DEMS
Department of Economics, Management and Statistics

DEMS WORKING PAPER SERIES

On the stability of global forecasting models

Marco Zanotti

No. 553 – June 2025

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

On the stability of global forecasting models

Marco Zanotti[a,*]

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

Abstract

Forecast stability, that is the consistency of predictions over time, is essential in business settings where sudden shifts in forecasts can disrupt planning and erode trust in predictive systems. Despite its importance, stability is often overlooked in favor of accuracy, particularly in global forecasting models. In this study, we evaluate the stability of point and probabilistic forecasts across different retraining frequencies and ensemble strategies using two large retail datasets (M5 and VN1). To do this, we introduce a new metric for probabilistic stability (MQC) and analyze ten different global models and four ensemble configurations. The results show that less frequent retraining not only preserves but often improves forecast stability, while ensembles, especially those combining diverse pool of models, further enhance consistency without sacrificing accuracy. These findings challenge the need for continuous retraining and highlight ensemble diversity as a key factor in reducing forecast stability. The study promotes a shift toward stability-aware forecasting practices, offering practical guidelines for building more robust and sustainable prediction systems.

Keywords: Time series, Demand forecasting, Forecasting competitions, Cross-learning, Global models, Forecast stability, Vertical stability, Machine learning, Deep learning, Conformal predictions

JEL: C53, C52, C55

1. Introduction

In recent years, global forecasting models, those trained across multiple time series simultaneously, have emerged as a powerful alternative to traditional local approaches, particularly in large-scale applications such as retail demand forecasting. While their accuracy and efficiency have been widely

*Corresponding author

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

studied, less attention has been paid to the stability of their forecasts over time. In many operational contexts, forecasts are produced regularly and serve as the foundation for critical business decisions, from inventory planning to resource allocation. In such settings, forecast stability becomes a key requirement; not only should forecasts be accurate, but they should also remain reasonably consistent over time. Indeed, forecast stability refers to the consistency of predictions produced by a forecasting model as new data becomes available. As defined by Godahewa, Bergmeir, Erkin Baz, Zhu, Song, García & Benavides (2025), forecast stability can be categorized into two primary forms: vertical stability, which concerns the consistency of forecasts for the same target date across different forecast origins, and horizontal stability, which addresses the smoothness of forecasts across the forecast horizon from a single origin. Both types play a distinct role: vertical stability helps avoid costly forecast revisions that can disrupt planning cycles, while horizontal stability prevents erratic behavior across time steps that could lead to inefficient operations and amplification of demand fluctuations, such as the bullwhip effect in supply chains (Lee, Padmanabhan & Whang, 1997). In particular, vertical stability may also be seen as temporal robustness, in the sense that a forecasting model is robust to updates (or retraining) as new observations are available, a property that is critical in practice for avoiding frequent changes in business decisions. Forecast instability, indeed, can have serious consequences. Unstable forecasts may lead to frequent and expensive adjustments in supply chain plans, diminish trust in the forecasting system, and complicating the decision-making processes, leading to suboptimal business outcomes.

Nevertheless, despite the importance of forecast stability, it is still common practice to evaluate forecasting models only on their accuracy, even in frequent retraining settings where stability is more likely to be compromised. This can also be due to an inherent perception that there exists a trade-off between accuracy and stability. As new information becomes available, updated forecasts may naturally differ from earlier versions, ideally improving in accuracy at the expense of stability. Conversely, overly stable forecasts might sacrifice accuracy by ignoring valuable new information. Ensembling techniques, which integrate multiple forecasts from different models, have been proposed as a mechanism to concurrently reduce bias and variance while enhancing stability (Wang, Hyndman, Li & Kang, 2023). By aggregating diverse predictions, ensembling methods can alleviate the impact of individual model volatility and potentially achieve a more favorable equilibrium between accuracy and stability. Even if the trade-off between forecasting accuracy and

stability presents a significant challenge for organizations seeking to balance predictive performance with operational consistency, recent research suggests that stability and accuracy are not necessarily in conflict, and both can be achieved at the same time ((Godahewa et al., 2025), (Van Belle, Crevits & Verbeke, 2023)). Moreover, Zanotti (2025b) demonstrated that the retraining frequency does not harm the forecast accuracy of global models, meaning that models estimated using this forecasting approach retain the same level of performance even when updated less frequently. This has a strong potential implication on the forecast stability: if we can reduce the retraining frequency of global models, then we should obtain more stable predictions, because stability should be a non-increasing function of the retraining frequency.

1.1. Research Question

We aim to address the question *"Are global forecasting models stable?"*. Purposely, we study whether the global modeling approach produces stable forecasts, and we try to understand the effects of retraining on the forecast stability, that is whether avoiding re-estimation for every new observation damages the stability of global models. To address this question, we rely on the two most recent and comprehensive retail forecasting datasets: the M5 and VN1 competition data.

To generally understand the stability (or instability) of the forecasting models, we consider ten distinct global forecasting methods (five from the traditional machine learning domain and five based on commonly used deep neural network architectures). Moreover, we examine stability across a range of retraining scenarios, from no retraining to continuous retraining, by exploring periodic strategies that broadly encompass the most practical and effective approaches.

We also investigate the use of ensembling, or forecasting combinations, as a mean to obtain more stable forecasts. This approach can indeed be significant to mitigate forecast instability, reducing the model variance and bias.

1.2. Contributions

Our contribution is fourfold:

- We provide the first comprehensive study of the forecast stability of global models, using 10 distinct methods, a diverse collection of real-world datasets, and evaluating both point and probabilistic predictions.

- We analyze the relationship between retraining frequency and forecast stability, comparing different scenarios to quantify the impact of frequent retraining in terms of the stability of forecasting.
- We suggest a new metric to evaluate the instability of probabilistic forecasts.
- We present practical guidelines for organizations and practitioners on when and how often to retrain global forecasting models to obtain stable forecasts.

By tackling these aspects, this paper contributes to both the forecasting and machine learning communities by providing insights into the stability 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, and the methods adopted for global forecasting are discussed in 3.2. The concepts related to model update and retrain scenario are explained in 3.3, together with the evaluation strategy adopted, while the metrics used to assess the stability of models are shown in 3.4. In Section 4 we discuss the empirical findings of our study, including forecast stability and ensembling on the different scenarios. Finally, Section 5 contains our summary and conclusions.

2. Related works

The cross-learning approach has seen substantial development in recent years. Today, most time series forecasting studies include at least a benchmark comparison involving global models (GMs), underscoring their growing importance in the field. Semenoglou, Spiliotis, Makridakis & Assimakopoulos (2021) demonstrated the high accuracy of GMs on the M4 competition dataset, while Hewamalage, Bergmeir & Bandara (2022) explored the conditions under which global models are competitive. Additionally, Montero-Manso & Hyndman (2021) and (Montero-Manso, 2023) provided theoretical support showing that GMs can match or surpass the accuracy of local models, with lower complexity and without assuming data similarity. Global models have proven to be the most accurate method in several forecasting domains, including retail demand (Spiliotis, Makridakis, Semenoglou & Assimakopoulos (2022), Bandara, Shi, Bergmeir, Hewamalage, Tran & Seaman

(2019), Juan R Trapero & Fildes (2015)), electricity demand (Buonanno, Caliano, Pontecorvo, Sforza, Valenti & Graditi, 2022), water demand (Groß & Hans, 2024), gas consumption (Gaweł & Paliński, 2024), and crop production (Ibañez & Monterola, 2023). Their effectiveness was particularly evident during the M5 competition (Makridakis, Spiliotis & Assimakopoulos, 2022a), where tree-based models employing cross-learning ranked among the top-performing solutions (Januschowski, Wang, Torkkola, Erkkilä, Hasson & Gasthaus, 2022). To further enhance GM performance, several strategies such as clustering (Godahewa, Bandara, Webb, Smyl & Bergmeir (2021a), Bandara, Bergmeir & Smyl (2020)) and data augmentation (Bandara, Hewamalage, Liu, Kang & Bergmeir, 2021) have been explored. Moreover, new machine learning (Godahewa, Webb, Schmidt & Bergmeir, 2023) and deep learning (Oreshkin, Carpov, Chapados & Bengio, 2020) architectures have been specifically designed to support cross-learning. Recently, research has begun focusing on improving GMs' ability to capture local patterns (Sen, Yu & Dhillon, 2019) and enhance their interpretability (Rajapaksha, Bergmeir & Hyndman, 2023).

From a forecasting stability perspective, the literature is very poor. Godahewa et al. (2025) introduced the first classification of the different types of forecasting stability, that is horizontal or vertical stability, proposing a new model-agnostic framework based on linear combinations of predictions to obtain more stable forecasts. Van Belle et al. (2023), instead, extended an existing deep learning architecture (NBEATS) to optimize forecasts from both a traditional forecast accuracy perspective as well as a forecast stability perspective, directly including an instability component into the loss function of the model. There is also active research on forecast stability within judgmental forecasting ground (Fildes & Goodwin, 2021). However, almost all the literature related to forecast stability is focused on point prediction stability only, possibly because it is not clear how to measure instability in terms of probabilistic forecasting, and maybe also because most machine learning and deep learning methods do not directly output probabilistic predictions (Makridakis, Spiliotis, Assimakopoulos, Chen, Gaba, Tsetlin & Winkler, 2022c). Nevertheless, in many forecasting applications, such as supply chain management, it is crucial to generate and evaluate predictions probabilistically, whether through prediction intervals, quantiles, or full predictive distributions (Fildes, Ma & Kolassa, 2022). Among the methods developed for uncertainty quantification, Vovk, Gammerman & Shafer (2005) introduced Conformal Inference, a model-agnostic framework that offers valid uncertainty estimates and can also be applied in time series forecasting settings

(Stankeviciute, M. Alaa & van der Schaar, 2021).

In the context of model retraining and updating strategies, the most comprehensive work in time series forecasting is by Zanotti (2025b), who showed the positive effects of reducing the retraining frequency of global models on both accuracy and forecasting costs. Spiliotis & Petropoulos (2024) extensively investigated the impact of various retraining scenarios and parameter update methods on model performance. However, their focus was limited to the exponential smoothing family, following the traditional local modeling approach. (Huber & Stuckenschmidt, 2020) touched on retraining within the retail demand domain, but the study was restricted to a small set of models, limited retraining strategies, and a proprietary daily dataset. While these findings are encouraging, there has been little direct investigation into whether global models specifically can retain stability with less frequent updates.

Building on these prior studies, our work directly examines the stability of global models. Moreover, by systematically evaluating a wide range of retraining strategies and their effects on the forecasting stability of various global models, we seek to offer both theoretical insights and practical guidance for promoting more stable forecasting practices.

3. Experimental design

This section presents the empirical analysis carried out to investigate the stability of global models and to determine whether less frequent retraining scenarios can yield stability outcomes comparable to those of the baseline scenario, which involves the most frequent retraining. We begin by describing the datasets used in our experiments, followed by an overview of the machine learning, deep learning, and ensemble models employed. Finally, we detail the instability measures, the various retraining scenarios considered, and the evaluation strategy applied to assess forecast performance.

3.1. Datasets

For our experiments, we employed two retail forecasting datasets: the M5 and the VN1 competition datasets. The M5 competition, part of the well-known M-competitions series organized by Spyros Makridakis and colleagues, aimed to benchmark forecasting methods in the context of retail demand (Makridakis, Spiliotis & Assimakopoulos, 2022b). The M5 dataset (Howard & Makridakis, 2020) includes 3,049 daily time series representing unit sales of Walmart products

across three categories (Food, Hobbies, and Household) sold in ten stores located in California, Texas, and Wisconsin. The data span from 2011 to 2016 and are characterized by high intermittency and a hierarchical structure, enabling forecasts at multiple aggregation levels (e.g., SKU, product category, store, and state). The dataset also includes exogenous variables such as prices, promotions, and special events (e.g., holidays), which can influence demand. The VN1 Forecasting - Accuracy Challenge, organized in October 2024 by Flieber, Syrup Tech, and SupChains, represents the first edition of a new competition series (Vandeput, 2024). The dataset comprises weekly sales data for 15,053 products sold by U.S.-based e-vendors between 2020 and 2024. Unlike the M5, which involves a single retailer (Walmart) and a limited number of physical stores, the VN1 dataset captures sales from 328 warehouses operated by 46 different retailers. As far as we are aware, our study is among the first to evaluate forecasting models on this dataset. Together, these two datasets offer the most recent and comprehensive time series collections related to retail demand, enabling good generalizability of our findings.

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 cases, our analysis focused on the most disaggregated level of the data (i.e., SKUs), where the potential advantages of reducing retraining frequency are likely to be most pronounced. To ensure consistency with the evaluation setup described in Section 3.3, we applied a filtering criterion: for the daily data (M5), only time series with at least two years of observations (730 data points) were retained; for the weekly data (VN1), we considered only those series with at least three years of data (157 observations).

3.2. Forecasting models

In this section, we present an overview of the global models used in our experiments.

Following Zanotti (2025b), let us define $\mathcal{Y}$ as the set of all available time series in a dataset, where each $Y_i \in \mathcal{Y}$ represents an individual time series. Let $\mathcal{F}$ denote the set of possible predictive functions, with each $F \in \mathcal{F}$ corresponding to a specific forecasting model. Without loss of generality,

we assume that all necessary information for prediction is contained within $\mathcal{Y}$. Under the local modeling paradigm, forecasts for a horizon h are generated by training a separate model for each individual time series. This implies that each series Y_i is associated with its own model, characterized by its own set of parameter values.

$$Y_i^h = F(Y_i, \theta_i). \tag{1}$$

In contrast, under the global modeling framework, forecasts for each individual time series are generated using a single model trained on the entire dataset. This approach leverages cross-series information, allowing the model to learn shared patterns and structures across all time series in $\mathcal{Y}$.

$$Y_i^h = F(\mathcal{Y}, \Theta). \tag{2}$$

It is important to note that in the cross-learning methodology, the model parameters Θ are not specific to individual time series but are shared across all series in the dataset. This parameter sharing is a key characteristic of global models, enabling them to generalize patterns across series and potentially reducing the forecast instability.

In our study, we focused exclusively on analyzing the stability of global methods, as our primary goal was to assess whether this modeling approach, unlike the traditional local one, can maintain performance with less frequent retraining. Cross-learning has become the standard in many industries dealing with large-scale time series data, particularly in retail demand forecasting, where regular forecasts for thousands of SKUs are required (Januschowski, Gasthaus, Wang, Salinas, Flunkert, Bohlke-Schneider & Callot, 2020). Therefore, to conduct a comprehensive evaluation of the stability of global models, we incorporated both traditional machine learning algorithms and state-of-the-art deep learning architectures. The selected models are well-established in the time series forecasting literature and represent a range of methodological paradigms, allowing for a broad and informative comparison.

As in Zanotti (2025b), we used five machine learning models and five deep learning models. Machine learning models have proven effective in forecasting tasks due to their ability to capture complex non-linear relationships in data. They are also relatively easy to train and tend to deliver strong performance when leveraging cross-learning techniques. In this study, we experimented with Linear (Pooled) Regression and four widely used tree-based methods. While machine learning

models are easier to train compared to deep learning alternatives, they often rely heavily on high-quality feature engineering (Januschowski et al., 2022). For our experiments, we adopted simplified versions of the feature engineering pipelines used in the top-performing M5 and VN1 solutions. We constructed time series features such as lags, rolling means, expanding means, and calendar features (e.g., year, month, week, day of the week). We also included static metadata such as store, product, category, and location identifiers based on the dataset's frequency. For the M5 dataset, additional external features like special events were also incorporated. Model hyperparameters were selected based on configurations used in top-performing competition solutions when available; otherwise, we relied on the recommended defaults from the respective libraries. We compared five machine learning models for time series forecasting. Linear Regression (LR) is a classical statistical model effective with appropriate feature engineering, often used as a benchmark in global model performance ((Montero-Manso & Hyndman, 2021), (Godahewa, Bergmeir, Webb, Hyndman & Montero-Manso, 2021b)). Random Forest (RF), an ensemble learning method, captures non-linear patterns and is robust to overfitting, making it suitable for demand forecasting ((Breiman, 2001), (Januschowski et al., 2022)). Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LGBM) are gradient-boosted models known for speed and accuracy, with LGBM excelling in large, high-dimensional datasets ((Chen & Guestrin, 2016), (Ke, Meng, Finley, Wang, Chen, Ma, Ye & Liu, 2017), Makridakis et al. (2022a)). Categorical Boosting (CatBoost) specializes in handling categorical data with minimal preprocessing (Prokhorenkova, Gusev, Vorobev, Dorogush & Gulin, 2018).

Deep learning models have gained significant traction in time series forecasting due to their ability to model long-range temporal dependencies and to learn hierarchical representations directly from raw input data (Goodfellow, Bengio & Courville, 2016). Unlike machine learning models, deep learning architectures typically do not require extensive manual feature engineering. They can autonomously learn lagged structures, trend, and seasonality directly from the raw series. Nevertheless, training deep models is often more complex due to the higher number of hyperparameters and their sensitivity to configuration choices, which can significantly affect performance (Smyl, 2020). In our implementation, we followed competition-proven practices by providing only static, calendar, and external covariates as inputs, while using top solutions' guidelines for setting the model hyperparameters. We compared five deep learning models for time series forecasting. Multi-Layer Perceptron

(MLP) is a versatile, efficient neural network used in many time series forecasting tasks (Rosenblatt, 1958), while Recurrent Neural Networks (RNNs), particularly LSTMs and GRUs, capture temporal dependencies in sequential data ((Cho, van Merriënboer, Gulcehre, Bahdanau, Bougares, Schwenk & Bengio, 2014), (Hochreiter & Schmidhuber, 1997)). Temporal Convolutional Networks (TCN) offer an alternative by using causal convolutions to capture long-range dependencies (Van den Oord, Dieleman, Zen, Simonyan, Vinyals, Graves, Kalchbrenner, Senior & Kavukcuoglu, 2016). Neural Basis Expansion Analysis for Time Series (NBEATS) and Neural Hierarchical Interpolation for Time Series (NHITS) are state-of-the-art models for time series forecasting, with NBEATS offering interpretable trend and seasonality components (Oreshkin et al., 2020), and NHITS improving upon NBEATS with hierarchical interpolation mechanisms (Challu, Olivares, Oreshkin, Garza, Mergenthaler-Canseco & Dubrawski, 2022).

Moreover, we also estimated four different ensemble learning models based on Zanotti (2025a). Ensemble learning combines the predictions of multiple base models to improve forecast accuracy and robustness, particularly when individual models capture complementary patterns in the data. In the context of global forecasting, ensembles are especially valuable for mitigating the risk of model instability and overfitting, which can occur when relying on a single method (Wang et al., 2023). In our study, we adopted a simple averaging strategy, which is widely used for its ease of implementation and effectiveness in improving forecast robustness (Claeskens, Magnus, Vasnev & Wang, 2016). Specifically, we constructed ensembles by averaging the forecasts of the top two, top three, top four, and top five most accurate individual models, as determined on out-of-sample performance. The ensemble configurations were designed to reflect practical setups and were validated using the same rolling origin evaluation framework described in Section 3.3. By incorporating ensemble learning into our analysis, we aimed to assess whether combining global models could further enhance forecasting stability under different retraining scenarios. Moreover, by evaluating multiple ensemble sizes, we aimed to understand how the inclusion of additional models affects the stability under different circumstances.

All global 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

Out-of-sample evaluation is a cornerstone of time series forecasting, providing a way to test a model's generalization ability beyond the training period, an essential step given that future data may diverge from historical patterns due to structural breaks, level shifts, or unexpected shocks (Tashman, 2000). Among the various evaluation strategies, the rolling origin evaluation has emerged as the most widely accepted and rigorous method (Bergmeir & Benítez, 2012). It respects the chronological nature of time series data while enabling repeated assessments across multiple forecast origins. In this framework, the time series is split into a training set and a test set, where the model is trained on the former and evaluated on the latter. Forecasts are generated for a defined horizon h, and at each iteration, the forecast origin is shifted forward by a specified step size, typically one, to simulate real-time forecasting. The model is then retrained on the updated training data, using either a fixed-length or expanding window. Stability metrics (see Section 3.4) are averaged over all iterations to provide a robust estimate of forecasting accuracy.

Compared to fixed origin evaluations, rolling origin offers a more nuanced view of a model's robustness by exposing it to a range of temporal conditions (including seasonal patterns, trends, and potential anomalies) that may not be captured in a single evaluation window (Bergmeir & Benítez, 2012). This iterative approach reduces the risk of overfitting to one specific train-test split, and is especially beneficial in operational contexts like retail, logistics, or finance, where forecasts are regularly updated in response to incoming data. It also aligns well with real-world decision-making processes that require forecasts to adapt over time.

In practice, the design of a rolling origin evaluation depends on several parameters: the size of the test set, the length of the forecast horizon, the step size, and the windowing strategy (fixed vs expanding). The expanding window strategy is particularly suited for short time series, as it maximizes the amount of historical data available at each iteration. This approach is widely used in applied settings (Petropoulos & et al., 2022), and it is the one we adopted in our study, especially given the relatively short length of the weekly VN1 series. Following Zanotti (2025b), Table 2 outlines the key parameters used in our experiments, including a test set spanning at least one full year to mitigate seasonal bias, horizons aligned with typical business use cases, and a step size of one to maximize evaluation frequency.

Furthermore, to investigate the impact of retraining frequency on forecasting stability, we

Table 2: 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

examined various retraining scenarios, or retraining windows, as defined in Zanotti (2025b). Each window r indicates the number of new observations after which the model is retrained. These retraining scenarios are tailored to the data frequency, daily for M5 and weekly for VN1, since frequency dictates both the forecast horizon and business review cycles. Table 2 summarizes the selected retraining windows, test periods, and forecast horizons. For instance, in the daily case, $r = 7$ reflects weekly updates. Across all scenarios, we trained global models on aligned datasets, updating the model either completely at each retraining step or not at all. We excluded hyperparameter tuning due to its high cost and marginal expected benefit. Following Zanotti (2025b), we treated $r = 1$ (or $r = 7$ for daily data) as the accuracy benchmark and $r = T$ as the no-retraining baseline, with intermediate values representing periodic retraining strategies.

3.4. Evaluation metrics

3.4.1. Stability metrics

The stability evaluation of point predictive models is a crucial topic in time series forecasting: very few metrics are available to capture models' vertical stability, and there is no consensus in the literature on what metric to use (Godahewa et al., 2025). Van Belle et al. (2023) proposed the symmetric Mean Absolute Percentage Change (sMAPC), which measures the change of one to h-step ahead forecasts obtained by two consecutive forecast origins, providing a measurement of up to which extent the forecasts generated at the first origin are unstable compared to those generated at the second origin.

$$\text{sMAPC} = \frac{200}{h-1} \sum_{t=n+1}^{n+h-1} \frac{|\hat{y}_{t,n} - \hat{y}_{t,n-1}|}{|\hat{y}_{t,n}| - |\hat{y}_{t,n-1}|}. \tag{3}$$

Here, $\hat{y}_{t,n}$ and $\hat{y}_{t,n-1}$ are the forecasts generated for period t with origins n and $n-1$ respectively. The instability across different pairs of consecutive forecasting origins can be obtained by a simple average of the sMAPC values. Lower values imply less unstable predictions.

Beyond point forecasts, our study also emphasizes the probabilistic stability of the models under different retraining strategies. Since most machine learning and deep learning models do not natively produce full predictive distributions, we employed Conformal Inference to construct prediction intervals around point forecasts. Conformal Inference is a flexible and robust framework that provides distribution-free, model-agnostic prediction intervals with guaranteed finite-sample coverage under mild assumptions (Vovk et al., 2005). Although initially developed for i.i.d. data, recent extensions have adapted it for time series applications by relaxing the exchangeability assumption (Stankeviciute et al., 2021). Its advantages (validity guarantees, minimal assumptions, computational simplicity, and effectiveness with limited data) make it especially suitable for our global forecasting setup, where different models are compared across heterogeneous datasets and varying levels of retraining frequency. Because no metric exists to evaluate the vertical stability (or instability) of the resulting probabilistic forecasts, we proposed a new measure, the Multi-Quantile Change (MQC, or Multi-Quantile Loss Change, MQLC), to comprehensively assess the stability of the probabilistic predictions. We defined the Quantile Change and the Multi-Quantile Change as:

$$\mathrm{QC} = \frac{1}{h-1} \sum_{t=n+1}^{n+h-1} \left(q \cdot (\hat{y}_{t,n-1} - \hat{y}_{t,n}) \cdot \mathbb{I}_{\hat{y}_{t,n-1} \geq \hat{y}_{t,n}} + (1-q) \cdot (\hat{y}_{t,n} - \hat{y}_{t,n-1}) \cdot \mathbb{I}_{\hat{y}_{t,n-1} < \hat{y}_{t,n}} \right), \quad (4)$$

$$\mathrm{MQC} = \frac{1}{\mathcal{Q}} \sum_{q \in \mathcal{Q}} \mathrm{QC}(q). \quad (5)$$

The Quantile Change (QC) metric is a measure of the change in forecasted quantiles between two consecutive forecast origins. Given two forecast origins, n and $n-1$, QC quantifies the average adjustment in the predicted quantiles across the forecast horizon h. The formula mirrors the Quantile Loss (Pinball Loss) but replaces the actual observations y_t, with previous forecasts $\hat{y}_{t,n-1}$, thereby shifting the focus from forecast accuracy to forecast stability. The intuition is straightforward: if the forecasts are stable across time (i.e., $\hat{y}_{t,n} \approx \hat{y}_{t,n-1}$), then QC will be small. Large QC values indicate that the model updates its distribution significantly with each new observation, suggesting a lack of temporal robustness in its probabilistic output. By aggregating this across a range of quantiles $\mathcal{Q}$, the Multi-Quantile Change (MQC) captures overall instability across the entire forecast distribution, providing a comprehensive metric for probabilistic stability. One drawback of the MQC is that it is based solely on a finite set of quantiles rather than the full predictive distribution. As a result, it captures changes only at selected points in the distribution (e.g., median, tails), potentially missing more subtle shifts in shape, skewness, or variance that occur

between those quantiles. However, in practical forecasting applications, especially in retail and supply chain domains where decisions are often made based on specific quantile levels (e.g., high quantiles for safety stock), this limitation is mitigated by the fact that the selected quantiles are usually those of greatest operational interest. Moreover, through Conformal Inference, it is possible to calculate as many quantiles as desired at no computational cost, making MQC a pragmatic and model-agnostic metric for capturing distributional instability across time.

We computed 13 quantiles: the median and 6 central prediction intervals (60%, 70%, 80%, 90%, 95% and 99%). Central intervals describe the forecast center, while wider intervals capture tail risks, essential for safety stock decisions in retail (Barrow & Kourentzes, 2016). To ensure reliable quantile estimates, conformal prediction intervals were computed on a validation set at least twice the forecast horizon, which constrained the number of time series used.

We normalized each evaluation metric relative to the baseline retraining scenario, defined by the dataset frequency, to enable consistent comparison across models and retraining windows. To statistically validate our findings, we applied the Friedman-Nemenyi test for multiple comparisons (Demšar, 2006).

3.4.2. Performance metrics

Evaluating point forecast accuracy in time series is a debated topic, with no consensus on the best metric (Hewamalage, Ackermann & Bergmeir, 2023). In our context of SKU-level demand forecasting, characterized by intermittent data, metrics based on absolute or percentage errors are suboptimal (Kolassa, 2020), and due to varying scales across series, a scaled accuracy metric has to be preferred. Hence, we adopted the Root Mean Squared Scaled Error (RMSSE) (Hyndman & Koehler, 2006),

$$\text{RMSSE} = \sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}(y_t - \hat{y}_t)^2}{\frac{1}{n-s}\sum_{t=s+1}^{n}(y_t - y_{t-s})^2}}. \tag{6}$$

which compares the model's squared error to that of a seasonal naive forecast, and it was the official accuracy metric in the M5 competition (Makridakis et al., 2022b). Lower values indicate better performance.

To assess probabilistic accuracy, we evaluated the quantiles produced through the Conformal Inference framework using the Quantile Loss (QL) and Multi-Quantile Loss (MQL).

$$\mathrm{QL} = \frac{1}{h} \sum_{t=n+1}^{n+h} \left(q \cdot (y_t - \hat{y}_t) \cdot \mathbb{I}_{y_t \geq \hat{y}_t} + (1-q) \cdot (\hat{y}_t - y_t) \cdot \mathbb{I}_{y_t < \hat{y}_t} \right), \tag{7}$$

$$\mathrm{MQL} = \frac{1}{\mathcal{Q}} \sum_{q \in \mathcal{Q}} \mathrm{QL}(q). \tag{8}$$

These proper scoring rules assess forecast distribution accuracy (Kolassa, 2016). Moreover, the MQL was the official metric of the M5 Uncertainty competition (Makridakis et al., 2022c).

For this study, we adopted a cloud computing machine NC6s_v3 hosted on Microsoft Azure, with Linux Ubuntu 24 operating system, 1 Graphical Processing Unit, 6 Computing Processing Units, 112GB of memory.

4. Results and discussion

This section discusses the empirical findings of our study, highlighting the interplay between accuracy, probabilistic performance, and stability cost across different retraining strategies and ensemble configurations. We extract insights from both the M5 and VN1 datasets to uncover consistent patterns as well as nuances unique to each dataset's characteristics.

Table 3 summarizes the performance and stability of all forecasting methods evaluated in this study, across both the M5 and VN1 datasets. For each method, the table reports four key metrics per dataset: RMSSE (Root Mean Squared Scaled Error) to evaluate point forecast accuracy, MQL (Multi-Quantile Loss) to assess probabilistic forecast performance, sMAPC (symmetric Mean Absolute Percentage Change) to compare the stability in point forecast settings, and MQC (Multi-Quantile Change) to understand the stability in probabilistic terms. Overall [1], the models evaluated achieved better stability results on the M5 dataset than on VN1. Several factors may account for this difference, such as the larger dataset size and higher frequency of the M5 time series, the availability of rich external regressors (e.g., promotions, special events), and the existence of well-established benchmark hyperparameter settings, many of which were unavailable for VN1 at the time of model training. Moreover, in terms of point forecasting, machine learning models are much more stable than deep learning methods on both datasets, implying that the former are less sensitive to frequent updates. On the contrary, in terms of probabilistic forecasting, deep learning

[1]The overall results are based on the baseline retraining scenario, $r = 7$ for M5 and $r = 1$ for VN1, which is commonly considered the standard in both theoretical literature and practical forecasting applications.

Table 3: Overall forecasting performance and stability for each method across datasets. RMSSE, MQL, sMAPC, and MQC. Minimum values per column are highlighted in bold.

Method	**M5**				**VN1**			
	RMSSE	**MQL**	**sMAPC**	**MQC**	**RMSSE**	**MQL**	**sMAPC**	**MQC**
LR	0.777	0.267	0.188	0.118	6.549	2.896	**0.085**	1.064
RF	–	–	–	–	1.868	2.590	0.092	1.322
XGBoost	**0.755**	0.258	0.075	0.115	1.890	2.469	0.146	1.101
LGBM	0.771	0.256	**0.054**	0.123	3.542	2.625	0.121	1.282
CatBoost	0.947	0.263	0.118	0.127	5.762	2.845	0.320	1.435
MLP	0.821	0.281	0.500	**0.109**	1.543	2.492	0.458	0.844
LSTM	–	–	–	–	1.913	2.843	0.558	1.020
TCN	0.865	0.290	0.499	0.114	1.913	2.843	0.600	**1.020**
NBEATSx	0.815	0.279	0.547	0.110	1.698	2.626	0.596	1.532
NHITS	0.828	0.284	0.566	0.112	1.699	2.632	0.605	1.592
Ens2A	0.757	0.255	0.059	0.117	**1.472**	**2.369**	0.503	1.079
Ens3A	0.757	0.256	0.089	0.117	1.517	2.410	0.514	1.200
Ens4A	0.758	**0.249**	0.095	0.113	1.524	2.386	0.167	1.180
Ens5A	0.763	0.251	0.099	0.111	1.544	2.375	0.164	1.153

architectures show less instability, meaning that these approaches can be a better option when retraining is performed frequently.

In general, ensembling seems to be a good solution to produce stable predictions. However, combining the most accurate models is often not sufficient to reduce the instability. Indeed, ensembles show increasing returns in stability as more models are added, as opposed to the well-known phenomenon in the ensemble literature, where accuracy improvements show diminishing returns from adding more models (Zanotti, 2025a). Interestingly, larger ensembles are usually more effective for stability purposes, implying that model diversity may be a more relevant criterion to create a forecast combinations, in contrast to model performance. This pattern is particularly true for probabilistic forecasting, but it is also evident in point predictions. For instance, in the VN1 dataset, the Ens4A and Ens5A combinations achieve good stability results simply because they contain both deep learning and machine learning models, extending the combination's diversity. Indeed, even if the deep learning models' instability is on average 60%, it is enough to add one single machine learning model to obtain an improvement of almost 45%.

Figure 1 shows the point forecast stability of each model along the different retraining scenarios for the M5 and the VN1 datasets. To allow comparisons, we display the results in relative terms with respect to the baseline scenarios, that is $r = 7$ for M5 and $r = 1$ for VN1. As expected, the sMAPC profiles are non-increasing functions of the retraining frequencies. The stability remains practically the same across update frequencies or even improves as the retraining period increases. Indeed, moving from high to low retraining frequencies, the stability of most global models improves compared to the baseline. In particular, for some models (i.e. CatBoost, LSTM, and TCN) the instability reduction is strong and consistent, reaching very low levels. These results imply that less frequent retraining may have a huge impact on the point forecast stability of global models. This can be explained by the fact that, when a forecasting origin is updated and the model is retrained on the new data, the model's predictions may differ consistently (even using the cross-learning approach) from those produced before the update. Therefore, less retraining is often synonymous with more stability.

Figures 2, and 3 statistically confirms the above results for the M5 and VN1 datasets. It is clear that most periodic retraining scenarios are statistically different to the continuous retraining strategy in terms of point forecast stability at the 5% level.

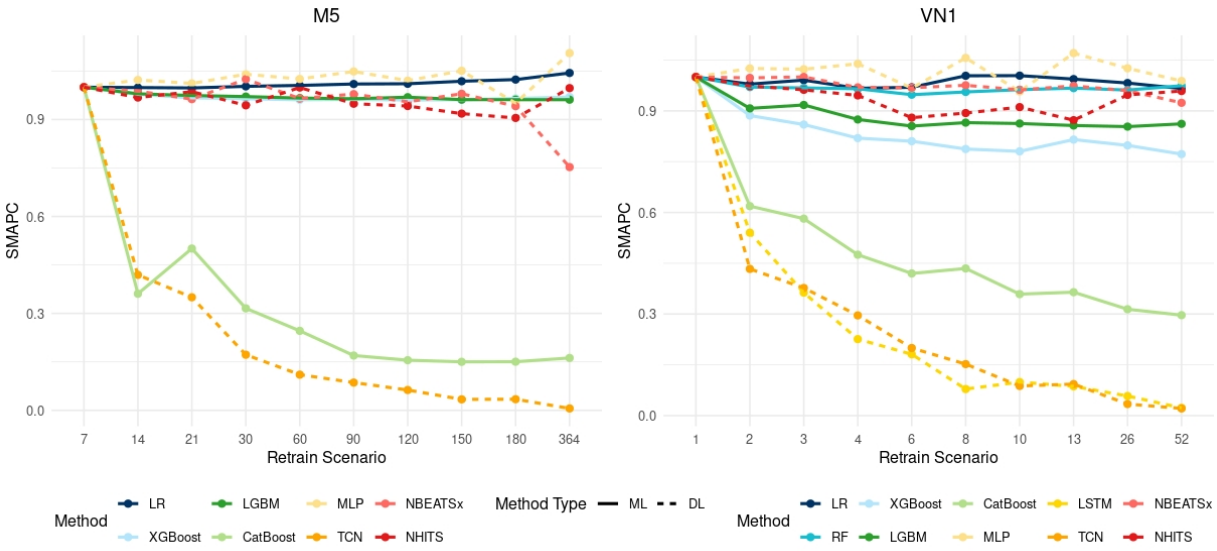

Figure 1: sMAPC results for each method and retrain scenario combination in relative terms with respect to the baseline scenario, $r = 7$ for the M5 dataset and $r = 1$ for the VN1 dataset.

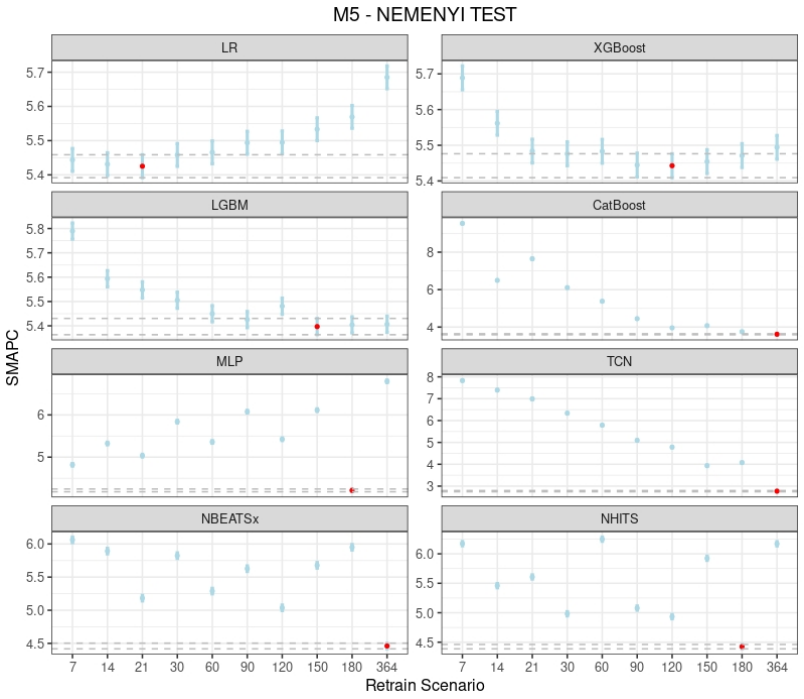

Figure 2: M5 Friedman-Nemenyi test results based on sMAPC.

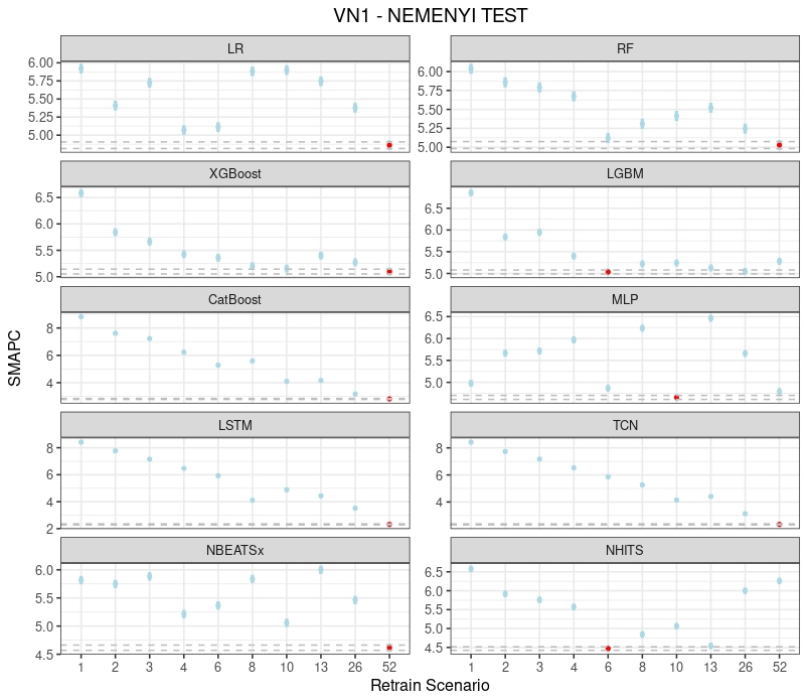

Figure 3: VN1 Friedman-Nemenyi test results based on sMAPC.

In a similar fashion, Figure 4 summarizes the relative stability in a probabilistic forecasting setting. In this context, we observe that, for the M5 dataset, the stability (as measured by Multi Quantile Change) paths are constant across almost every retraining scenario, meaning that less frequent updates neither harm nor improve the probabilistic forecasting stability of most global models. For the VN1 dataset, instead, we observe an almost convex relationship between the stability and the retraining frequency. On average, models' stability improves for low retraining scenarios and then starts deteriorating around $r = 10$, implying that it is possible to avoid retraining up to 10 weeks without negatively impacting on the forecast stability. For both datasets, these results are confirmed by the statistical tests on MQC over the different retrain periods (see supplementary materials).

Overall, the results from Figures 1 and 4, supported by the Friedman-Nemenyi tests, indicate that reducing the retraining frequency of global models does not negatively impact (and may even enhance) forecast stability, for both point and probabilistic forecasts. When considered alongside the findings of Zanotti (2025b), this provides strong evidence against the practice of continuous retraining in global forecasting models. Lower retraining frequencies preserve both accuracy and

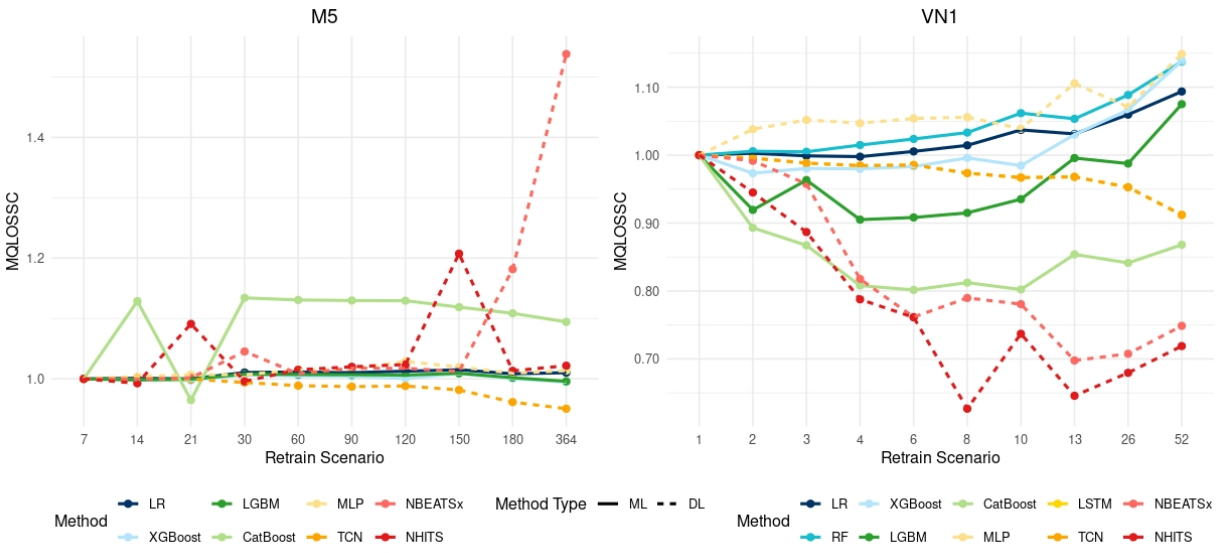

Figure 4: MQC results for each method and retrain scenario combination in relative terms with respect to the baseline scenario, $r = 7$ for the M5 dataset and $r = 1$ for the VN1 dataset.

stability, while offering substantial savings in computational resources.

Figure 5 shows the evolution of point forecast stability, across different retraining scenarios for various ensemble configurations. The results on the M5 dataset indicate that reducing retraining frequency may negatively affect forecast stability. However, this impact is below 1% for most retraining scenarios. Indeed, statistical tests (in supplementary material) on the significance of these differences show that most retraining scenarios are not statistically different from each other, with periodic retraining strategies being at least as good as continuous retraining. The results on the VN1, instead, indicate that there exists a clear benefit in lowering the retraining frequency in terms of forecast stability. Notably, on the M5 dataset the smallest ensemble (Ens2A) consistently achieves the lowest sMAPC values, since it is obtained by combining XGBoost and LGBM that where both the most accurate and the most stable global forecasting models on that set of data. Nonetheless, larger ensembles (such as Ens4A and Ens5A) stability profiles tend to outperform smaller ones, especially at lower retraining frequencies, suggesting that model diversity plays a critical role in stabilizing ensemble forecasts when. Figure 6 further supports these observations in the probabilistic domain. First of all, probabilistic forecasts of ensembles are generally more stable than that of the base models across retraining frequencies, especially in M5, where MQC variations are within 1 percentage point. Moreover, in the VN1 dataset, the almost convex relationship between

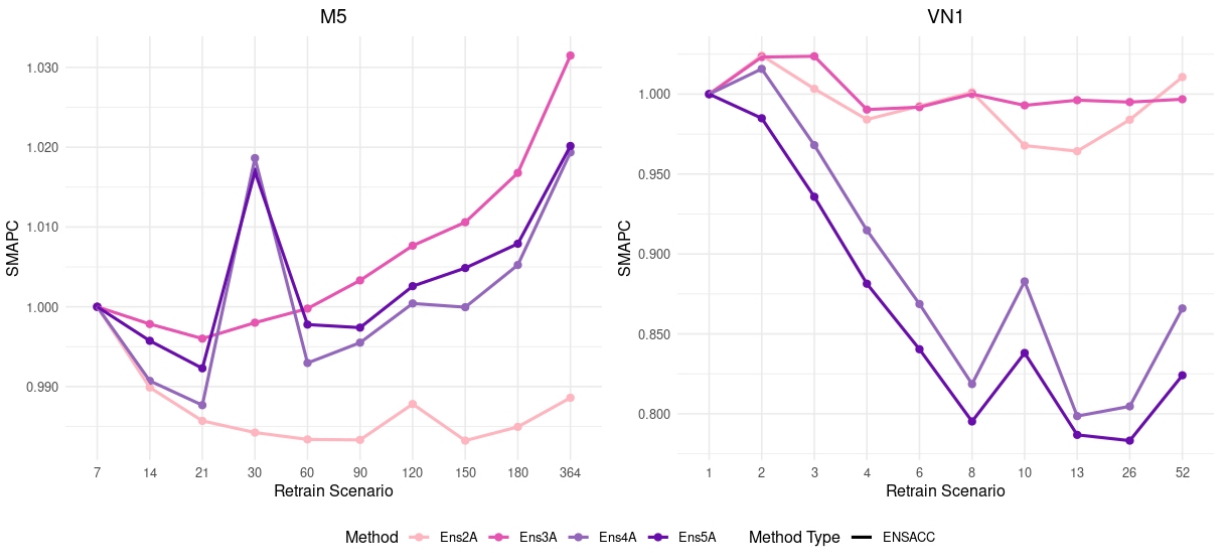

Figure 5: sMAPC results for each ensemble and retrain scenario combination in relative terms with respect to the baseline scenario, $r = 7$ for the M5 dataset and $r = 1$ for the VN1 dataset.

stability and the retraining frequency is flattened, because the variations are more restrained as the retraining period increases, and the decline in stability happens only from $r = 26$ onwards. Then, larger ensembles provide consistently more stable probabilistic performance across most retraining scenarios, confirming the role of the diversity composition of ensembles even in the probabilistic setting. Most importantly, the effect of ensembling on the profiles of stability, as measured by sMAPC or MQC, is that it greatly reduces the variations across the different retraining frequencies, reducing both positive and negative stability changes.

Figures 7 and 8 illustrate the fundamental trade-off between forecast accuracy and stability, comparing RMSSE with sMAPC and MQL with MQC, respectively. These plots reveal that high accuracy does not guarantee high stability. In particular, several deep learning models, achieve low RMSSE or MQL scores but exhibit high instability, especially in VN1, highlighting their sensitivity to changes in the training data and their volatility across forecast origins. Conversely, models like Random Forest, XGBoost, and LGBM strike a more favorable balance, maintaining relatively low RMSSE and sMAPC values. Nevertheless, it is the ensemble models, and especially those with larger size like Ens4A and Ens5A, that most consistently approach the optimality of stability and accuracy. In both the point and probabilistic domains, even very simple ensembles manage to control variance and mitigate instability without sacrificing much, if any, predictive performance.

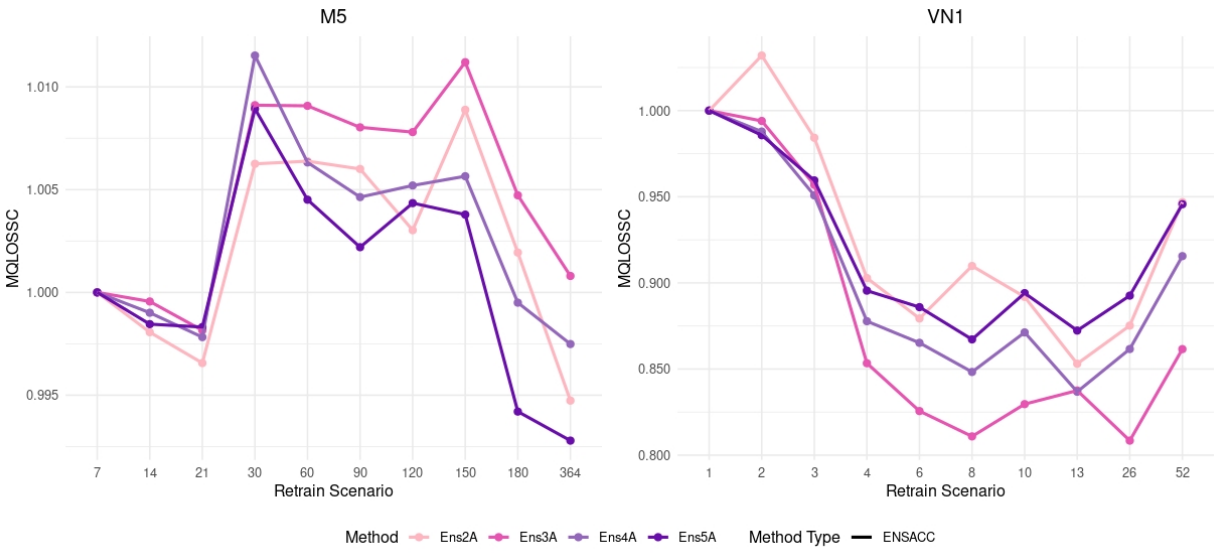

Figure 6: MQC results for each ensemble and retrain scenario combination in relative terms with respect to the baseline scenario, $r = 7$ for the M5 dataset and $r = 1$ for the VN1 dataset.

The overall picture that emerges from these figures reinforces the idea ensembling offers a practical and effective strategy to counteract forecast instability, particularly when constructed from a diverse pool of models.

This set of results has clear practical implications: in operational forecasting systems, especially in domains such as retail and supply chain management, the use of ensembles combined with lower retraining frequencies may offer a robust forecasting process, balancing performance with reliability.

5. Conclusions

This study provides a comprehensive empirical assessment of forecast stability in global models, focusing on the effects of retraining frequency and ensembling across both point and probabilistic forecasting contexts. Our findings demonstrate that global forecasting models are able to produce very stable predictions and that they can achieve even more stability under reduced retraining frequencies, challenging the prevailing wisdom that frequent model updates are necessary to maintain forecast reliability.

In terms of point forecasting, the results consistently show that retraining less frequently does not harm, and often improves, forecast stability. This trend is particularly evident in the M5 dataset, where many models exhibit monotonic or flat stability profiles as the retraining interval increases.

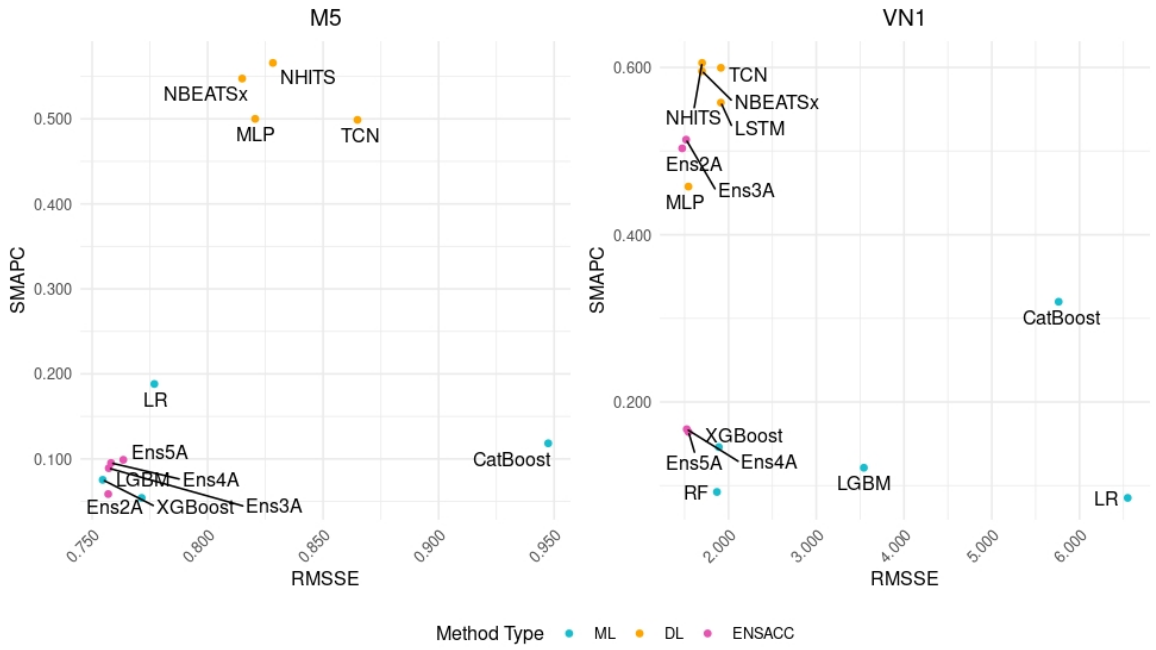

Figure 7: The accuracy-stability trade-off. RMSSE vs sMAPC for the M5 and VN1 datasets.

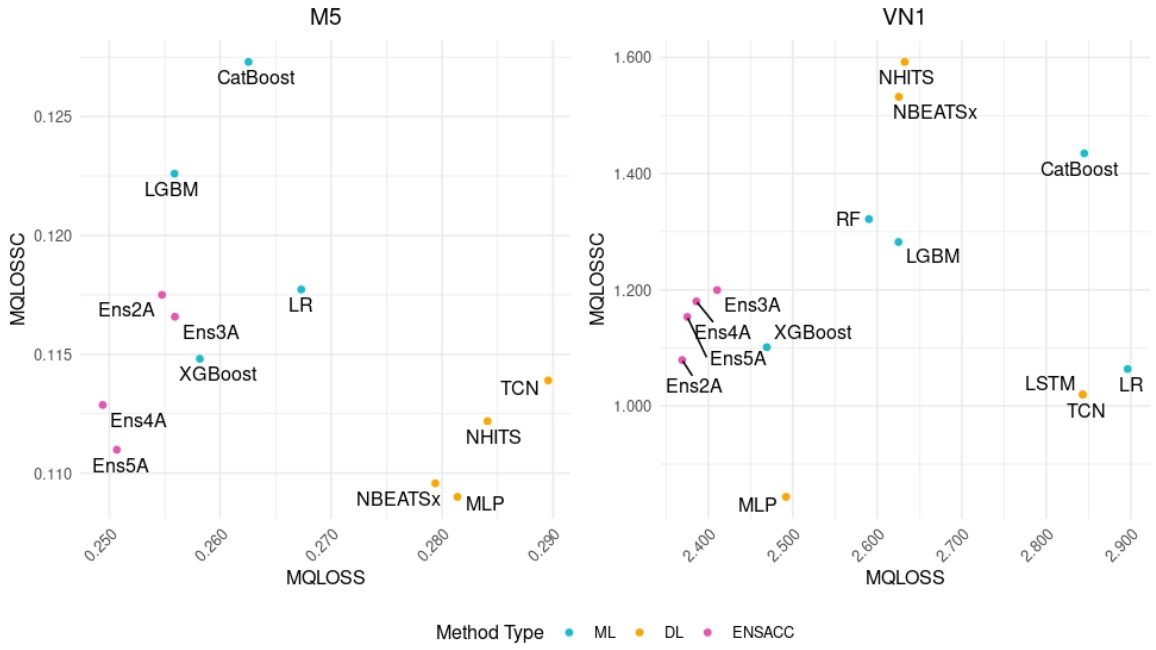

Figure 8: The accuracy-stability trade-off. MQL vs MQC for the M5 and VN1 datasets.

Similar patterns, although more nuanced, are observed in the VN1 dataset. Probabilistic forecast stability, as measured by the newly defined Multi-Quantile Change (MQC), shows a slightly different behavior. While in the M5 dataset forecast stability remains largely unaffected by retraining frequency, in VN1 we observe a convex relationship, indicating that retraining up to every 10 weeks can preserve stability, but beyond that point the quality of probabilistic forecasts begins to deteriorate. These results are statistically validated via Friedman-Nemenyi tests and indicate that continuous retraining is often not the optimal strategy.

Ensembling emerges as a powerful strategy to enhance forecast stability. While combining only the most accurate models offers moderate gains, the most substantial improvements in stability, both point and probabilistic, are achieved by ensembles that incorporate a diverse mix of forecasting algorithms. This effect is especially pronounced in the VN1 dataset, where larger and more heterogeneous ensembles significantly reduce forecast volatility across retraining scenarios. Notably, ensemble models not only improve stability metrics but also smooth the variation in stability across retraining frequencies, further reinforcing their role as a stabilizing force in global forecasting systems. These results suggest that model diversity should be prioritized over raw model performance when building ensembles.

The joint analysis of retraining and ensembling reveals that the two approaches act as complementary levers for stabilizing forecasts. Less frequent retraining reduces abrupt changes in model behavior across forecasting origins, while ensembling mitigates the instability arising from any single model's sensitivity to input changes. When combined, these strategies offer a robust forecasting setup that minimizes the instability of both point estimates and prediction intervals without sacrificing predictive accuracy. Figures comparing the accuracy–stability trade-off clearly show that ensembles lie closer to the optimality, offering an optimal balance between these competing objectives.

From a practical standpoint, these findings have several implications. First, businesses that rely on large-scale forecasting systems, such as those in retail, inventory management, and supply chain planning, can adopt lower retraining frequencies not only to conserve computational resources but also to increase operational stability, without compromising on forecast quality. Second, incorporating diverse ensembles into global forecasting frameworks provides an effective hedge against model instability, thereby increasing the reliability of forecasts used in decision-making

processes. Finally, these results support a shift in emphasis from pure accuracy to stability-aware model selection and evaluation, in those contexts that are more sensitive to forecast reviews.

Nevertheless, some limitations of the present study should be acknowledged. While the analysis covers a broad range of global models and two state-of-the-art retail forecasting datasets, the generalizability of the findings to other domains or data types (e.g., macroeconomic indicators, financial series, or healthcare time series) remains an open question. Additionally, the study focuses solely on full retraining schemes and does not explore incremental or online learning strategies, which may offer a different trade-off between accuracy and stability. Future research could extend this work by incorporating such adaptive methods, assessing alternative ensembling techniques (e.g., weighted or stacking ensembles), and introducing business-oriented stability metrics that link forecast instability directly to operational cost or risk. Further theoretical development of stability measures, particularly for probabilistic forecasts, may also contribute to a deeper understanding of how models behave across time and under varying data regimes.

In conclusion, this work provides strong empirical evidence that frequent retraining is not necessary for maintaining stability in global forecasting models, and that ensembling, particularly when built on diverse model pools, is a highly effective strategy for achieving both stable and accurate forecasts. These findings contribute to a growing body of research promoting sustainable and robust forecasting practices in modern data-intensive environments.

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.

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In T. Gedeon, K. W. Wong, & M. Lee (Eds.), *Neural Information Processing* (pp. 462–474). Cham: Springer International Publishing.

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.

Buonanno, A., Caliano, M., Pontecorvo, A., Sforza, G., Valenti, M., & Graditi, G. (2022). Global vs. local models for short-term electricity demand prediction in a residential/lodging scenario. *Energies*, *15*. URL: https://www.mdpi.com/1996-1073/15/6/2037. doi:10.3390/en15062037.

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.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics. URL: https://aclanthology.org/D14-1179/. doi:10.3115/v1/D14-1179.

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., & Goodwin, P. (2021). Stability in the inefficient use of forecasting systems: A case study in a supply chain company. *International Journal of Forecasting*, *37*, 1031–1046. URL: https://www.sciencedirect.com/science/article/pii/S0169207020301801. doi:https://doi.org/10.1016/j.ijforecast.2020.11.004.

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.

Gaweł, B., & Paliński, A. (2024). Global and local approaches for forecasting of long-term natural gas consumption in poland based on hierarchical short time series. *Energies*, *17*. URL: https://www.mdpi.com/1996-1073/17/2/347. doi:10.3390/en17020347.

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., Bergmeir, C., Erkin Baz, Z., Zhu, C., Song, Z., García, S., & Benavides, D. (2025). On forecast stability. *International Journal of Forecasting*, . URL: https://www.sciencedirect.com/science/article/pii/S0169207025000068. doi:https://doi.org/10.1016/j.ijforecast.2025.01.006.

Godahewa, R., Webb, G. I., Schmidt, D., & Bergmeir, C. (2023). Setar-tree: a novel and accurate tree algorithm for global time series forecasting. *Mach. Learn.*, *112*, 2555–2591. URL: https://doi.org/10.1007/s10994-023-06316-x. doi:10.1007/s10994-023-06316-x.

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-.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Groß, M., & Hans, L. (2024). Leveraging potentials of local and global models for water demand forecasting. *Engineering Proceedings*, *69*. URL: https://www.mdpi.com/2673-4591/69/1/129. doi:10.3390/engproc2024069129.

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.

Hewamalage, H., Bergmeir, C., & Bandara, K. (2022). Global models for time series forecasting: A simulation study. *Pattern Recognition*, *124*, 108441. URL: https://www.sciencedirect.com/science/article/pii/S0031320321006178. doi:https://doi.org/10.1016/j.patcog.2021.108441.

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.

Ibañez, S. C., & Monterola, C. P. (2023). A global forecasting approach to large-scale crop production prediction with time series transformers. *Agriculture*, *13*. URL: https://www.mdpi.com/2077-0472/13/9/1855. doi:10.3390/agriculture13091855.

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.

Juan R Trapero, N. K., & Fildes, R. (2015). On the identification of sales forecasting models in the presence of promotions. *Journal of the Operational Research Society*, *66*, 299–307. URL: https://doi.org/10.1057/jors.2013.174. doi:10.1057/jors.2013.174. arXiv:https://doi.org/10.1057/jors.2013.174.

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. (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.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science*, *43*, 546–558. URL: http://www.jstor.org/stable/2634565.

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.

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.

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.

Rajapaksha, D., Bergmeir, C., & Hyndman, R. J. (2023). Lomef: A framework to produce local explanations for global model time series forecasts. *International Journal of Forecasting*, *39*, 1424–1447. URL: https://www.sciencedirect.com/science/article/pii/S0169207022000978. doi:https://doi.org/10.1016/j.ijforecast.2022.06.006.

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.

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.

Sen, R., Yu, H.-F., & Dhillon, I. S. (2019). Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. volume 32. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/3a0844cee4fcf57de0c71e9ad3035478-Paper.pdf.

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.

Spiliotis, E., Makridakis, S., Semenoglou, A., & Assimakopoulos, V. (2022). Comparison of statistical and machine learning methods for daily sku demand forecasting. *Operational Research*, *22*, 3037–3061. doi:10.1007/s12351-020-00605-2. Publisher Copyright: © 2020, Springer-Verlag GmbH Germany, part of Springer Nature.

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. The M3- Competition.

Van Belle, J., Crevits, R., & Verbeke, W. (2023). Improving forecast stability using deep learning. *International Journal of Forecasting*, *39*, 1333–1350. URL: https://www.sciencedirect.com/science/article/pii/S016920702200098X. doi:https://doi.org/10.1016/j.ijforecast.2022.06.007.

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. (2025a). The cost of ensembling: is it always worth combining? URL: https://arxiv.org/abs/2506.04677. arXiv:2506.04677.

Zanotti, M. (2025b). 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. Table 4 shows the composition of ensembles across the different datasets.

Table 4: Composition of ensembles for the M5 and VN1 datasets.

Dataset	**Ensemble**	**Models Included**
M5	ENS2A	XGBoost, LGBM
	ENS3A	XGBoost, LGBM, LR
	ENS4A	XGBoost, LGBM, LR, NBEATSx
	ENS5A	XGBoost, LGBM, LR, NBEATSx, MLP
VN1	ENS2A	MLP, NBEATSx
	ENS3A	MLP, NBEATSx, NHITS
	ENS4A	MLP, NBEATSx, NHITS, RF
	ENS5A	MLP, NBEATSx, NHITS, RF, XGBoost

The Tables 5, 6, 8, and 7 show the forecast instability of the different models and ensembles along the examined retrain scenarios for the M5 daily dataset.

The Tables 9, 10, 12, and 11 show the forecast instability of the different models and ensembles along the examined retrain scenarios for the VN1 weekly dataset.

Figures 9, and 10 show the results of the Friedman-Nemenyi test in the context of probabilistic instability for the M5 and VN1 datasets.

Figures 11, 12, 13, and 14 show the results of the Friedman-Nemenyi test in the context of both point and probabilistic instability for the different ensemble models across retrain scenarios and for each dataset.

Method	7	14	21	30	60	90	120	150	180	364
LR	0.188	0.188	0.188	0.189	0.189	0.190	0.190	0.192	0.193	0.196
XGBoost	0.075	0.074	0.073	0.073	0.072	0.072	0.072	0.072	0.073	0.073
LGBM	0.054	0.053	0.053	0.053	0.052	0.052	0.052	0.052	0.052	0.052
CatBoost	0.118	0.043	0.059	0.037	0.029	0.020	0.018	0.018	0.018	0.019
MLP	0.500	0.511	0.506	0.520	0.513	0.524	0.510	0.525	0.475	0.553
TCN	0.499	0.209	0.174	0.086	0.055	0.043	0.031	0.017	0.017	0.003
NBEATSx	0.547	0.541	0.527	0.561	0.528	0.536	0.523	0.536	0.515	0.412
NHITS	0.566	0.548	0.558	0.534	0.565	0.537	0.533	0.520	0.512	0.564

Table 5: M5 sMAPC values for each method and retrain scenario combination.

Method	7	14	21	30	60	90	120	150	180	364
Ens2A	0.059	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058
Ens3A	0.089	0.089	0.089	0.089	0.089	0.089	0.090	0.090	0.091	0.092
Ens4A	0.095	0.094	0.094	0.097	0.095	0.095	0.095	0.095	0.096	0.097
Ens5A	0.099	0.098	0.098	0.101	0.099	0.099	0.099	0.099	0.100	0.101

Table 6: M5 sMAPC values for each ensemble and retrain scenario combination.

Method	7	14	21	30	60	90	120	150	180	364
LR	0.118	0.118	0.118	0.119	0.119	0.119	0.119	0.119	0.119	0.119
XGBoost	0.115	0.115	0.115	0.116	0.115	0.115	0.115	0.116	0.115	0.114
LGBM	0.123	0.122	0.123	0.124	0.123	0.123	0.123	0.124	0.123	0.122
CatBoost	0.127	0.144	0.123	0.144	0.144	0.144	0.144	0.142	0.141	0.139
MLP	0.109	0.109	0.110	0.110	0.111	0.111	0.112	0.111	0.110	0.110
TCN	0.114	0.114	0.114	0.113	0.113	0.112	0.113	0.112	0.110	0.108
NBEATSx	0.110	0.109	0.110	0.115	0.110	0.112	0.111	0.111	0.129	0.169
NHITS	0.112	0.111	0.122	0.112	0.114	0.114	0.115	0.135	0.114	0.115

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

Method	7	14	21	30	60	90	120	150	180	364
Ens2A	0.117	0.117	0.117	0.118	0.118	0.118	0.118	0.119	0.118	0.117
Ens3A	0.117	0.117	0.116	0.118	0.118	0.118	0.117	0.118	0.117	0.117
Ens4A	0.113	0.113	0.113	0.114	0.114	0.113	0.113	0.114	0.113	0.113
Ens5A	0.111	0.111	0.111	0.112	0.111	0.111	0.111	0.111	0.110	0.110

Table 8: M5 MQC values for each ensemble and retrain scenario combination.

Method	1	2	3	4	6	8	10	13	26	52
LR	0.085	0.084	0.085	0.082	0.083	0.086	0.086	0.085	0.084	0.082
RF	0.092	0.090	0.089	0.089	0.088	0.088	0.089	0.089	0.089	0.090
XGBoost	0.146	0.129	0.126	0.120	0.118	0.115	0.114	0.119	0.117	0.113
LGBM	0.121	0.110	0.111	0.106	0.104	0.105	0.105	0.104	0.104	0.105
CatBoost	0.320	0.198	0.186	0.152	0.134	0.139	0.115	0.117	0.100	0.095
MLP	0.458	0.469	0.468	0.475	0.443	0.483	0.438	0.490	0.469	0.452
LSTM	0.558	0.301	0.203	0.126	0.101	0.044	0.055	0.048	0.032	0.013
TCN	0.600	0.260	0.226	0.177	0.120	0.091	0.052	0.056	0.020	0.013
NBEATSx	0.596	0.594	0.596	0.577	0.577	0.581	0.573	0.580	0.571	0.550
NHITS	0.605	0.589	0.583	0.572	0.533	0.541	0.552	0.528	0.573	0.580

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

Method	7	14	21	30	60	90	120	150	180	364
Ens2A	0.503	0.515	0.505	0.495	0.499	0.504	0.487	0.485	0.495	0.509
Ens3A	0.514	0.526	0.526	0.509	0.509	0.514	0.510	0.512	0.511	0.512
Ens4A	0.167	0.170	0.162	0.153	0.145	0.137	0.148	0.134	0.135	0.145
Ens5A	0.164	0.162	0.154	0.145	0.138	0.131	0.138	0.129	0.129	0.135

Table 10: VN1 sMAPC values for each ensemble and retrain scenario combination.

Method	1	2	3	4	6	8	10	13	26	52
LR	1.064	1.067	1.063	1.061	1.070	1.079	1.103	1.097	1.127	1.163
RF	1.322	1.329	1.328	1.341	1.353	1.366	1.403	1.392	1.439	1.504
XGBoost	1.101	1.072	1.080	1.079	1.083	1.097	1.084	1.135	1.175	1.254
LGBM	1.282	1.179	1.235	1.161	1.165	1.173	1.199	1.277	1.266	1.379
CatBoost	1.435	1.281	1.244	1.159	1.150	1.166	1.151	1.225	1.207	1.246
MLP	0.844	0.876	0.887	0.883	0.889	0.891	0.876	0.933	0.903	0.969
LSTM	1.020	1.015	1.008	1.004	1.005	0.993	0.986	0.987	0.972	0.930
TCN	1.020	1.015	1.008	1.004	1.005	0.993	0.986	0.987	0.972	0.930
NBEATSx	1.532	1.519	1.467	1.253	1.167	1.210	1.196	1.069	1.084	1.147
NHITS	1.592	1.505	1.412	1.254	1.212	0.998	1.174	1.028	1.082	1.145

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

Method	7	14	21	30	60	90	120	150	180	364
Ens2A	1.079	1.114	1.062	0.974	0.949	0.982	0.962	0.921	0.944	1.022
Ens3A	1.200	1.192	1.148	1.024	0.990	0.973	0.995	1.005	0.970	1.034
Ens4A	1.180	1.166	1.122	1.036	1.021	1.001	1.028	0.988	1.017	1.080
Ens5A	1.153	1.137	1.107	1.033	1.022	1.000	1.031	1.006	1.029	1.091

Table 12: VN1 MQC values for each ensemble and retrain scenario combination.

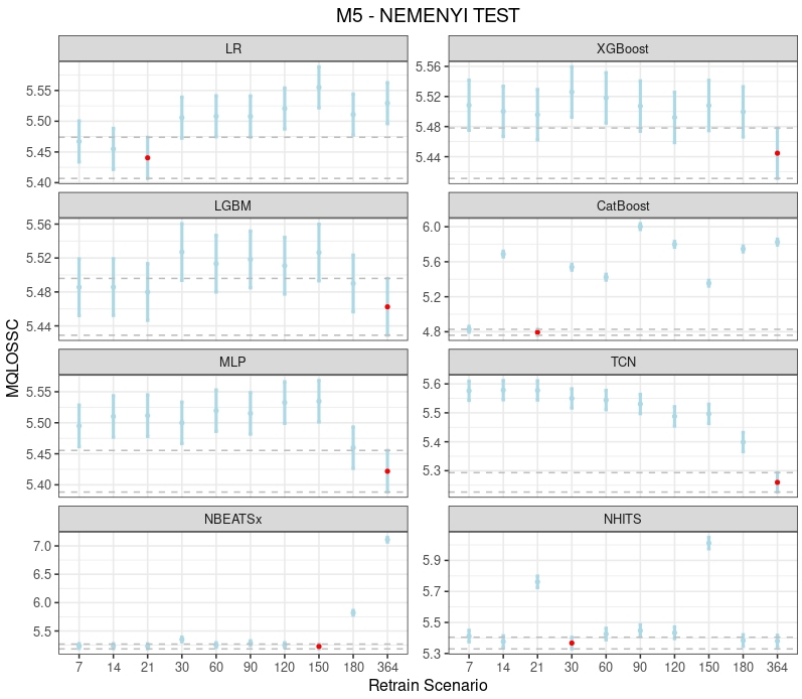

Figure 9: M5 Friedman-Nemenyi test results based on MQC.

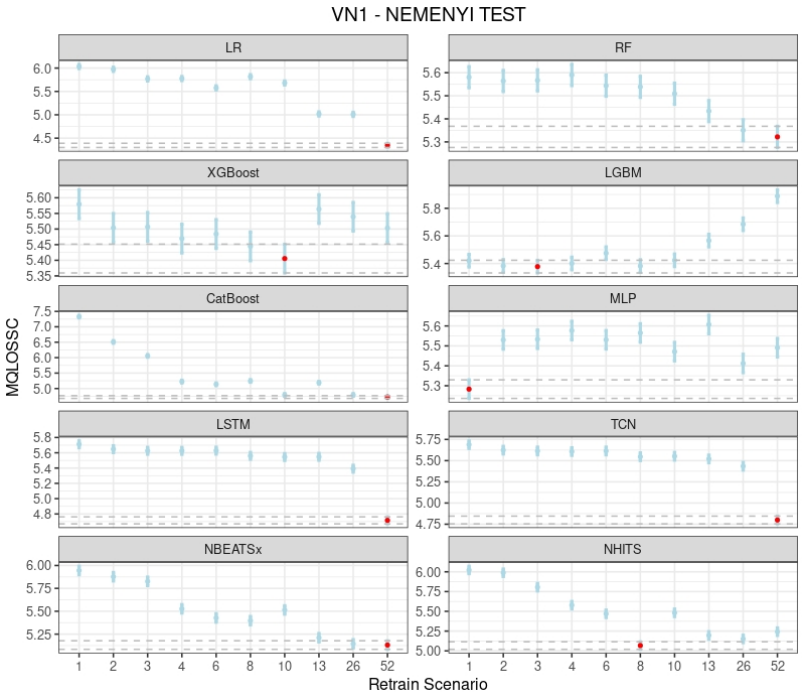

Figure 10: VN1 Friedman-Nemenyi test results based on MQC.

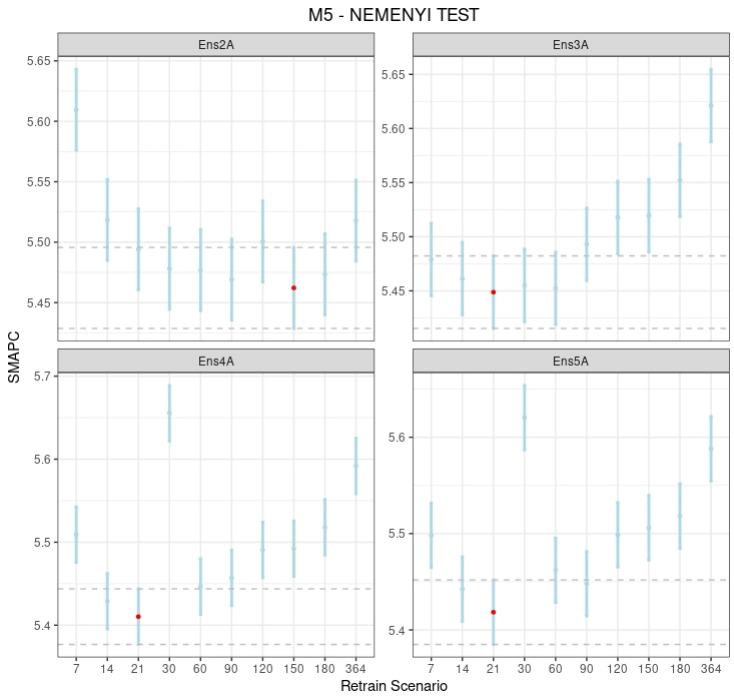

Figure 11: M5 Friedman-Nemenyi test results on ensembles based on sMAPC.

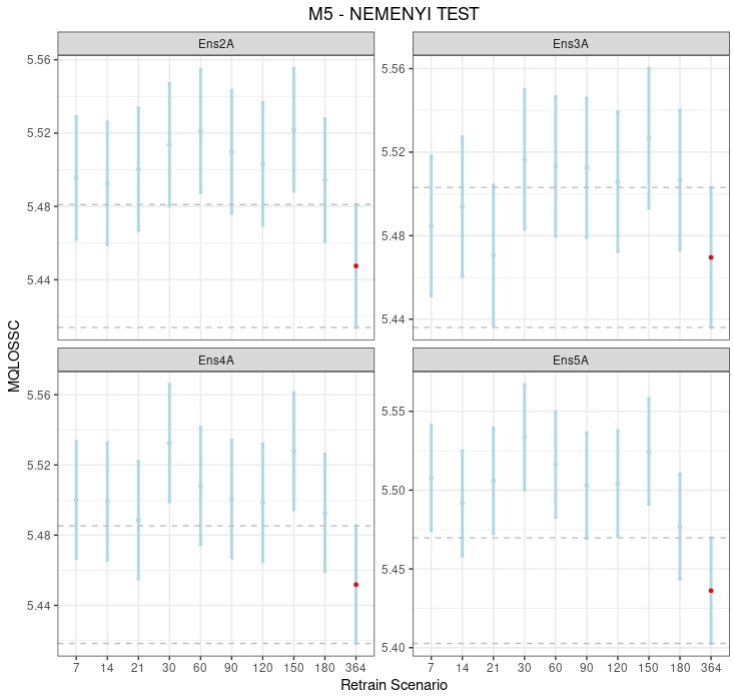

Figure 12: M5 Friedman-Nemenyi test results on ensembles based on MQC.

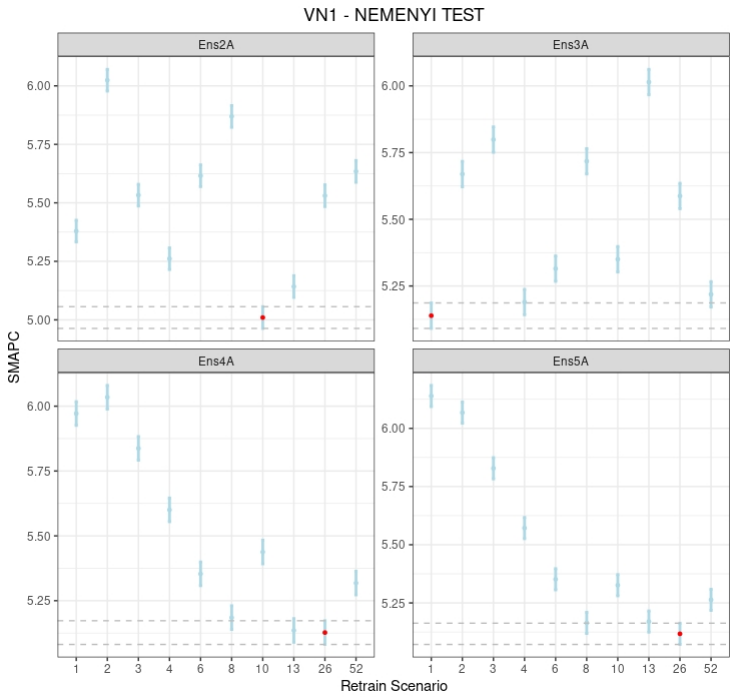

Figure 13: VN1 Friedman-Nemenyi test results on ensembles based on sMAPC.

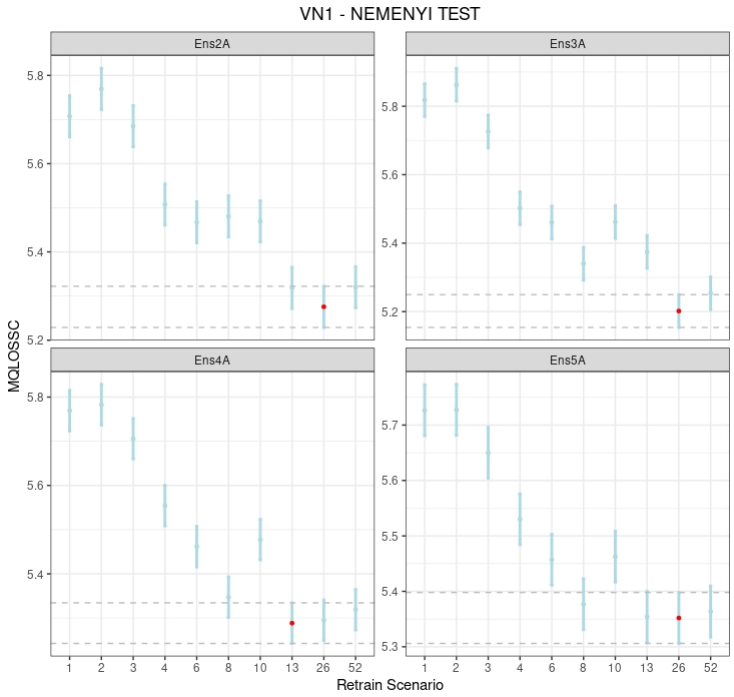

Figure 14: VN1 Friedman-Nemenyi test results on ensembles based on MQC.