Special Issue: M5 competition.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. volume 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.
- Kolassa, S. (2011). Combining exponential smoothing forecasts using akaike weights. *International Journal of Forecasting*, 27, 238–251. URL: <https://www.sciencedirect.com/science/article/pii/S0169207010001032>. doi:<https://doi.org/10.1016/j.ijforecast.2010.04.006>.
- Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, 32, 788–803. URL: <https://www.sciencedirect.com/science/article/pii/S0169207016000315>. doi:<https://doi.org/10.1016/j.ijforecast.2015.12.004>.
- Kolassa, S. (2020). Why the “best” point forecast depends on the error or accuracy measure. *International Journal of Forecasting*, 36, 208–211. URL: <https://www.sciencedirect.com/science/article/pii/S0169207019301359>. doi:<https://doi.org/10.1016/j.ijforecast.2019.02.017>. M4 Competition.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022a). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38, 1346–1364. URL: <https://www.sciencedirect.com/science/article/pii/S0169207021001874>. doi:<https://doi.org/10.1016/j.ijforecast.2021.11.013>. Special Issue: M5 competition.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022b). The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38, 1325–1336. URL: <https://www.sciencedirect.com/science/article/pii/S0169207021001187>. doi:<https://doi.org/10.1016/j.ijforecast.2021.07.007>. Special Issue: M5 competition.
- Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., & Winkler, R. L. (2022c). The m5 uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting*, 38, 1365–1385. URL: <https://www.sciencedirect.com/science/article/pii/S0169207021001722>. doi:<https://doi.org/10.1016/j.ijforecast.2021.10.009>. Special Issue: M5 competition.
- Montero-Manso, P. (2023). How to leverage data for time series forecasting with artificial intelligence models: Illustrations and guidelines for cross-learning. In M. Hamoudia, S. Makridakis, & E. Spiliotis (Eds.), *Forecasting with Artificial Intelligence: Theory and Applications* (pp. 123–162). Cham: Springer Nature Switzerland. URL: https://doi.org/10.1007/978-3-031-35879-1_6. doi:10.1007/978-3-031-35879-1_6.
- Montero-Manso, P., & Hyndman, R. J. (2021). Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37, 1632–1653. URL: <https://www.sciencedirect.com/science/article/pii/S0169207021000558>. doi:<https://doi.org/10.1016/j.ijforecast.2021.03.004>.
- Newbold, P., & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society. Series A (General)*, 137, 131–165. URL: <http://www.jstor.org>.

org/stable/2344546.

- Nikolopoulos, K., & Petropoulos, F. (2018). Forecasting for big data: Does suboptimality matter? *Computers & Operations Research*, 98, 322–329. URL: <https://www.sciencedirect.com/science/article/pii/S0305054817301193>. doi:<https://doi.org/10.1016/j.cor.2017.05.007>.
- Nixtla (2022). Nixtla - state-of-the-art time series and forecasting software. <https://nixtlaverse.nixtla.io/>. Nixtla.
- Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499. URL: <http://dblp.uni-trier.de/db/journals/corr/corr1609.html#OordDZSVGKSK16>.
- Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=r1ecqn4YwB>.
- Pawlikowski, M., & Chorowska, A. (2020). Weighted ensemble of statistical models. *International Journal of Forecasting*, 36, 93–97. URL: <https://www.sciencedirect.com/science/article/pii/S0169207019301190>. doi:<https://doi.org/10.1016/j.ijforecast.2019.03.019>. M4 Competition.
- Petropoulos, F., & et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, 38, 705–871. URL: <https://www.sciencedirect.com/science/article/pii/S0169207021001758>. doi:<https://doi.org/10.1016/j.ijforecast.2021.11.001>.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems NIPS'18* (p. 6639–6649). Red Hook, NY, USA: Curran Associates Inc.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 386–408. doi:10.1037/h0042519.
- Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green ai. *Commun. ACM*, 63, 54–63. URL: <https://doi.org/10.1145/3381831>. doi:10.1145/3381831.
- Semenoglou, A.-A., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, 37, 1072–1084. URL: <https://www.sciencedirect.com/science/article/pii/S0169207020301850>. doi:<https://doi.org/10.1016/j.ijforecast.2020.11.009>.
- Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36, 75–85. URL: <https://www.sciencedirect.com/science/article/pii/S0169207019301153>. doi:<https://doi.org/10.1016/j.ijforecast.2019.03.017>. M4 Competition.
- Smyl, S., & Hua, N. G. (2019). Machine learning methods for gefcom2017 probabilistic load forecasting. *International Journal of Forecasting*, 35, 1424–1431. URL: <https://www.sciencedirect.com/science/article/pii/S0169207019300172>. doi:<https://doi.org/10.1016/j.ijforecast.2019.02.002>.
- Spiliotis, E., & Petropoulos, F. (2024). On the update frequency of univariate forecasting models. *European Journal of Operational Research*, 314, 111–121. URL: <https://www.sciencedirect.com/science/article/pii/S0377221723006859>. doi:<https://doi.org/10.1016/j.ejor.2023.08.056>.

- Stankeviciute, K., M. Alaa, A., & van der Schaar, M. (2021). Conformal time-series forecasting. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems* (pp. 6216–6228). Curran Associates, Inc. volume 34. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/312f1ba2a72318edaaa995a67835fad5-Paper.pdf.
- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16, 437–450. URL: <https://www.sciencedirect.com/science/article/pii/S0169207000000650>. doi:[https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0). The M3- Competition.
- Vandeput, N. (2024). Vn1 forecasting - accuracy challenge. <https://www.datasource.ai/en/home/data-science-competitions-for-startups/phase-2-vn1-forecasting-accuracy-challenge/description>. DataSource.ai.
- Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Berlin, Heidelberg: Springer-Verlag.
- Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 39, 1518–1547. URL: <https://www.sciencedirect.com/science/article/pii/S0169207022001480>. doi:<https://doi.org/10.1016/j.ijforecast.2022.11.005>.
- Zanotti, M. (2025). Do global forecasting models require frequent retraining? URL: <https://arxiv.org/abs/2505.00356>. arXiv:2505.00356.

Supplementary material

In this section, we provide tables and figures related to the empirical results of the M5 and VN1 datasets.

The Tables 6 and 7 show the forecast accuracy of the different ensembles along the examined retrain scenarios for the M5 daily dataset, while Table 8 depicts the computing time in seconds.

Method	7	14	21	30	60	90	120	150	180	364
Ens2A	0.757	0.757	0.757	0.757	0.757	0.757	0.757	0.757	0.757	0.757
Ens3A	0.757	0.757	0.757	0.757	0.757	0.757	0.757	0.757	0.757	0.757
Ens4A	0.758	0.758	0.758	0.759	0.758	0.758	0.758	0.758	0.759	0.759
Ens5A	0.763	0.763	0.763	0.764	0.763	0.764	0.763	0.764	0.764	0.764
Ens2T	0.814	0.827	0.814	0.826	0.822	0.830	0.826	0.827	0.828	0.824
Ens3T	0.776	0.786	0.776	0.786	0.784	0.787	0.785	0.786	0.787	0.785
Ens4T	0.764	0.773	0.764	0.773	0.772	0.774	0.772	0.773	0.773	0.772
Ens5T	0.763	0.771	0.762	0.772	0.771	0.772	0.770	0.771	0.772	0.771

Table 6: M5 RMSSE values for each method and retrain scenario combination.

The Tables 9 and 10 show the forecast accuracy of the different ensembles along the examined retrain scenarios for the VN1 weekly dataset, while Table 11 depicts the computing time in seconds.

Figures 11 and 12 show the results of the Friedman-Nemenyi test in the context of point forecast accuracy for the M5 and VN1 datasets respectively. Figures 13 and 14 present the test results related to the probabilistic evaluation.

The costs tables 12 and 13 show the estimated cost in real values of each scenario for M5 daily data.

The costs tables 14 and 15 show the estimated cost in real values of each scenario for VN1 weekly data.

Method	7	14	21	30	60	90	120	150	180	364
Ens2A	0.255	0.255	0.256	0.256	0.258	0.258	0.259	0.259	0.259	0.262
Ens3A	0.256	0.256	0.257	0.257	0.259	0.259	0.260	0.260	0.261	0.263
Ens4A	0.249	0.250	0.251	0.251	0.253	0.254	0.255	0.255	0.256	0.259
Ens5A	0.251	0.251	0.252	0.253	0.255	0.256	0.257	0.257	0.259	0.261
Ens2T	0.230	0.265	0.228	0.267	0.267	0.269	0.268	0.269	0.269	0.270
Ens3T	0.223	0.257	0.221	0.258	0.260	0.261	0.261	0.261	0.262	0.263
Ens4T	0.218	0.253	0.216	0.255	0.256	0.257	0.257	0.258	0.259	0.261
Ens5T	0.220	0.254	0.218	0.257	0.258	0.259	0.260	0.260	0.262	0.264

Table 7: M5 MQL values for each method and retrain scenario combination.

Method	7	14	21	30	60	90	120	150	180	364
Ens2A	59846	32015	23124	17659	10899	8679	7551	7717	6374	5417
Ens3A	71219	38084	27513	20935	12911	10197	8962	9076	7507	6382
Ens4A	92445	54416	42478	34519	25299	22166	20632	20758	18966	17648
Ens5A	110029	68460	55343	46738	36710	33322	31713	31730	29678	28287
Ens2T	21797	11591	8297	6178	3705	2803	2494	2472	2025	1674
Ens3T	37214	19780	14141	10591	6272	4826	4243	4223	3427	2834
Ens4T	54797	33824	27005	22811	17684	15982	15323	15196	14139	13473
Ens5T	76024	50156	41971	36394	30071	27951	26994	26878	25598	24739

Table 8: M5 CT values for each method and retrain scenario combination.

Method	1	2	3	4	6	8	10	13	26	52
Ens2A	1.472	1.479	1.480	1.453	1.444	1.443	1.447	1.431	1.428	1.430
Ens3A	1.517	1.520	1.504	1.462	1.451	1.435	1.450	1.429	1.429	1.430
Ens4A	1.524	1.528	1.515	1.472	1.461	1.445	1.459	1.436	1.435	1.436
Ens5A	1.544	1.543	1.541	1.495	1.484	1.468	1.480	1.459	1.459	1.455
Ens2T	5.719	5.687	5.778	5.701	5.671	5.696	5.639	5.708	5.797	5.939
Ens3T	5.536	5.492	5.556	5.400	5.326	5.369	5.112	5.246	5.497	5.782
Ens4T	4.472	4.426	4.643	4.296	4.256	4.310	4.162	4.159	4.412	4.659
Ens5T	3.672	3.615	3.807	3.513	3.481	3.523	3.410	3.406	3.603	3.798

Table 9: VN1 RMSSE values for each method and retrain scenario combination.

Method	1	2	3	4	6	8	10	13	26	52
Ens2A	2.369	2.371	2.370	2.326	2.343	2.358	2.356	2.360	2.405	2.467
Ens3A	2.410	2.399	2.399	2.338	2.335	2.353	2.351	2.362	2.411	2.459
Ens4A	2.386	2.399	2.376	2.326	2.341	2.355	2.349	2.344	2.402	2.440
Ens5A	2.375	2.371	2.383	2.318	2.334	2.341	2.357	2.346	2.405	2.446
Ens2T	2.545	2.551	2.575	2.576	2.594	2.601	2.601	2.608	2.660	2.694
Ens3T	2.588	2.582	2.652	2.595	2.613	2.634	2.617	2.591	2.660	2.725
Ens4T	2.419	2.431	2.473	2.439	2.440	2.466	2.463	2.446	2.532	2.572
Ens5T	2.339	2.335	2.355	2.341	2.361	2.380	2.391	2.380	2.456	2.492

Table 10: VN1 MQL values for each method and retrain scenario combination.

Method	1	2	3	4	6	8	10	13	26	52
Ens2A	2205	1228	940	748	607	508	460	460	361	312
Ens3A	3456	1932	1477	1175	952	796	722	721	569	492
Ens4A	28318	14312	10192	7352	5307	3880	3183	3291	1831	1094
Ens5A	28849	14586	10391	7498	5414	3962	3252	3361	1876	1125
Ens2T	766	395	287	212	157	120	102	104	67	48
Ens3T	1572	807	574	419	306	229	192	197	118	79
Ens4T	2533	1341	983	747	571	452	394	398	277	217
Ens5T	3660	1980	1477	1141	893	724	641	652	477	392

Table 11: VN1 CT values for each method and retrain scenario combination.

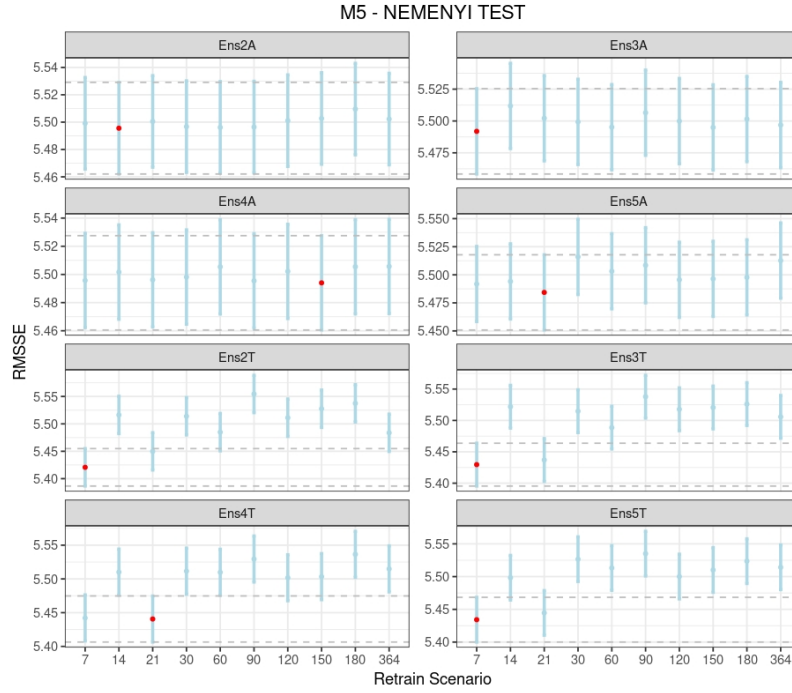


Figure 11: M5 Friedman-Nemenyi test results based on RMSSE.

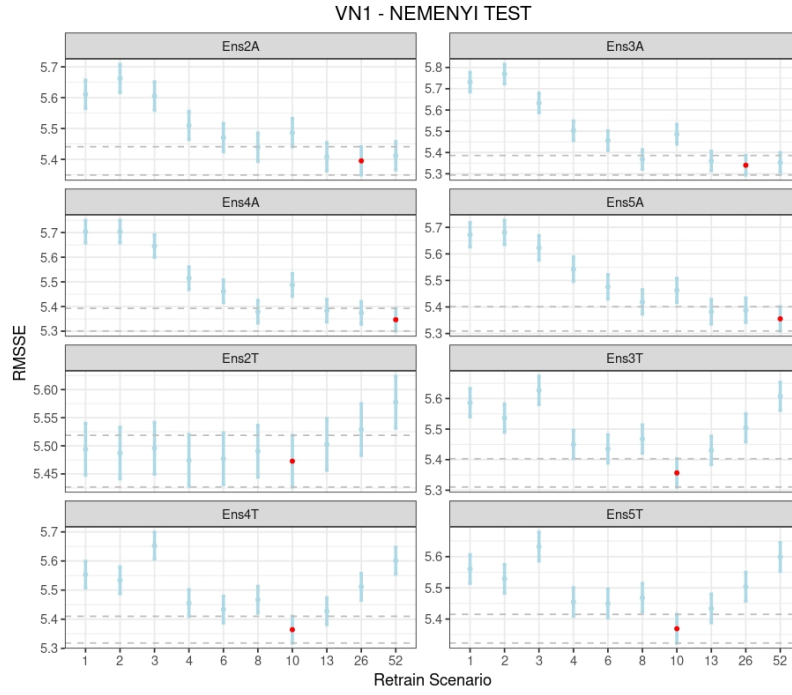


Figure 12: VN1 Friedman-Nemenyi test results based on RMSSE.

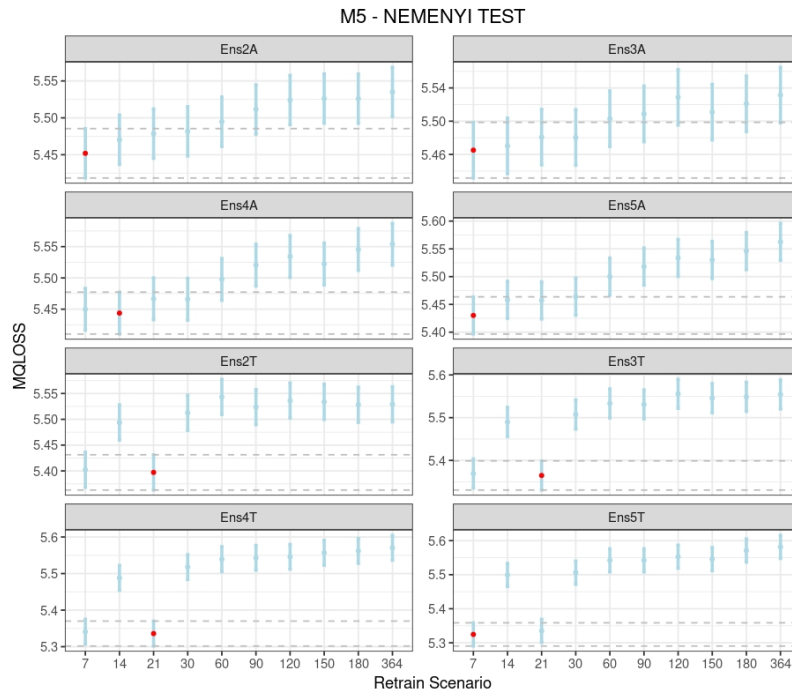


Figure 13: M5 Friedman-Nemenyi test results based on MQL.

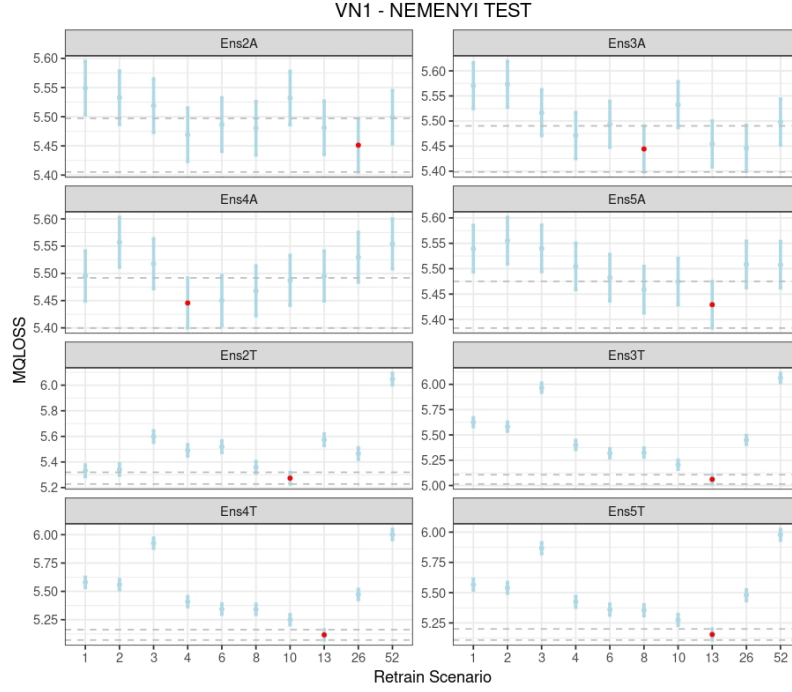


Figure 14: VN1 Friedman-Nemenyi test results based on MQL.

Method	7	14	21	30	60	90	120	150	180	364
LR	390,732	208,499	150,791	112,547	69,132	52,131	48,474	46,678	38,937	33,151
XGBoost	529,679	281,358	200,778	151,635	88,200	69,495	60,082	60,155	48,189	39,861
LGBM	1,526,424	818,569	593,676	455,075	286,262	228,698	199,337	204,972	170,802	146,252
CatBoost	358,123	189,714	134,258	99,699	58,167	44,164	37,206	38,266	30,622	24,362
MLP	604,120	482,505	441,981	419,823	392,046	383,287	380,697	376,968	368,023	365,525
TCN	1,146,256	960,626	890,924	862,432	815,006	801,612	791,801	791,348	784,201	776,181
NBEATSx	729,263	561,105	514,166	466,683	425,593	411,232	400,953	401,352	393,673	387,058
NHITS	754,783	573,233	510,655	470,077	425,684	409,844	402,731	402,394	393,071	385,470
Average	754,922	509,451	429,654	379,746	320,011	300,058	290,160	290,267	278,440	269,733

Table 12: M5 estimated costs for each base method and retrain scenario combination (in \$).

Method	7	14	21	30	60	90	120	150	180	364
Ens2A	2,056,103	1,099,927	794,454	606,710	374,461	298,193	259,419	265,127	218,991	186,113
Ens3A	2,446,835	1,308,426	945,244	719,256	443,594	350,323	307,893	311,805	257,928	219,264
Ens4A	3,176,097	1,869,531	1,459,411	1,185,940	869,187	761,555	708,845	713,157	651,601	606,322
Ens5A	3,780,218	2,352,036	1,901,392	1,605,763	1,261,233	1,144,843	1,089,543	1,090,125	1,019,624	971,847
Ens2T	748,854	398,213	285,049	212,246	127,299	96,295	85,680	84,944	69,559	57,513
Ens3T	1,278,533	679,571	485,827	363,881	215,499	165,789	145,762	145,099	117,748	97,375
Ens4T	1,882,653	1,162,076	927,808	783,704	607,545	549,077	526,459	522,067	485,770	462,900
Ens5T	2,611,916	1,723,181	1,441,975	1,250,387	1,033,138	960,309	927,412	923,419	879,444	849,958
Average	2,247,651	1,324,120	1,030,144	840,985	616,494	540,797	506,376	506,967	462,583	431,411

Table 13: M5 estimated costs for each ensemble method and retrain scenario combination (in \$).

Method	1	2	3	4	6	8	10	13	26	52
LR	15,234	7,839	5,731	4,292	3,205	2,508	2,140	2,190	1,463	1,099
RF	1,605,768	799,597	562,896	398,954	281,246	199,214	158,959	165,975	81,529	38,865
XGBoost	34,254	17,687	12,796	9,413	6,944	5,262	4,473	4,534	2,863	2,005
LGBM	505,304	260,721	186,904	138,460	101,775	76,738	65,013	66,335	40,721	27,701
CatBoost	52,021	26,594	18,515	13,366	9,611	7,022	5,798	5,977	3,311	2,023
MLP	62,102	34,489	26,431	21,149	17,121	14,391	13,027	13,015	10,255	8,891
LSTM	82,927	49,019	38,851	31,962	27,022	23,465	21,829	21,932	18,415	16,652
TCN	72,763	41,291	31,910	25,464	20,791	17,579	15,983	16,365	12,936	11,284
NBEATSx	80,337	44,837	34,283	27,179	22,057	18,387	16,686	16,702	13,085	11,277
NHITS	80,776	45,458	34,688	27,585	22,336	18,607	16,888	16,855	13,405	11,601
Average	259,149	132,753	95,301	69,782	51,211	38,317	32,080	32,988	19,798	13,140

Table 14: VN1 estimated costs for each base method and retrain scenario combination (in \$).

Method	1	2	3	4	6	8	10	13	26	52
Ens2A	142,439	79,326	60,714	48,328	39,177	32,778	29,713	29,717	23,340	20,167
Ens3A	223,214	124,784	95,402	75,913	61,513	51,385	46,602	46,572	36,745	31,769
Ens4A	1,828,982	924,381	658,299	474,867	342,760	250,599	205,561	212,547	118,274	70,634
Ens5A	1,863,237	942,068	671,095	484,280	349,703	255,861	210,033	217,080	121,137	72,638
Ens2T	49,488	25,526	18,527	13,705	10,149	7,769	6,613	6,724	4,327	3,103
Ens3T	101,509	52,120	37,042	27,071	19,760	14,792	12,410	12,700	7,637	5,126
Ens4T	163,611	86,609	63,473	48,220	36,880	29,183	25,437	25,715	17,892	14,017
Ens5T	236,374	127,900	95,383	73,684	57,671	46,762	41,420	42,080	30,828	25,301
Average	576,106	295,339	212,491	155,758	114,701	86,140	72,223	74,141	45,022	30,344

Table 15: VN1 estimated costs for each ensemble method and retrain scenario combination (in \$).