Forking-Sequences

Anonymous authors
Paper under double-blind review

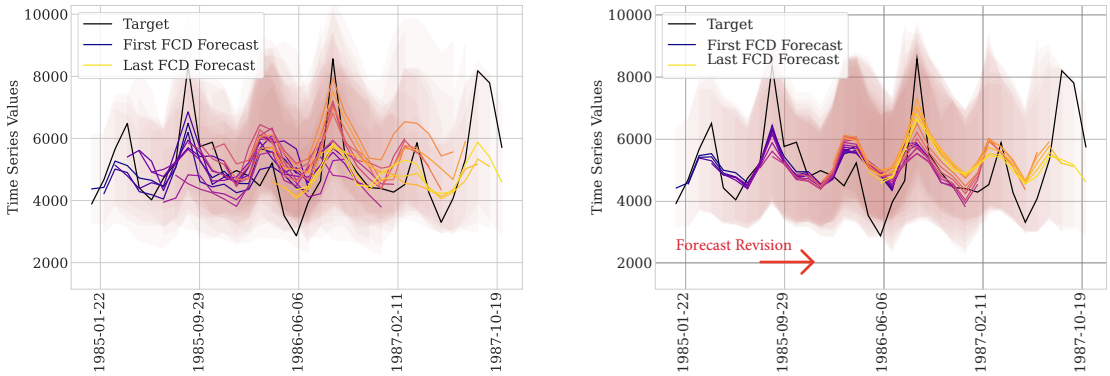
Abstract

While accuracy is a critical requirement for time series forecasting, an equally important (yet often overlooked) desideratum is forecast stability across forecast creation dates (FCDs). Even highly accurate models can produce erratic revisions between FCDs, undermining stakeholder trust and disrupting downstream decision-making. To improve forecast stability of such revisions, several state-of-the-art models including `MQCNN`, `MQT`, and `SPADE` employ a powerful yet underexplored neural network architectural design known as *forking-sequences*. This architectural design jointly encodes and decodes the entire time series across all FCDs, producing an entire multi-horizon forecast grid in a single forward pass. This approach contrasts with conventional statistical and neural forecasting methods that process FCDs independently, generating only a single multi-horizon forecast per forward pass. In this work, we formalize the forking-sequences design and motivate its broader adoption by introducing a metric for quantifying excess volatility in forecast revisions and by providing theoretical and empirical analysis. We theoretically motivate three key benefits of forking-sequences: (i) increased forecast stability through ensembling; (ii) gradient variance reduction, leading to more stable and consistent training steps; and (iii) improved computational efficiency during inference. We validate the benefits of forking-sequences compared to baseline window-sampling on the M-series benchmark, using 16 datasets from the M1, M3, M4, and Tourism competitions. We observe median accuracy improvements across datasets of 29.7%, 46.2%, 49.3%, 28.6%, 24.7%, and 6.4% for MLP, RNN, LSTM, CNN, Transformer, and `State Space`-based architectures, respectively. We then show that forecast ensembling during inference can improve median forecast stability by 10.8%, 13.2%, 13.0%, 10.9%, 10.2%, and 11.2% for these respective models trained with forking-sequences, while maintaining accuracy.

1 Introduction

Forecasting plays an important role in a wide range of domains, including energy systems (Hong et al., 2016), finance (Maung and Swanson, 2025), economics (Elliott and Timmermann, 2008), product supply chain (Wen et al., 2017; Eisenach et al., 2021; Wolff et al., 2024), and healthcare analytics (Nikolopoulos et al., 2021; Potosnak et al., 2025). It typically serves as a foundation for predicting uncertain events, enabling informed decision-making in various scenarios, including demand anticipation, production scheduling, and resource allocation. To meet diverse operational needs and reduce the accumulation of forecast errors, multi-horizon forecasting tools have become the default, offering improved support for short, medium, and long-term planning. However, as these systems operate over time, they generate multiple, overlapping forecasts for each target date. This sequence of revisions raises considerations regarding the consistency of these forecasts, which we refer to as *forecast stability*.

The balance between forecast accuracy (of a single prediction) and forecast stability (as predictions are updated) presents a new challenge: while new temporal information should naturally refine predictions, excessive fluctuations in forecasts can complicate operational planning and decision-making. For example, electrical grid operators rely on load forecasts to determine the power supply. In the event of an upcoming intense heat wave, forecast revisions from 45 to 65 GW reflect a useful revision that enables preparations such as activating reserve plants. Conversely, if forecasts are overly stable (i.e., not revised), it can result in inadequate power supply where insufficient power when needed can cause preventable blackouts affecting



(a) Original Forecasts **(b)** Forking-sequence Ensembled Forecasts

Figure 1: Forecast comparison for a series from the `M1` dataset (Makridakis et al., 1982). (a) Forecasts generated without the forking-sequences ensemble. (b) Forecasts with the forking-sequences ensemble applied. Augmenting the model with the forking-sequences ensemble significantly reduces forecast variability across FCDs, resulting in more stable and consistent forecast distributions. We marked the direction of forecast revisions in red. The lines show P50 (median) forecasts across different FCDs. By reusing encoder computations, forking-sequences enables computationally efficient ensembling with negligible additional computational cost.

large populations. Forecast stability raises two research questions: 1) how can we measure forecast stability in a way that captures benign revisions while identifying problematic volatility?; and 2) are there neural forecasting architectural designs that can improve forecast stability, with minimal impact or improvements on forecast accuracy?

In this paper, we formally introduce *forking-sequences*, an encoder-agnostic network architectural design that improves both forecast accuracy and forecast stability. The forking-sequences design offers two distinct use paradigms: during training, it augments loss computation with additional samples, leading to reduced gradient variance and better preservation of gradients for early time steps in recurrent architectures; and during inference, it enables efficient ensemble forecasting across multiple FCDs, improving forecast stability.

(i) **Forking-Sequences Gradient Improvements.** We introduce forking-sequences and formally analyze its gradient variance reduction properties and its ability to ameliorate vanishing gradients. We show that, under mild temporal correlation assumptions, the gradient variance decreases linearly with the number of FCDs ($\mathcal{O}(1/T)$, echoing results from the weak law of large numbers). These gradient improvements lead to more stable training steps, which in turn accelerates convergence in optimization, in both convex and nonconvex loss landscapes. We empirically demonstrate that these optimization benefits also translate into improved generalization.

(ii) **Forking-Sequences Computational Efficiency.** We show that forking-sequences enable an order of magnitude improved computational regime for encoder-decoder architectures by supporting efficient inference across FCDs in a single forward pass. Unlike standard approaches that recompute the encoder for each FCD, forking-sequences reuse encoder outputs across all FCDs, significantly reducing redundant computation. This results in orders-of-magnitude gains in inference efficiency, from quadratic to linear complexity ($\mathcal{O}(T^2)$ vs. $\mathcal{O}(T)$).

(iii) **Forking-Sequences Ensembling.** By generating multi-horizon forecasts for all FCDs in a single forward pass, forking-sequences naturally produce overlapping forecasts for each target date, which we demonstrate can be ensembled to improve forecast stability with no impact on accuracy.

(iv) **Novel Forecast Stability Metrics.** We introduce two novel metrics for assessing forecast stability across FCDs: *Scaled Excess Volatility* (sEV) and *Symmetric Quantile Percentage Change* (sQPC). Improving upon sQPC, sEV assesses probabilistic forecasts and distinguishes accuracy-improving revisions from detrimental ones, going beyond simply measuring the magnitude of forecast changes.

(v) **Forking-Sequences Empirical Validation.** We evaluate both accuracy and forecast stability across FCDs, comparing forking-sequences against the standard window-sampling approach on the M-series benchmark: 16 large-scale datasets from the Tourism competition (Athanasopoulos et al., 2011) and the M-forecast competitions (Makridakis et al., 1982; Makridakis and Hibon, 2000; Makridakis et al., 2020). We demonstrate that for `MLP`, `RNN`, `LSTM`, `CNN`, and `Transformer`-based architectures, median forecast accuracy across datasets improves by up to 49.3%.

The remainder of this paper is organized as follows: Section 2 explores related work and relevant literature. Section 3 introduces the architectural design of forking-sequences and its properties. Section 4 presents the main empirical results. Finally, Section 5 discusses our key findings, and outlines future research directions.

2 Related Work

Ideas similar to forking-sequences have been present in the forecasting literature for decades, despite not inspiring neural network architectural design or training. Most notably, Hart (1994) introduced time-series cross-validation as a model selection technique that accounts for temporal dynamics and data dependencies, addressing the limitations of classical cross-validation; this approach assumes independent FCD observations, and forking-sequences bears several similarities with it (Bergmeir and Benítez, 2012). (Figure 7 provides a depiction of the temporal cross-validation evaluation.)

More recently, forking-sequences has informed the design of MQForecaster neural networks for industrial applications, with notable examples including `MQCNN`, `MQT`, and `SPADE` (Wen et al., 2017; Olivares et al., 2023; Eisenach et al., 2021; Wolff et al., 2024). However, this prior work largely adopted forking-sequences as a practical heuristic, without providing formal presentation, theoretical justification, or systematic empirical validation of its benefits. In some cases, its theoretical motivation was attributed to the martingale properties of forecast revisions (Chen et al., 2022; Foster and Stine, 2021), rather than recognizing its connections with cross-validation and its simpler and more natural role in reducing variance during training and inference.

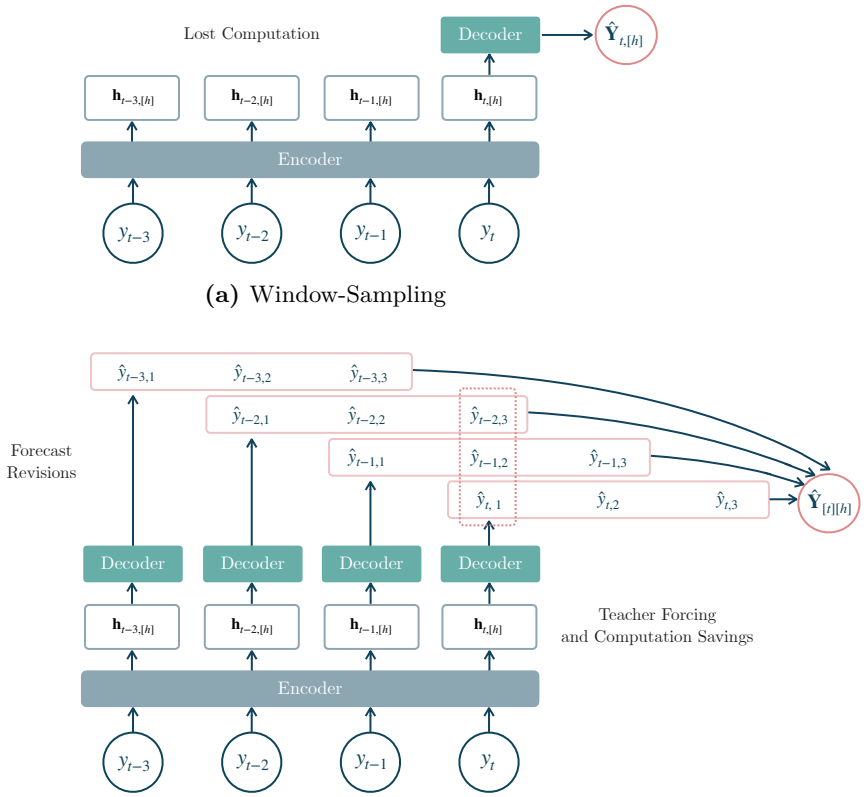
Beyond forecasting; in the context of healthcare care and natural language processing (NLP), the concept of *target replication* has been considered to train LSTM time series classifiers for medical diagnosis (Lipton et al., 2016); and a similar scheme to train text document classifiers has also been considered (Dai and Le, 2015). We distinguish target replication and the standard *teacher forcing* technique (Williams and Zipser, 1989).

The first method for training a recurrent network in a fully online, continuous manner was the Real-Time Recurrent Learning (RTRL) algorithm, which updates the network's weights at every time step rather than waiting for the full sequence to finish (Williams and Zipser, 1989). Williams and Zipser later coined the term *teacher forcing* to describe the practice of feeding ground-truth inputs to the recurrent network during training, instead of using its own predictions. This approach improved the training stability of models trained with backpropagation through time by reducing the accumulation of prediction errors across time steps.

The *teacher forcing* method, most notably through the next-token prediction objective (Bengio et al., 2003), is an autoregressive learning strategy which trains models to predict the next token given all previous ones, and it laid the foundation for modern language models. It was later popularized at scale by generative pre-trained transformers such as GPT-1 and GPT-3 (Radford et al., 2018; Brown et al., 2020).

A key difference between the standard *teacher-forcing* technique used in NLP autoregressive models and the *forking-sequences* approach used in neural forecasting is the implicit target replication that arises in multi-horizon forecasts. Whereas NLP models are typically trained to predict only the next token, forecasting models must predict longer future trajectories, generating multiple future steps for each FCD. This temporal overlap in targets is precisely what exacerbates the need for forecast stability.

Neural forecasting has focused almost exclusively on improving predictive accuracy without concern for forecast stability. The few works that address stability typically rely on specialized loss (Belle et al., 2023), dynamic loss weighting algorithms (Caljon et al., 2025), or post-processing techniques such as linear interpolation (Belle et al., 2023) that require additional parameters and processing steps. In contrast, *forking-sequences* offers a much simpler way to efficiently ensemble forecasts for overlapping target dates generated across multiple FCDs in a single forward pass, without modifying the loss or neural architecture.



(a) Window-Sampling

(b) Forking-Sequences

Figure 2: (a) A *window-sampling* model, where its encoder outputs a single hidden state $\mathbf{h}_{t,[h]}$ (blue rectangles) for each input sequence $\mathbf{y}_{[t]}$. Blue rectangles indicate encoded series, green rectangles denote the decoders, and the red circle denotes the prediction target. (b) A *forking-sequences* model, where its encoder outputs hidden states $\mathbf{h}_{[t][h]}$, reusing computations from $\mathbf{h}_{[t-1][h]}$. During training, forking-sequences collects multi-horizon errors from all intermediate FCDs. Additionally, decoded forecasts across FCDs can be easily ensembled (dotted lines), saving most of the encoding computational costs and improving forecasts revision stability.

3 Methodology

3.1 Multi-Horizon Forecasting

We start by introducing the general multi-horizon forecasting task notation. Let the FCDs collection be $[t] = [1, \ldots, t]$ and the forecast horizon be denoted by $[h] = [1, \ldots, h]$. The target random variables are

$$\text{single target} \quad \mathbf{Y}_{t,[h]} = [Y_{t+1}, \ldots, Y_{t+h}] \in \mathbb{R}^h, \tag{1}$$

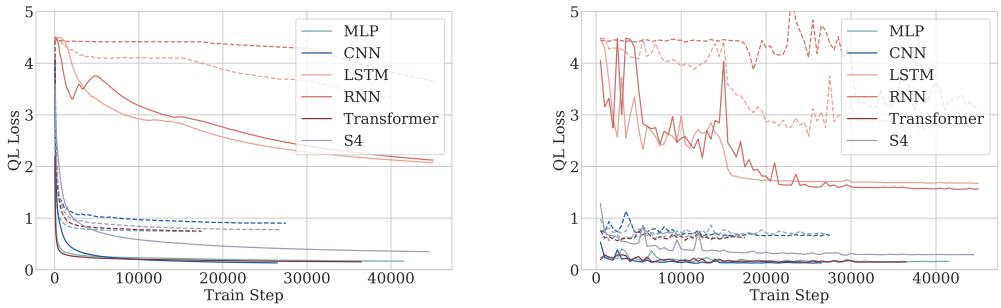
$$\text{and grid target} \quad \mathbf{Y}_{[t][h]} \in \mathbb{R}^{t \times h}. \tag{2}$$

The forecasting task we tackle aims to estimate the following conditional distribution:

$$\mathbb{P}\left(\mathbf{Y}_{t,[h]} \mid \boldsymbol{\theta},\ \mathbf{y}_{[t]}\right), \tag{3}$$

where $\boldsymbol{\theta}$ denotes the model parameters and $\mathbf{y}_{[t]} = [y_1, y_2, \ldots, y_t]$ denotes the autoregressive realizations until t. In our work, we denote forecasts for the target variable $\mathbf{Y}_{t,[h]}$ as $\hat{\mathbf{Y}}_{\mathbf{t},[\mathbf{h}]}$.

Forecast revisions. Our work focuses on multi-horizon forecasting across FCDs ($H > 1$). This setting is essential for studying forecasts' stability. In single-horizon forecasting, each prediction corresponds to a unique future time point, so forecasts never overlap, and revisions cannot occur. In contrast, multi-horizon



(a) Train Set **(b)** Validation Set

Figure 3: MQForecaster model optimization convergence on the M4 hourly data, using either *forking-sequences* (solid) or *window-sampling* (dashed) techniques. Quantile loss on the **(a)** train set and **(b)** validation set as a function of train steps are consistently lower for forking-sequences models. Across architectures forking sequences show validation accuracy improvements. Quantile loss for the train set of other frequency-specific datasets is shown in Fig. 12.

forecasting produces overlapping predictions: the same target value is forecast multiple times at successive FCDs as new information arrives. This naturally forms a sequence of revisions for each target value. Formally,

$$\text{For the target random variable} \quad Y_{t+h}, \quad \hat{\mathbf{Y}}_{t+1,h} \quad \text{is a revision of} \quad \hat{\mathbf{Y}}_{t,h+1}. \tag{4}$$

Parameter Learning. To tackle the multi-horizon forecasting task, we learn neural network parameters $\boldsymbol{\theta}$ by minimizing the empirical expectation of the Multi Quantile Loss (QL; (Koenker and Bassett, 1978)),

$$\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\Theta} \hat{\mathbb{E}} \left[\frac{1}{|Q|} \sum_{[q]\in Q} \mathrm{QL}_q(Y, \hat{Y}^{(q)}(\boldsymbol{\theta})) \right], \tag{5}$$

where QL is the quantile loss for a given quantile level $q \in \mathcal{Q} = \{0.1, 0.2, \ldots, 0.9\}$

$$\mathrm{QL}_q(y, \hat{y}^{(q)}) = q(y - \hat{y}^{(q)})_+ + (1-q)(\hat{y}^{(q)} - y)_+. \tag{6}$$

3.2 Forking-Sequences Architecture Design

Most neural forecasting models use a standard (and intuitive) architectural design known as *window-sampling* (Oreshkin et al., 2020; Challu et al., 2023; Nie et al., 2023; Salinas et al., 2020; Lim et al., 2021; Das et al., 2024; Zhou et al., 2020; Ansari et al., 2024; Woo et al., 2024; Garza and Mergenthaler-Canseco, 2023). The window-sampling approach segments the series into windows of size L; and windows are treated independently across FCDs. This approach is illustrated in Figure 2a, and it is formalized in Equation 7:

$$\mathbf{h}_{t,[h]} = \mathrm{Encoder}\left(\mathbf{Y}_{[t]}\right) \qquad \text{and} \qquad \hat{\mathbf{Y}}_{t,[h]} = \mathrm{Decoder}\left(\mathbf{h}_{t,[h]}\right). \tag{7}$$

In contrast, *forking-sequences* is a neural architectural design that jointly encodes and decodes the time series across all FCDs. Specifically, the model computes a single hidden representation $\mathbf{h}_{[t][h]}$ that is then decoded into forecasts $\hat{\mathbf{y}}_{[t][h]}$ for all FCDs $[1, \ldots, T]$. This approach is illustrated in Figure 2b, and it is formalized in Equation 8:

$$\mathbf{h}_{[t][h]} = \mathrm{Encoder}\left(\mathbf{Y}_{[t]}\right) \qquad \text{and} \qquad \hat{\mathbf{Y}}_{[t][h]} = \mathrm{Decoders}\left(\mathbf{h}_{[t][h]}\right). \tag{8}$$

3.3 Forking-Sequences Gradient Improvements

In this subsection, we analyze how the *forking-sequences* design influences parameter learning. We theoretically motivate the design by showing that leveraging each forecast across multiple FCDs reduces the variance of the

gradient estimator in a manner analogous to the statistical efficiency gains obtained from full autoregressive factorization in *teacher-forcing* (Williams and Zipser, 1989; Bengio et al., 2003). Beyond variance reduction, we demonstrate that forking-sequences provide denser supervision signals, which accelerate learning, improve convergence, and mitigate vanishing-gradient effects across a range of encoder architectures.

During training, the gradients of the model parameters are computed for SGD updates to minimize the loss in Equation 5. The gradient is computed on a B-length minibatch. With a window-sampling architecture, multi-horizon losses are gathered in a batch $\mathcal{B}$ for individual FCDs:

$$\nabla\mathcal{L} = \frac{1}{B \times H} \sum_{b=1}^{B} \sum_{h=1}^{H} \nabla\mathcal{L}\left(y_{b,t,h},\ \hat{y}_{b,t,h}\right). \tag{9}$$

Using a forking-sequences design, multi-horizon losses are gathered across all FCDs:

$$\nabla\mathcal{L}_T = \frac{1}{B \times T \times H} \sum_{b=1}^{B} \sum_{t=1}^{T} \sum_{h=1}^{H} \nabla\mathcal{L}\left(y_{b,t,h},\ \hat{y}_{b,t,h}\right). \tag{10}$$

The proof of the following result is included in Appendix A.

Theorem 1. *(*Forking-Sequences Gradient Variance Reduction*) Consider the forking-sequences gradient from Equation 10. If the sequence of gradient samples T is M-dependent (Shumway and Stoffer, 2017) then the estimator converges in probability to the full gradient, and its variance decreases linearly $\mathcal{O}\left(1/T\right)$.*

Forking-sequences implicitly act as a form of data augmentation, not by perturbing inputs, but by enriching the number of effective training signals extracted from each series. As with *teacher forcing* in autoregressive language models, which turns one text sequence into many next-token supervision events, forking-sequences convert series into a full set of multi-horizon targets. Although traditional window sampling could provide similar enrichment, it would require multiple forward passes through the encoder. In contrast, forking-sequences reuse a single encoded representation, while expanding the number of training losses by a factor of T, multiplying the available supervision at negligible computational cost. This data augmentation has a twofold effect. First, model training with forking-sequences reduces the variance of the stochastic gradient around its mean at a linear rate ($\mathcal{O}(1/T)$), with FCDs; see Fig. 9b in Appendix A. Second, forking-sequences help preserve gradients in early FCDs, thus helping alleviate vanishing gradient issues, as shown in Appendix B.

As shown in Figure 3, a natural consequence of the reduced gradient variance is a faster and more stable convergence of the training procedure. In Appendix C, we verify this effect using linear autoregressive models optimized with convex quantile loss on synthetic data, demonstrating consistent improvements in both convex and non-convex regimes. Practical implications of these results are that models trained with forking-sequences can tolerate larger learning rates, require fewer iterations to reach comparable loss levels, and tend to follow smoother optimization trajectories. This leads to training procedures that are more robust to hyperparameter choices, less sensitive to initialization, and more computationally efficient.

Table 1: Theoretical Encoder Complexity

	MLP	CNN	RNN	Att.	S4
FS	$\mathcal{O}(T)$	$\mathcal{O}(T)$	$\mathcal{O}(T)$	$\mathcal{O}(T^2)$	$\mathcal{O}(T \log T)$
WS restr.	$\mathcal{O}(TL)$	$\mathcal{O}(TL)$	$\mathcal{O}(TL)$	$\mathcal{O}(TL^2)$	$\mathcal{O}(TL \log L)$
WS full	$\mathcal{O}(T^2)$	$\mathcal{O}(T^2)$	$\mathcal{O}(T^2)$	$\mathcal{O}(T^3)$	$\mathcal{O}(T^2 \log T)$

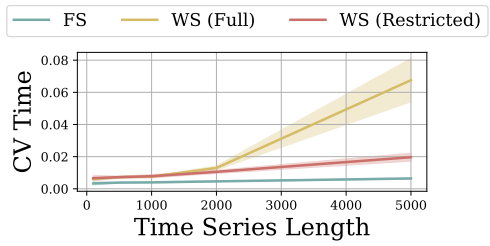
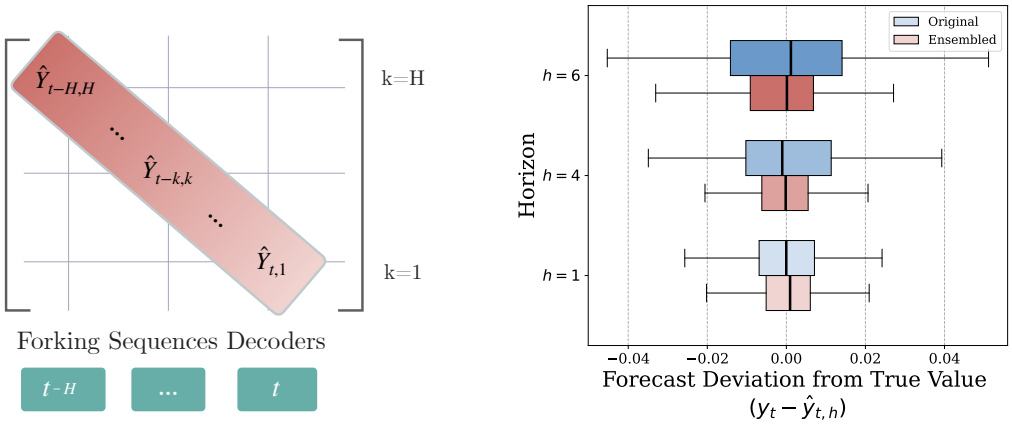


Figure 4: Empirical validation of computational complexity through wall-clock measurements across different inference methods. T represents the time series length, and L denotes the window size in restricted window-sampling.

3.4 Forking-Sequences Computational Complexity

As illustrated in Figure 2, the forking-sequences approach enhances efficiency by reusing the encoder outputs across all forecast creation dates (FCDs), which allows the model to propagate hidden states $\mathbf{h}_{[t][h]}$ from $\mathbf{h}_{[t-1][h]}$). In contrast to standard methods that must recompute the encoder's hidden states separately for each FCD, incurring repeated and redundant encoder passes, this design achieves its efficiency by reusing



(a) Forking-Sequences Ensemble **(b)** Forecast Stability Improvements

Figure 5: a) We adapt forking-sequences during inference to ensemble multiple forecasts of the same future date τ by computing a function (ex., moving average) across predictions generated from previous FCDs. b) Forking-sequences ensembling improves forecasts' stability, reducing the estimators variance with a linear convergence rate analogous to the weak law of large numbers.

intermediate computations within a single forward pass. This
substantially reduces the total number of encoder operations required to generate the full set of multi-horizon forecasts.

Forking-sequences naturally suit encoders like CNNs, LSTMs, RNNs, and Transformers capable of outputting intermediate encoder outputs, while also being adaptable to MLPs through appropriate input reformatting.

We show that this architectural design yields a computational speedup factor of T, where T is the number of FCDs. See the complexity summary in Table 1 and complete calculation in Appendix D. Figure 4 empirically validates these complexity results by measuring the cross-validation inference time across series of different lengths. This experiment compares the computational cost of an MQCNN architecture cross validation inference when using three different variants: the standard full windows-sampling, the more efficient restricted windows-sampling, and our proposed forking-sequences approach.

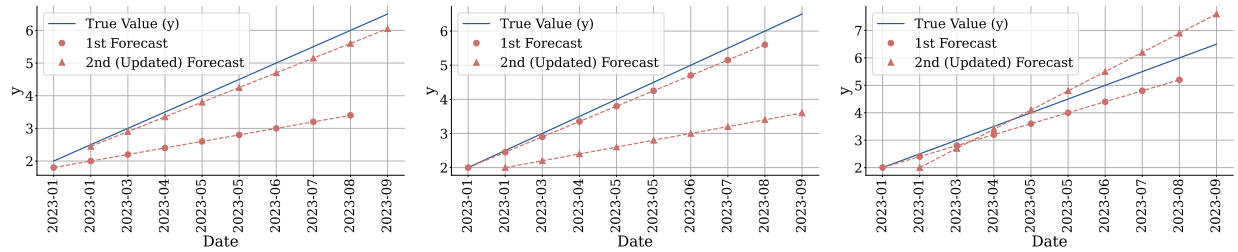
In the case of convolutional models, the forking-sequences scale linearly with the time series length, achieving a complexity of $\mathcal{O}(T)$ by producing forecasts at each FCD while reusing previous computations. In contrast, window-sampling incurs a higher computational cost, scaling quadratically as $\mathcal{O}(T^2)$ in the WS (full) case, or linearly as $\mathcal{O}(TL)$ when restricting the encoder to a window of length L in the WS (restricted) case.

As stated earlier, the gradient variance reduction shown in **Theorem 1** could also be achieved simply by increasing the batch size, but forking-sequences achieve this improvement with significantly lower computational cost, as shown in Figure 4. Specifically, increasing the batch size by a factor of T would result in a computational cost of $\mathcal{O}(BT) = \mathcal{O}(T^2)$, assuming $B = T$. In contrast, forking-sequences achieve a comparable variance reduction with only $\mathcal{O}(T)$ complexity, by reusing computations across FCDs.

3.5 Forking-Sequences Ensembling

The forking-sequences architectural design can be extended to serve as an efficient forecasting ensemble technique to help improve stability of the forecast revisions in Equation 4. By generating predictions for all FCDs in a single forward pass, forking-sequences naturally produce overlapping predictions for each target date t, which can be assembled to improve the stability of the forecast as follows:

$$\widetilde{\mathbf{Y}}_{t,h} = \frac{1}{H} \sum_{k=0}^{H} \widehat{\mathbf{Y}}_{t-k,h+k} \qquad \text{for } t \geq H. \tag{11}$$



(a) Improving Revision **(b)** Deteriorating Revision **(c)** Overshooting Revision

Figure 6: Example penalty behavior of the Excess Volatility (EV) metric. EV distinguishes accuracy-improving revisions from accuracy-degrading ones, assigning no penalty when revisions improve accuracy, while asymmetrically penalizing both deteriorating and overshooting revisions according to their impact on accuracy.

The forecast revisions are illustrated in Figure 5a as the diagonal within the forecast grid $\hat{\mathbf{Y}}_{[t][h]}$, and the temporally overlapped forecasts in Figure 2b.

Although it is tempting to expect a variance-reduction behavior similar to the results of Theorem 1, it is important to recognize that forecast variance naturally increases the further a forecast is from its corresponding observation. As a result, there is an inherent limit to how much ensembling can improve stability: older forecast revisions carry substantially higher uncertainty, whereas more recent revisions are both more accurate and less variable. This makes it desirable for an ensemble to place greater weight on newer forecasts rather than treating all revisions equally. We investigate alternative ensembling strategies that account for this effect in Appendix H. Building on this discussion, ensemble techniques such as Equation 11 can also be applied to window-sampling variants, though at additional computational cost. We detail the impact of ensembling on forecast stability for forking-sequences models in Appendix H.1 and for window-sampling models in Appendix I.

We acknowledge that ensembling can be integrated during both training and inference with forking-sequences, and could be further extended with learnable parameters (as explored in (Eisenach et al., 2021)). We defer the exploration of incorporating forecast ensembling directly into the training phase to future work.

3.6 Excess Volatility

Here we introduce a novel metric for assessing forecast stability across FCDs: *Excess Volatility* (EV). EV is designed to reward accuracy-improving forecast revisions while distinguishing them from harmful volatility—addressing a limitation of measures such as forecast percentage change, which quantify revisions without considering forecast accuracy. Excess Volatility is defined as

$$\mathrm{EV}(y,\ \hat{\mathbf{y}}_1,\ \hat{\mathbf{y}}_2) = \mathrm{QL}(\hat{\mathbf{y}}_2, \hat{\mathbf{y}}_1) - (\mathrm{QL}(y, \hat{\mathbf{y}}_1) - \mathrm{QL}(y, \hat{\mathbf{y}}_2)). \tag{12}$$

For two consecutive forecasts $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$, the EV metric in Equation 12 evaluates the quality of forecast revisions by asymmetrically rewarding revisions that improve accuracy and penalizing those that degrade it. Unlike percentage-change metrics such as sQPC, EV incorporates ground-truth values to judge whether a revision is beneficial or harmful. Metrics that allow one to distinguish accuracy-improving revisions from detrimental ones are more informative and helpful improving forecasting systems (Chen et al., 2022).

The proof of the following result is included in Appendix E.

Theorem 2. (*Excess Volatility Metric Properties*) Taking the ground truth y and two consecutive forecasts $\hat{\mathbf{y}}_1$, $\hat{\mathbf{y}}_2$, the EV metric has the following properties:

- **Positivity:** $\mathrm{EV}(y,\ \hat{\mathbf{y}}_1,\ \hat{\mathbf{y}}_2) \geq 0$
- **Zero-Penalty for Improving Revisions:** When the revision from $\hat{\mathbf{y}}_1$ to $\hat{\mathbf{y}}_2$ represents a direct proportional improvement and $\hat{\mathbf{y}}_2$ lies on the linear path from y to $\hat{\mathbf{y}}_1$, then $\mathrm{EV}(y,\ \hat{\mathbf{y}}_1,\ \hat{\mathbf{y}}_2) = 0$.

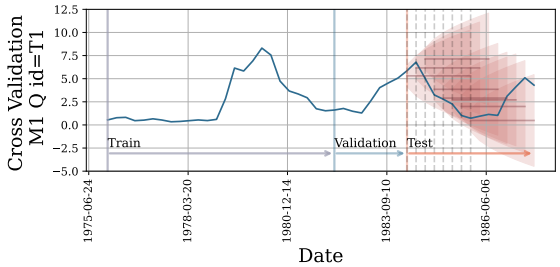


Figure 7: Example time series from the M1 competition dataset. The train and validation sets consist of all observations preceding the first dotted line. The cross validation test set is comprised of a set of FCDs with lengths defined in Table 6. To leverage all the FCDs within train for model estimation while preventing temporal leakage between splits, we apply masking across training and test sets.

- **Maximum penalty for Deteriorating Revisions:** When $\hat{\mathbf{y}}_1$ lies on the direct path between y and $\hat{\mathbf{y}}_2$; that is, when the revision from $\hat{\mathbf{y}}_1$ to $\hat{\mathbf{y}}_2$ moves linearly away from the truth, then $\mathrm{EV}(y,\ \hat{\mathbf{y}}_1,\ \hat{\mathbf{y}}_2) = \mathrm{QL}(y, \hat{\mathbf{y}}_2) - \mathrm{QL}(y, \hat{\mathbf{y}}_1)$. In this fully misdirected case, EV assigns the maximum penalty, equal to the entire accuracy degradation.

- **Overshoot-Revision Penalty Property:** When the ground truth y lies on the line segment between $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$; that is, when $\hat{\mathbf{y}}_2$ revises $\hat{\mathbf{y}}_1$ in the correct direction but overshoots, then $\mathrm{EV} = \mathrm{QL}(y, \hat{\mathbf{y}}_2)$. In this excess-revision case, EV penalizes only the overshoot.

Example penalty behavior of the Excess Volatility (EV) metric is shown in Fig. 6.

4 Experiments

We now evaluate the forking-sequences architecture design along five key axes. (i) Accuracy: Do models trained with forking-sequences outperform those trained with window-sampling? (ii) Stability: To what extent do forking-sequences ensembles improve forecast stability? (iii) Accuracy–Stability Tradeoff: Can ensembling enhance stability without degrading accuracy? (iv) Architecture Interactions: How do the benefits of forking-sequences vary across different encoder architectures? (v) Ablations: How do alternative ensembling techniques affect accuracy and stability, and how does forking-sequences mitigate vanishing gradients? Can we extend forking-sequences ensembling to work with established architectures such as `NBEATS` and `PatchTST`. Below, we describe the benchmarks, baselines, and training/testing protocols, followed by quantitative results and references to the ablation studies addressing each question.

Benchmark Datasets. To measure the forking-sequences effects, we rely on the M-series benchmark. We use 16 large-scale time-series forecasting benchmarks, containing over 100,000 time series, drawn from well-known forecasting competitions: `M1` (Makridakis et al., 1982), `M3` (Makridakis and Hibon, 2000), `M4` (Makridakis et al., 2020), and `Tourism` (Athanasopoulos et al., 2011). We adopt the data handling and pre-processing practices established in prior work on cross-frequency transfer learning (Alexandrov et al., 2020; Olivares et al., 2025). To enable temporal cross-validation in our experiments, we expand the original test sets of a single forecast horizon to include earlier timesteps for n prediction windows consecutively shifted by one FCD as depicted in Fig. 7. We adjust the validation set accordingly to equal a single forecast horizon preceding the test set with the train set preceding the validation set.

Let h be the forecast horizon as specified in Table 6 and depicted in Figure 7, we define the train, validation, and test partitions, using numpy notation, of the data as follows:

$$\mathbf{Y}_{\text{train}} = \mathbf{Y}_{[\,:\,-3h+1]}, \qquad \mathbf{Y}_{\text{validation}} = \mathbf{Y}_{[-3h+1\,:\,-2h+1]}, \qquad \text{and} \qquad \mathbf{Y}_{\text{test}} = \mathbf{Y}_{[-2h+1\,:\,t]}. \tag{13}$$

We train frequency-specialized models combining `M1`, `M3`, `M4`, and `Tourism` frequency-specific datasets. This enables us to leverage a larger training data corpus, while still focusing the forecasting tasks to frequency-

specific prediction horizons (see Table 6). We default to these frequency-specific horizon values for the Tourism dataset as well for consistency across experiments.

Forecast Baselines For our main experiments, we evaluated a curated set of baseline models. These include the classic statistical model AutoARIMA (Hyndman and Khandakar, 2008; Hyndman et al., 2025) and several neural forecasting models designed for controlled comparisons. We fix the overall MQForecaster (Wen et al., 2017; Olivares et al., 2023) architecture and vary only the encoder and the training scheme. We select six fundamental architectures that represent the main neural backbones used throughout the time series forecasting literature: multilayer perceptron (`MLP`), recurrent neural network (`RNN`), long-short-term memory (`LSTM`), convolutional neural network (`CNN`), `Transformer` and `State Space`. Our work employs and adapts a dilated RNN implementation (Chang et al., 2017; Olivares et al., 2022). The `Transformer` encoder uses skip connections and residual layers with dilated SelfAttention layers inspired by dilated convolutions (van den Oord et al., 2016). The `State Space` encoder uses the architecture from the Structured State Space sequence model (`S4`) proposed by Gu et al. (2022). For each encoder variant of MQForecaster, we train models with forking-sequences and compare their generalization performance against models trained with window-sampling. We train models with optimized hyperparameters, including batch size and learning rate, across different encoder architectures. Details on hyperparameter selection are provided in the Appendix G.

4.1 Accuracy and Stability Metrics

To assess forecast accuracy, we use the *scaled Continuous Ranked Probability Score* (sCRPS, (Gneiting and Raftery, 2007)); and the *Mean Absolute Error* (MAE). The main paper reports probabilistic forecasts' accuracy using sCRPS while additional MAE-based point forecast results are reported in Appendix J. sCRPS is a scaled version of the CRPS and is defined in Equation 14:

$$\mathrm{sCRPS}\left(\mathbf{y}_{[b][t][h]},\ \hat{\mathbf{y}}_{[b][t][h]}\right) = \frac{\sum_{b,t,h} \mathrm{CRPS}(y_{b,t,h},\ \hat{\mathbf{y}}_{b,t,h})}{\sum_{b,t,h} |y_{b,t,h}|}. \tag{14}$$

where CRPS is defined as an integral over quantiles:

$$\mathrm{CRPS}(y, \hat{\mathbf{y}}) = \int_0^1 \mathrm{QL}(y, \hat{y}^{(q)}) dq \approx \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathrm{QL}(y, \hat{y}^{(q)}). \tag{15}$$

In practice, we approximate the integral above with a finite sum over equally spaced quantile levels, denoted by the set $\mathcal{Q}$. We use nine quantiles $q \in \{0.1, 0.2, \ldots, 0.9\}$ in our evaluation.

To asses the forecast stability we use *Scaled Excess Volatility* (sEV) and *Symmetric Quantile Percentage Change* (sQPC). Like sCRPS, sEV is a scaled measurement that enables a comparable evaluation of the forecast stability between datasets. The main body of the paper emphasizes sEV due to its compatibility with multi-quantile forecasting, while conventional sQPC-based point-forecast results are reported in Appendix J. sEV is defined as:

$$\mathrm{sEV}\left(\mathbf{y}_{[b][t][h]},\ \hat{\mathbf{y}}_{[b][t][h]}\right) = \frac{\sum_{b,t,h} \mathrm{EV}(y_{b,t,h},\ \hat{\mathbf{y}}_{b,t,h+1},\ \hat{\mathbf{y}}_{b,t+1,h})}{\sum_{b,t,h} |y_{b,t,h}|}. \tag{16}$$

4.2 Main Empirical Results

Table 2 reports mean sCRPS values, averaged over five runs, for all dataset–frequency combinations and model architectures, including statistical baselines. For `LSTM` encoder models, the forking-sequences training scheme reduced sCRPS by 49.3%, on average across datasets, compared to the window-sampling scheme, confirming earlier observations by (Lipton et al., 2016; Dai and Le, 2015; Lee et al., 2015). Consistent median gains across datasets are observed for `MLP` (29.7%), `RNN` (46.2%), and `CNN` (28.6%) encoders, while the `Transformer` (24.7%) and `State Space` (6.4%) encoders showed smaller improvement.

In Appendix B we present evidence that this difference stems from the superior gradient flow of these encoders capable of reducing vanishing gradient issues of recurrent architectures. The results of Table 2 are summarized

Table 2: Empirical evaluation of probabilistic forecasts. Mean scaled continuous ranked probability score (sCRPS) averaged over 5 runs. Lower measurements are preferred. The methods without standard deviation have deterministic solutions. For the MQForecaster architecture we vary the type of encoder, and the training scheme between forking-sequences (FS) and window-sampling (WS).

		MLP		RNN		LSTM		CNN		Transformer		S4		StatsForecast	
	Freq	FS	WS	FS	WS	FS	WS	FS	WS	FS	WS	FS	WS	ETS	ARIMA
M1	M	0.127	0.197	0.654	0.89	0.659	0.863	0.128	0.209	0.143	0.191	0.217	0.213	0.142	0.151
		(0.002)	(0.016)	(0.086)	(0.092)	(0.075)	(0.13)	(0.002)	(0.017)	(0.008)	(0.014)	(0.029)	(0.003)	(-)	(-)
	Q	0.111	0.123	0.643	0.89	0.675	0.954	0.114	0.125	0.119	0.137	0.138	0.129	0.114	0.128
		(0.002)	(0.005)	(0.051)	(0.032)	(0.085)	(0.09)	(0.007)	(0.009)	(0.004)	(0.02)	(0.011)	(0.002)	(-)	(-)
	Y	0.104	0.159	0.995	0.998	0.995	0.997	0.102	0.102	0.11	0.127	0.111	0.115	0.11	0.107
		(0.006)	(0.096)	(0.001)	(0.001)	(0.001)	(0.002)	(0.007)	(0.01)	(0.009)	(0.02)	(0.005)	(0.008)	(-)	(-)
M3	O	0.036	0.067	0.067	0.231	0.114	0.331	0.038	0.048	0.034	0.079	0.052	0.042	0.033	0.034
		(0.003)	(0.037)	(0.037)	(0.105)	(0.136)	(0.114)	(0.008)	(0.007)	(0.001)	(0.039)	(0.004)	(0.0004)	(-)	(-)
	M	0.091	0.127	0.123	0.227	0.117	0.233	0.089	0.14	0.096	0.127	0.1	0.133	0.105	0.106
		(0.001)	(0.007)	(0.006)	(0.081)	(0.005)	(0.08)	(0.001)	(0.023)	(0.005)	(0.01)	(0.004)	(0.0003)	(-)	(-)
	Q	0.069	0.09	0.085	0.136	0.087	0.214	0.071	0.094	0.071	0.116	0.076	0.088	0.077	0.078
		(0.001)	(0.004)	(0.003)	(0.043)	(0.005)	(0.056)	(0.005)	(0.006)	(0.003)	(0.055)	(0.004)	(0.001)	(-)	(-)
	Y	0.138	0.198	0.165	0.348	0.176	0.352	0.126	0.168	0.135	0.177	0.14	0.149	0.149	0.155
		(0.006)	(0.065)	(0.019)	(0.135)	(0.028)	(0.088)	(0.002)	(0.006)	(0.013)	(0.033)	(0.007)	(0.003)	(-)	(-)
M4	H	0.031	0.133	0.304	0.637	0.315	0.581	0.031	0.159	0.04	0.133	0.043	0.155	0.068	0.031
		(0.003)	(0.01)	(0.079)	(0.03)	(0.092)	(0.036)	(0.003)	(0.006)	(0.007)	(0.011)	(0.007)	(0.002)	(-)	(-)
	D	0.021	0.022	0.025	0.208	0.026	0.228	0.021	0.052	0.022	0.022	0.022	0.022	0.023	0.023
		(0.0003)	(0.0001)	(0.001)	(0.157)	(0.003)	(0.141)	(0.0004)	(0.028)	(0.0003)	(0.001)	(0.0002)	(0.0004)	(-)	(-)
	W	0.044	0.061	0.061	0.196	0.064	0.305	0.043	0.069	0.046	0.064	0.052	0.06	0.054	0.051
		(0.002)	(0.001)	(0.003)	(0.055)	(0.003)	(0.057)	(0.001)	(0.013)	(0.006)	(0.005)	(0.004)	(0.0004)	(-)	(-)
	M	0.089	0.115	0.114	0.299	0.11	0.304	0.087	0.129	0.092	0.119	0.096	0.117	0.097	0.099
		(0.001)	(0.002)	(0.006)	(0.131)	(0.005)	(0.125)	(0.0005)	(0.026)	(0.003)	(0.01)	(0.003)	(0.001)	(-)	(-)
	Q	0.077	0.097	0.092	0.173	0.097	0.335	0.079	0.101	0.081	0.127	0.085	0.094	0.081	0.085
		(0.001)	(0.005)	(0.002)	(0.038)	(0.007)	(0.102)	(0.005)	(0.007)	(0.005)	(0.065)	(0.004)	(0.001)	(-)	(-)
	Y	0.115	0.181	0.185	0.428	0.21	0.411	0.108	0.138	0.115	0.151	0.12	0.128	0.124	0.132
		(0.005)	(0.086)	(0.039)	(0.135)	(0.066)	(0.119)	(0.002)	(0.003)	(0.007)	(0.027)	(0.004)	(0.002)	(-)	(-)
Tourism	M	0.088	0.261	0.66	0.838	0.653	0.82	0.096	0.275	0.131	0.25	0.241	0.281	0.12	0.148
		(0.005)	(0.023)	(0.073)	(0.05)	(0.068)	(0.064)	(0.006)	(0.015)	(0.033)	(0.021)	(0.036)	(0.002)	(-)	(-)
	Q	0.098	0.168	0.808	0.931	0.818	0.943	0.104	0.164	0.099	0.226	0.174	0.172	0.104	0.119
		(0.004)	(0.003)	(0.023)	(0.008)	(0.048)	(0.015)	(0.013)	(0.008)	(0.003)	(0.11)	(0.011)	(0.002)	(-)	(-)
	Y	0.17	0.24	0.982	0.99	0.982	0.989	0.163	0.193	0.172	0.216	0.185	0.187	0.141	0.17
		(0.004)	(0.081)	(0.002)	(0.004)	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)	(0.038)	(0.003)	(0.003)	(-)	(-)

in Figure 14a which shows the percentage change of the sCRPS metric for models with forking-sequences over that of models with window-sampling, averaged across datasets. All encoder variants with forking-sequences have improved sCRPS and generally outperform statistical baselines. Point forecasting results are reported in Appendix J.

Based on median sCRPS across datasets in Table 2, the top-performing models are forking-sequences variants of `MLP`, `CNN`, and `Transformer` encoders, in that order, all outperforming the statistical baselines `AutoETS` and `AutoARIMA`. The `S4` encoder with forking-sequences ranks fifth, following `AutoETS` but ahead of `AutoARIMA`. Among forking-sequences models, only the `RNN` and `LSTM` encoders do not outperform statistical baselines.

Table 3 presents sEV values, averaged over five runs, for all dataset–frequency combinations and model architectures, including statistical baselines. Improved median sEV across datasets is observed for `MLP` (21.7%), `CNN` (35.1%), and `S4` (2.8%) encoders with forking-sequences, while `RNN` (-43.6%), `LSTM` (-75.1%), and `Transformer` (-11.7%) encoders achieve better median sEV with window-sampling. Table 2 is summarized in Fig. 14e which shows the percentage change of the sEV metric for models with forking-sequences over that of models with window-sampling, averaged across datasets. We present sQPC results in Table 15.

We demonstrate that forecast ensembling during inference improves forecast stability across all encoders trained with forking-sequences in Fig. 13. Using exponential smoothing ($\alpha = 0.9$) during inference, models trained with forking-sequences show median sEV percentage changes relative to window-sampling of 29.7%, -24.5%, -52.4%, 42.2%, 1.0%, and 13.7% for `MLP`, `RNN`, `LSTM`, `CNN`, `Transformer`, and `S4` encoders, respectively, without degrading accuracy as shown in Figs 14b and 14f.

Table 3: Empirical evaluation of probabilistic forecasts stability. Mean *scaled Excess Volatility* (sEV) averaged over 5 runs. Lower measurements are preferred. The methods without standard deviation have deterministic solutions. For the `MQForecaster` architecture we vary the type of encoder, and the training scheme between forking-sequences (FS) and window-sampling (WS).

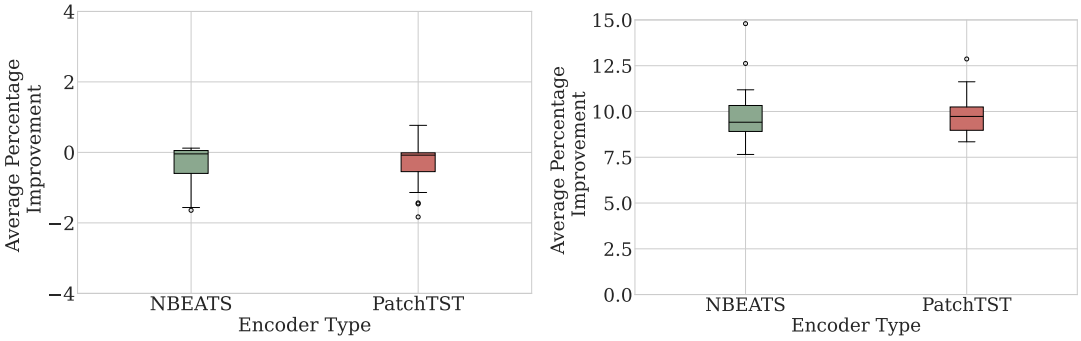
		`MLP`		`RNN`		`LSTM`		`CNN`		`Transformer`		`S4`		StatsForecast	
	Freq	FS	WS	FS	WS	FS	WS	FS	WS	FS	WS	FS	WS	`ETS`	`ARIMA`
M1	M	0.017	0.034	0.007	0.003	0.005	0.002	0.015	0.044	0.02	0.016	0.04	0.05	0.007	0.009
		(0.002)	(0.021)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)	(0.006)	(0.002)	(0.017)	(0.004)	(0.001)	(-)	(-)
	Q	0.019	0.019	0.005	0.002	0.003	0.001	0.016	0.019	0.014	0.019	0.027	0.023	0.013	0.024
		(0.001)	(0.007)	(0.003)	(0.0004)	(0.001)	(0.0001)	(0.001)	(0.002)	(0.003)	(0.009)	(0.002)	(0.0004)	(-)	(-)
	Y	0.026	0.012	0.0001	0.0001	0.0001	0.0001	0.023	0.016	0.023	0.012	0.018	0.016	0.012	0.017
		(0.002)	(0.008)	(0)	(0)	(0)	(0)	(0.002)	(0.002)	(0.004)	(0.006)	(0.003)	(0.002)	(-)	(-)
M3	O	0.005	0.002	0.004	0.004	0.002	0.001	0.005	0.003	0.005	0.002	0.004	0.003	0.004	0.004
		(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.0002)	(0.0002)	(0.0004)	(0.002)	(0.001)	(0.0004)	(-)	(-)
	M	0.011	0.017	0.018	0.012	0.015	0.01	0.01	0.022	0.012	0.009	0.015	0.025	0.005	0.006
		(0.001)	(0.011)	(0.003)	(0.01)	(0.001)	(0.008)	(0.0003)	(0.003)	(0.001)	(0.009)	(0.002)	(0.001)	(-)	(-)
	Q	0.009	0.013	0.011	0.015	0.01	0.005	0.009	0.013	0.008	0.013	0.012	0.016	0.007	0.009
		(0.0005)	(0.005)	(0.001)	(0.002)	(0.002)	(0.003)	(0.0005)	(0.002)	(0.002)	(0.006)	(0.001)	(0.0005)	(-)	(-)
	Y	0.023	0.013	0.016	0.009	0.015	0.008	0.023	0.016	0.022	0.014	0.019	0.017	0.013	0.021
		(0.002)	(0.008)	(0.003)	(0.002)	(0.003)	(0.002)	(0.001)	(0.002)	(0.003)	(0.006)	(0.002)	(0.002)	(-)	(-)
M4	H	0.002	0.004	0.002	0.005	0.001	0.003	0.002	0.012	0.003	0.004	0.005	0.011	0.004	0.001
		(0.0004)	(0.003)	(0.0004)	(0.002)	(0.001)	(0.001)	(0.0002)	(0.001)	(0.001)	(0.003)	(0.001)	(0.0003)	(-)	(-)
	D	0.002	0.002	0.002	0.001	0.002	0.0005	0.002	0.013	0.002	0.002	0.002	0.002	0.002	0.002
		(0.0002)	(0)	(0.0004)	(0.001)	(0.0005)	(0.0002)	(0.0002)	(0.02)	(0.0002)	(0.0001)	(0.0001)	(0)	(-)	(-)
	W	0.005	0.007	0.006	0.005	0.005	0.003	0.005	0.008	0.005	0.005	0.007	0.008	0.005	0.004
		(0.0005)	(0.003)	(0.001)	(0.001)	(0.0002)	(0.001)	(0.0002)	(0.001)	(0.002)	(0.003)	(0.0003)	(0.0001)	(-)	(-)
	M	0.009	0.012	0.012	0.007	0.009	0.005	0.008	0.015	0.01	0.006	0.011	0.017	0.008	0.008
		(0.0005)	(0.007)	(0.001)	(0.005)	(0.0004)	(0.004)	(0.0002)	(0.002)	(0.001)	(0.006)	(0.001)	(0.001)	(-)	(-)
	Q	0.011	0.013	0.012	0.011	0.011	0.004	0.01	0.013	0.009	0.014	0.013	0.016	0.008	0.009
		(0.0001)	(0.005)	(0.001)	(0.001)	(0.001)	(0.002)	(0.0005)	(0.001)	(0.002)	(0.006)	(0.001)	(0.0005)	(-)	(-)
	Y	0.019	0.01	0.011	0.008	0.01	0.007	0.019	0.014	0.018	0.011	0.015	0.014	0.011	0.019
		(0.001)	(0.006)	(0.001)	(0.002)	(0.001)	(0.002)	(0.0004)	(0.002)	(0.003)	(0.005)	(0.002)	(0.002)	(-)	(-)
Tourism	M	0.013	0.034	0.008	0.004	0.006	0.003	0.013	0.043	0.018	0.018	0.032	0.047	0.009	0.009
		(0.002)	(0.018)	(0.002)	(0.003)	(0.001)	(0.002)	(0.001)	(0.005)	(0.005)	(0.016)	(0.003)	(0.001)	(-)	(-)
	Q	0.017	0.041	0.006	0.002	0.005	0.001	0.019	0.04	0.014	0.041	0.051	0.05	0.019	0.02
		(0.001)	(0.019)	(0.001)	(0.0003)	(0.001)	(0.001)	(0.004)	(0.008)	(0.003)	(0.02)	(0.002)	(0.001)	(-)	(-)
	Y	0.027	0.013	0.0002	0.0002	0.0002	0.0001	0.031	0.018	0.022	0.014	0.019	0.017	0.023	0.028
		(0.002)	(0.007)	(0)	(0.0001)	(0)	(0.0001)	(0.002)	(0.002)	(0.003)	(0.005)	(0.004)	(0.002)	(-)	(-)

4.3 Ablation Studies

In Table 2, we observe that the performance gains from forking-sequences vary between encoder architectures. In controlled experiments where only the encoder module differs, `Transformer` and `State Space` encoder models show smaller improvements, which we hypothesized may stem from their inherently stronger gradient propagation compared to recurrent architectures. Appendix B presents both empirical and theoretical exploration of how forking-sequences and window-sampling influence gradient flow in LSTMs and Transformers. By analyzing the Jacobian products that compose backpropagation-through-time, we find that window-sampling LSTM encoders suffer from vanishing gradients with respect to the series length, whereas Transformer encoders do not. Moreover, forking-sequences LSTM encoder does not have the vanishing gradient problem. We complement these findings with explicit measurements of gradient norms between positions in the context window (Figure 10).

Our ablation studies in Appendix H reveal a trade-off between forecast stability and accuracy when applying different ensembling functions as shown in Figs. 1, 13, 15, and 16. Exponential smoothing with $\alpha = 0.9$ emerges as an effective approach, successfully improving forecast stability while maintaining model accuracy. Although lower α values improve forecast stability in terms of sEV, they do so at the cost of accuracy. This trade-off is intuitive since α controls the weighting of future forecasts: higher α values give more weight to target date predictions closer in the horizon, which are generally more accurate than predictions for the same target date farther in the horizon. As such, this weighting scheme helps preserve the accuracy of the forecast while improving stability.

In Appendix I, we examine whether the forking-sequences ensembling benefits extend beyond architectures specifically designed for it. We evaluated `NBEATS` (Oreshkin et al., 2020) and `PatchTST` (Nie et al., 2023), both trained as foundation models on synthetic data and tested in a cross-frequency transfer-learning setting. Our results show that these architectures can be seamlessly augmented with forking-sequences ensembles without



(a) sCRPS Effects **(b)** sEV Effects

Figure 8: Ablation study on applying forking-sequences ensembles to window-sampling foundation forecasting models. a) Ensembling is capable of maintaining the original accuracy of the foundation forecasting models. b) We observe significant improvements in their forecast stability measured with the scaled Excessive Volatility.

compromising their original accuracy (less than 0.1% difference), while achieving almost 10% improvement in forecast stability across the M-series benchmark. We summarize these findings in Figure 8.

In our final ablation study in Appendix K, we investigate the effects of forecast ensembles on the probabilistic coverage of forecasts across the M1 and M4 datasets. We found no consistent calibration effect attributable to the ensemble. On M1 dataset, increasing the ensemble's window size increased under-coverage for P50 and P90 while improving P10 coverage. Conversely, on the M4 dataset, increasing the window size improved P10 and P50 over-coverage, shifting P10 and P50 toward under-coverage. These mixed, dataset- and quantile-dependent results suggest the coverage impact of simple ensembling is inconclusive, highlighting the need for future research to explore more sophisticated, conditionally-aware aggregation methods, such as those inspired by Quantile Regression Averaging (Liu et al., 2017) or weighted ensembles with quantile-specific weights (Fakoor et al., 2023).

5 Discussion and Conclusion

Central Findings. We re-introduced *forking-sequences* and proposed a novel *excess volatility* metric to measure forecast stability. We provided extensive theoretical and empirical validation of the main benefits of forking-sequences, showing: 1) improved gradients through variance reduction and mitigation of vanishing gradients, 2) an order-of-magnitude improvement in the computational efficiency of cross-validation inference compared with window-sampling, and 3) a natural ability to ensemble overlapping multi-horizon forecasts. Using the M-series benchmark; 16 datasets from the M1, M3, M4, and Tourism competitions, we demonstrated significant improvements in forecast stability with no accuracy tradeoff, both for MQForecaster architectures and for other foundation forecasting models.

Limitations and Research Directions. This forking-sequences study focused only on univariate time series forecasting, but the technique can be readily extended to multivariate settings. We chose the univariate case to isolate the theoretical and empirical effects of the paradigm in the most general forecasting scenario, leaving multivariate extensions as a promising direction for future work. Additionally, we focus on simple forecast ensemble strategies (see the Appendix H), but accuracy-stability tradeoffs could be further improved by integrating forecasting-specific inductive biases in the ensemble design or by learning the ensemble weights directly (Liu et al., 2017; Eisenach et al., 2021; Fakoor et al., 2023). We restricted our analysis to frequency-specialized models, but foundation forecasting models are typically trained in a cross-frequency transfer learning setting (Ness et al., 2023; Olivares et al., 2025). It remains an open question whether the benefits of forking-sequences extend to scenarios where series vary widely in scale and length. A foreseeable limitation is the increase in memory usage when applying the forking-sequences to high-frequency data, where time series tend to be of longer length.

In computer vision and natural language processing, the forking-sequences architectural design and its associated training paradigms are the default, yet they remain underexplored in the forecasting domain. The limited adoption of forking-sequences is partly explained by the dominance of simple window-sampling pipelines in major neural forecasting libraries `TSLib` (Wu et al., 2023), `Darts` (Herzen et al., 2022), `GluonTS` (Alexandrov et al., 2020), `PytorchForecasting` (Beitner, 2020), and `NeuralForecast` (Olivares et al., 2022). As mainstream neural forecast libraries grow in complexity, integrating forking-sequence–based training and inference becomes increasingly difficult. We hope that our results motivate the rediscovery of these techniques within the forecasting community and encourage library maintainers to adopt forking-sequences.

References

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Tarkmen, and Yuyang Wang. GluonTS: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020. URL `http://jmlr.org/papers/v21/19-820.html`.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024.

George Athanasopoulos, Rob J. Hyndman, Haiyan Song, and Doris C. Wu. The Tourism forecasting competition. *International Journal of Forecasting*, 27(3):822–844, 2011. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2010.04.009. URL `https://www.sciencedirect.com/science/article/pii/S016920701000107X`. Special Section 1: Forecasting with Artificial Neural Networks and Computational Intelligence. Special Section 2: Tourism Forecasting.

Jan Beitner. PyTorchForecasting: Forecasting with neural networks made simple. GitHub Repository, 2020. URL `https://pytorch-forecasting.readthedocs.io/en/stable/`.

Jente Van Belle, Ruben Crevits, and Wouter Verbeke. Improving forecast stability using deep learning. *International Journal of Forecasting*, 39(3):1333–1350, 2023.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

Christoph Bergmeir and José M Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Daan Caljon, Jeff Vercauteren, Simon De Vos, Wouter Verbeke, and Jente Van Belle. Using dynamic loss weighting to boost improvements in forecast stability. *International Journal of Forecasting*, 2025.

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramirez, Max Mergenthaler-Canseco, and Artur Dubrawski. NHITS: neural hierarchical interpolation for time series forecasting. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial*

Intelligence, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25854. URL https://doi.org/10.1609/aaai.v37i6.25854.

Shiyu Chang, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, and et al. Dilated recurrent neural networks. In *31st Conference on Neural Information Processing Systems*, 2017.

Kevin C. Chen, Lee Dicker, Carson Eisenach, and Dhruv Madeka. MQTransformer: Multi-Horizon Forecasts with Context Dependent Attention and Optimal Bregman Volatility. In Maria Florina Balcan and Marina Meila, editors, *In: Proceedings of 8th SIGKDD International Workshop on Mining and Learning From Time Series - Deep Forecasting: Models, Interpretability, and Applications (KDD' 22)*. ACM. Association for Computing Machinery, 8 2022. URL https://kdd-milets.github.io/milets2022/papers/MILETS_2022_paper_9623.pdf.

Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems, NIPS*, 2015.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

Carson Eisenach, Yagna Patel, and Dhruv Madeka. MQTransformer: Multi-Horizon Forecasts with Context Dependent and Feedback-Aware Attention. In Maria Florina Balcan and Marina Meila, editors, *Submitted to Proceedings of the 38th International Conference on Machine Learning*. PMLR. Working Paper version available at arXiv:2009.14799, 8 2021.

Graham Elliott and Allan Timmermann. Economic forecasting. *Journal of Economic Literature*, 46(1):3–56, March 2008. doi: 10.1257/jel.46.1.3. URL https://www.aeaweb.org/articles?id=10.1257/jel.46.1.3.

Rasool Fakoor, Taesup Kim, Jonas Mueller, Alexander J. Smola, and Ryan J. Tibshirani. Flexible model aggregation for quantile regression. *Journal of Machine Learning Research*, 24(162):1–45, 2023. URL http://jmlr.org/papers/v24/22-0799.html.

Dean P. Foster and Robert A. Stine. Threshold martingales and the evolution of forecasts, 2021. URL https://arxiv.org/abs/2105.06834.

Azul Garza and Max Mergenthaler-Canseco. Timegpt, 2023. URL https://arxiv.org/abs/2310.03589.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.00396.

Jeffrey D. Hart. Automated kernel smoothing of dependent data by using time series cross- validation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):529–542, 1994. ISSN 00359246. URL http://www.jstor.org/stable/2346125.

Julien Herzen, Francesco LÃ¤ssig, Samuele Giuliano Piazzetta, Thomas Neuer, LÃ©o Tafti, Guillaume Raille, Tomas Van Pottelbergh, Marek Pasieka, Andrzej Skrodzki, Nicolas Huguenin, Maxime Dumonal, Jan KoÅ›cisz, Dennis Bader, FrÃ©dÃ©rick Gusset, Mounir Benheddi, Camila Williamson, Michal Kosinski, Matej Petrik, and GaÃ«l Grosch. Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research*, 23(124):1–6, 2022. URL http://jmlr.org/papers/v23/21-1177.html.

Charles C Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *(O.N.R. Memorandum No. 52)*, 1957. URL https://doi.org/10.1016/j.ijforecast.2003.09.015.

Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913, 2016.

Rob J. Hyndman and Baki Billah. Unmasking the theta method. *International Journal of Forecasting*, 19 (2):287–290, 2003. ISSN 0169-2070. doi: https://doi.org/10.1016/S0169-2070(01)00143-1. URL https://www.sciencedirect.com/science/article/pii/S0169207001001431.

Rob J. Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software, Articles*, 27(3):1–22, 2008. ISSN 1548-7660. doi: 10.18637/jss.v027.i03. URL https://www.jstatsoft.org/v027/i03.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688, 2006. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2006.03.001. URL http://www.sciencedirect.com/science/article/pii/S0169207006000239.

Rob J Hyndman, George Athanasopoulos, Azul Garza, Cristian Challu, Max Mergenthaler, and Kin G. Olivares. *Forecasting: Principles and Practice, the Pythonic Way*. OTexts, Melbourne, Australia, 2025. available at https://otexts.com/fppy/.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913643.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 562–570, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/lee15a.html.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2021.03.012. URL https://www.sciencedirect.com/science/article/pii/S0169207021000637.

Zachary Chase Lipton, David C. Kale, Charles Elkan, and Randall C. Wetzel. Learning to diagnose with LSTM recurrent neural networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.03677.

Bidong Liu, Jakub Nowotarski, Tao Hong, and Rafał Weron. Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Transactions on Smart Grid*, 8(2):730–737, 2017. doi: 10.1109/TSG.2015.2437877.

S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2):111–153, 1982. doi: https://doi.org/10.1002/for.3980010202. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980010202.

Spyros Makridakis and Michèle Hibon. The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476, 2000. ISSN 0169-2070. doi: https://doi.org/10.1016/S0169-2070(00)00057-1. URL https://www.sciencedirect.com/science/article/pii/S0169207000000571. The M3- Competition.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.04.014. URL https://www.sciencedirect.com/science/article/pii/S0169207019301128. M4 Competition.

Albert W. Marshall and Ingram Olkin. Multivariate Chebyshev Inequalities. *The Annals of Mathematical Statistics*, 31(4):1001 – 1014, 1960. doi: 10.1214/aoms/1177705673. URL https://doi.org/10.1214/aoms/1177705673.

Kenwin Maung and Norman R. Swanson. A survey of models and methods used for forecasting when investing in financial markets. *International Journal of Forecasting*, 41(4):1355–1382, 2025.

Mike Van Ness, Huibin Shen, Hao Wang, Xiaoyong Jin, Danielle C. Maddix, and Karthick Gopalswamy. Cross-frequency time series meta-forecasting, 2023. URL https://arxiv.org/abs/2302.02077.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Konstantinos Nikolopoulos, Sushil Punia, Andreas Schäfers, Christos Tsinopoulos, and Chrysovalantis Vasilakis. Forecasting and planning during a pandemic: Covid-19 growth rates, supply chain disruptions, and governmental decisions. *European Journal of Operational Research*, 290(1), 2021.

Kin G. Olivares, Cristian Challú, Federico Garza, Max Mergenthaler Canseco, and Artur Dubrawski. NeuralForecast: User friendly state-of-the-art neural forecasting models. PyCon Salt Lake City, Utah, US 2022, 2022. URL https://github.com/Nixtla/neuralforecast.

Kin G. Olivares, Nganba Meetei, Ruijun Ma, Rohan Reddy, Mengfei Cao, and Lee Dicker. Probabilistic hierarchical forecasting with deep poisson mixtures. *International Journal of Forecasting, accepted*, Preprint version available at arXiv:2110.13179, 2023. URL https://arxiv.org/abs/2110.13179.

Kin G. Olivares, Malcolm Wolff, Tatiana Konstantinova, Shankar Ramasubramanian, Boris Oreshkin, Andrew Gordon Wilson, Andres Potapczynski, Willa Potosnak, Mengfei Cao, Michael W. Mahoney, and Dmitry Efimov. A realistic evaluation of cross-frequency transfer learning and foundation forecasting models. In *Thirty-Ninth Annual Conference on Neural Information Processing Systems NeurIPS 2025*, volume Recent Advances in Time Series Foundation Models Have We Reached the 'BERT Moment'?, San Diego, USA, 2025. NeurIPS 2025. URL https://arxiv.org/abs/2411.05852.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: neural basis expansion analysis for interpretable time series forecasting. In *8th International Conference on Learning Representations, ICLR 2020*, 2020. URL https://openreview.net/forum?id=r1ecqn4YwB.

Willa Potosnak, Cristian Ignacio Challu, Kin G. Olivares, Keith A Dufendach, and Artur Dubrawski. Global deep forecasting with patient-specific pharmacokinetics. In *Proceedings of Machine Learning Research*, 2025. URL https://raw.githubusercontent.com/mlresearch/v287/main/assets/potosnak25a/potosnak25a.pdf.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.07.001. URL https://www.sciencedirect.com/science/article/pii/S0169207019301888.

R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications: With R Examples. Fifth Edition. Appendix A. On Large Sample Theory.* Springer Texts in Statistics. Springer International Publishing, 2017. ISBN 9783319524528. URL https://books.google.com/books?id=sfFdDwAAQBAJ.

Slawek Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 07 2019. doi: 10.1016/j.ijforecast.2019.03.017.

M Syntetos, John Boylan, and JD Croston. On the categorization of demand patterns. *Journal of the Operational Research Society*, 56, 05 2005. doi: 10.1057/palgrave.jors.2601841.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *Computer Research Repository*, abs/1609.03499, 2016. URL http://arxiv.org/abs/1609.03499.

Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A Multi-horizon Quantile Recurrent Forecaster. In *31st Conference on Neural Information Processing Systems NIPS 2017, Time Series Workshop*, 2017. URL https://arxiv.org/abs/1711.11053.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989. doi: 10.1162/neco.1989.1.2.270.

Malcolm Wolff, Kin G. Olivares, Boris Oreshkin, Sunny Ruan, Sitan Yang, Abhinav Katoch, Shankar Ramasubramanian, Youxin Zhang, Michael W. Mahoney, Dmitry Efimov, and Vincent Quenneville-Bélair. ♠ SPADE ♠: Split Peak Attention DEcomposition. In *Thirty-Eighth Annual Conference on Neural Information Processing Systems NeurIPS 2024*, volume Time Series in the Age of Large Models Workshop, Vancouver, Canada, 2024. NeurIPS 2024. URL https://arxiv.org/abs/2411.05852.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

Haixu Wu, Yong Liu, Huikun Weng, Yuxuan Wang, Tengge Hu, Haoran Zhang, and Jiawei Guo. Time Series Library (TSLib). GitHub Repository, 2023. URL https://github.com/thuml/Time-Series-Library.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *The Association for the Advancement of Artificial Intelligence Conference 2021 (AAAI 2021).*, abs/2012.07436, 2020. URL https://arxiv.org/abs/2012.07436.

A Forking-Sequences Theoretical Foundations

Here we prove the gradient variance reduction guaranties for the forking-sequences training scheme. The gradient is computed with respect to the model parameters by minimizing Equation 5. We show that this scheme reduces the variance of the stochastic gradient around its mean, assuming short-range dependence among gradients across forecast creation dates. Our arguments are based on techniques and definitions from Shumway and Stoffer (2017) and Marshall and Olkin (1960). M-dependence is illustrated in Figure 9a.

For simplicity of the arguments' in the section we simplify the notation to refer to the forking-sequences gradient estimator:

$$\bar{\nabla} L_T = \frac{1}{B \times H} \sum_{b=1}^{B} \sum_{h=1}^{H} \nabla \mathcal{L}\left(y_{b,t,h},\ \hat{y}_{b,t,h}\right). \tag{17}$$

Definition 1. Let $\{\nabla L_t\} \subset \mathbb{R}^P$, with P the dimension of the network parameters $\boldsymbol{\theta}$, be a sequence of random variables, the sequence is M-dependent if:

- The gradient estimators are stationary $\mathbb{E}[\nabla L_t] = \mu$, with μ its expected gradient.
- with bounded stationary covariance $\Sigma = \mathrm{Cov}(\nabla L_s, \nabla L_t) = \gamma(|s-t|) \in \mathbb{R}^{P \times P}$.

  where $\gamma(b) = \mathrm{Cov}(\nabla L_0, \nabla L_b)$ and $\gamma(b) = 0$ for all $|b| > M$,

  that is, the gradient estimators $\{\nabla L_t\}$ are not correlated after M steps.

Lemma 1. Consider the forking sequences gradient estimator $\bar{\nabla} L_T = \frac{1}{T} \sum_t \nabla L_t$. If the gradient samples are M-dependent, the covariance estimator almost surely converges to 0 as T grows, that is, $\mathrm{Cov}\left(\bar{\nabla} L_T\right) \xrightarrow{a.s.} \mathbf{0}$.

We prove that with probability 1, the sequence of covariance estimator random variables converges to 0. That is $\mathbb{P}(\lim_{T\to\infty} \mathrm{Cov}\left(\bar{\nabla} L_T\right) = 0) = 1$

Proof.

$$\begin{aligned}
\mathrm{Cov}(\bar{\nabla} L_T) &= \frac{1}{T^2} \mathrm{Cov}\left(\sum_{t=1}^{T} \nabla L_t\right) \\
&= \frac{1}{T^2} \sum_{s=1}^{T} \sum_{t=1}^{T} \mathrm{Cov}(\nabla L_s, \nabla L_t) \\
&= \frac{1}{T^2} \sum_{b=-M}^{M} (T - |b|)\gamma(b) \\
&= \frac{1}{T} \sum_{b=-M}^{M} \left(1 - \frac{|b|}{T}\right) \gamma(b) \\
&\xrightarrow{a.s.} 0 \times \left(\sum_{b=-M}^{M} \gamma(b)\right) = 0 \times \Sigma
\end{aligned}$$

Using the M-dependence assumption, we know M is finite. Since $\frac{|b|}{T} \to 0$ then the finite sum converges to the stationary covariance. Finally using the product of almost sure convergences we conclude.

□

Theorem 1. (Forking-Sequences Gradient Variance Reduction) Consider the forking-sequences gradient estimator $\bar{\nabla} L_T = \frac{1}{T}\sum_t \nabla L_t$. If the gradient samples are M-dependent, then the estimator converges in probability to the true gradient μ at a rate $\mathcal{O}\left(\frac{1}{T}\right)$.

Proof. Using a change of variable and the Markov inequality, we have that

$$\mathbb{P}\left(\sqrt{(\bar{\nabla} L_T - \mu)^T \Sigma^{-1} (\bar{\nabla} L_T - \mu)} > \epsilon\right) = \mathbb{P}\left(\sqrt{v} > \epsilon\right) = \mathbb{P}\left(v > \epsilon^2\right) \leq \frac{\mathbb{E}[v]}{\epsilon^2}.$$

To bound the numerator, using cyclic property of the trace and the Lemma we know that

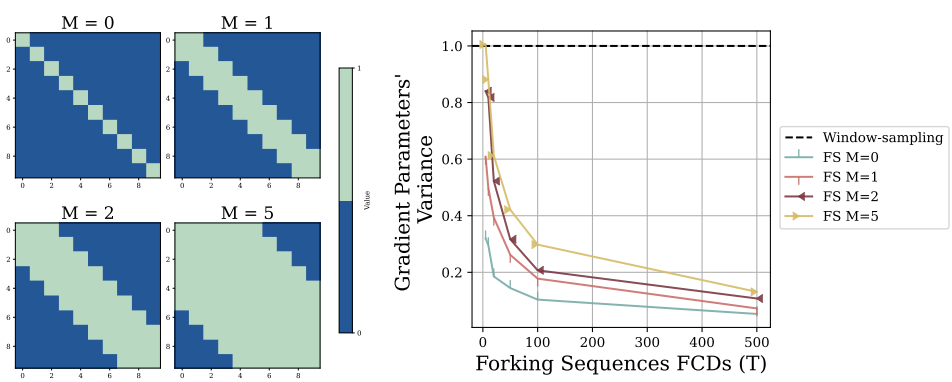
$$\begin{aligned}
\mathbb{E}\left[v\right] &= \mathbb{E}\left[(\bar{\nabla} L_T - \mu)^T \Sigma^{-1} (\bar{\nabla} L_T - \mu)\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left(\Sigma^{-1} (\bar{\nabla} L_T - \mu)(\bar{\nabla} L_T - \mu)^T\right)\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left(\Sigma^{-1} \mathrm{Cov}(\bar{\nabla} L_T)\right)\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left(\Sigma^{-1}\left(\frac{1}{T}\sum_{b=-K}^{K}\left(1 - \frac{|b|}{T}\right)\gamma(b)\right)\right)\right] \\
&\leq \frac{1}{T}\mathbb{E}\left[\mathrm{tr}\left(\Sigma^{-1}\Sigma\right)\right] \\
&= \frac{P}{T}.
\end{aligned}$$

By combining the two bounds, we obtain that

$$\mathbb{P}\left(\sqrt{(\bar{\nabla} L_T - \mu)^T \Sigma^{-1} (\bar{\nabla} L_T - \mu)} > \epsilon\right) \leq \frac{P}{T\epsilon^2}.$$

□

Here we empirically validate the theoretical results, with synthetic data. We control the M-dependence of a multivariate normal random variable using the covariances in Figure 9a, and then measure the variance of a mean estimator that we report in Figure 9b.



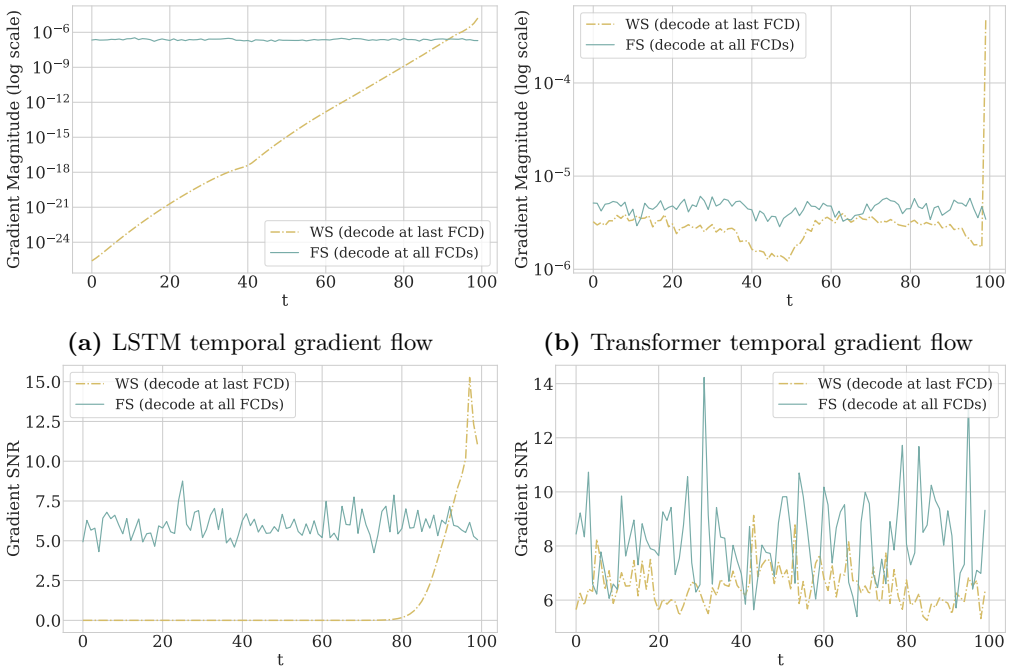
(a) M-dependence Covariance **(b)** Gradient Variance

Figure 9: a) Visualization of the covariance of M-dependent random variables. Green sections indicate correlated samples, while blue indicate uncorrelated variables. b) Mean estimator variance reduction as a function of the samples of the forking-sequences training scheme for different levels of M-dependence.

B Encoder Gradient Study

We observe that forking-sequences yields the largest performance improvement for the `LSTM` encoder compared to window-sampling. In contrast, forking-sequences has minimal impact on the `Transformer` encoder. We hypothesize that this difference stems from the Transformer's superior gradient flow: with window-sampling, Transformers encode entire context windows at once via self-attention, eliminating sequential information bottlenecks. LSTMs, however, suffer from vanishing gradients during backpropagation through time. Forking-sequences mitigates this issue for the `LSTM` encoder by decoding across all FCDs, which preserves stronger gradients to early timesteps (see Proposition 1). We demonstrate this effect empirically in Fig. 10 and provide theoretical justification in Propositions 1 and 2.

B.1 Empirical Gradient Analysis



(a) LSTM temporal gradient flow

(b) Transformer temporal gradient flow

(c) LSTM gradient SNR

(d) Transformer gradient SNR

Figure 10: (a) For the LSTM encoder, forking-sequences (FS) maintains stable gradients across timesteps, while window-sampling (which decodes only at the final FCD) exhibits exponential gradient decay. **(b)** Forking-sequences yields higher gradient signal-to-noise ratio (SNR) for LSTMs, indicating more stable training signals. **(c)** For the Transformer encoder, when losses are averaged by a factor of $\frac{1}{T}$, forking-sequences yields gradient magnitudes comparable to window-sampling. **(d)** Transformer gradient SNR remains relatively stable for both forking-sequences and window-sampling, as Transformers do not suffer from vanishing gradients.

To isolate gradient flow effects, we generate 100 synthetic time series of length $T = 100$ following $y(t) = 0.02t + \sin(t) + 0.5\sin(2t) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.01)$ and $t \in [0, 4\pi]$. We forecast $h = 12$ timesteps ahead using a single-layer LSTM encoder (hidden dimension 32) and a 2-layer Transformer encoder ($d_{\text{model}} = 32$, 4 heads, feedforward dimension 128), both paired with a 2-layer MLP decoder (hidden dimension 64). We measure temporal gradient flow by computing $\left\|\frac{\partial \mathcal{L}}{\partial x_t}\right\|$ at each input position $t \in \{0, \ldots, T-1\}$ after backpropagating loss against synthetic random targets. To assess gradient stability, we repeat this 20 times and compute signal-to-noise ratio (SNR) as $\text{SNR}_t = \mu\left(\left\|\frac{\partial \mathcal{L}}{\partial x_t}\right\|\right) / \left[\sigma\left(\left\|\frac{\partial \mathcal{L}}{\partial x_t}\right\|\right) + \epsilon\right]$ where $\epsilon = 10^{-8}$, with higher SNR indicating more consistent gradient signals. We find that forking-sequences preserves gradient flow for the `LSTM` encoder as shown in Fig. 10. For the `Transformer` encoder, window-sampling and forking-sequences have comparable SNR, which is expected as Transformers are not affected by vanishing gradient issues.

B.2 Theoretical Gradient Analysis

We supplement these empirical findings with theoretical analysis of gradient flow to early timesteps, specifically gradients at h_0, for `LSTM` and `Transformer` encoders under both forking-sequences and window-sampling. We begin with the following assumptions:

1. The number of decoded FCDs satisfies $T = 1$ for window-sampling and $T > 1$ for forking-sequences.
2. All loss functions $\mathcal{L}_t$ at FCD $t \in \{1, \ldots, T\}$ have similar magnitudes: $\left\| \frac{\partial \mathcal{L}_t}{\partial h_t} \right\| \approx c$ for some constant $c > 0$.
3. For the `LSTM` encoder, the spectral norm of the hidden state Jacobian satisfies $\left\| \frac{\partial h_t}{\partial h_{t-1}} \right\| \leq \gamma < 1$ for all timesteps $t \in \{1, \ldots, T\}$.

Proposition 1. (LSTM gradient preservation with forking-sequences.) For an LSTM encoder, forking-sequences with T decoded FCDs provides gradient magnitudes to early timesteps that are bounded by $\sum_{t=1}^{T} \gamma^t$ compared to γ^T for window-sampling.

Proof. For window-sampling,

$$\frac{\partial \mathcal{L}}{\partial h_0} = \frac{\partial \mathcal{L}}{\partial h_T} \frac{\partial h_T}{\partial h_{T-1}} \frac{\partial h_{T-1}}{\partial h_{T-2}} \cdots \frac{\partial h_1}{\partial h_0} \tag{18}$$

$$= \frac{\partial \mathcal{L}}{\partial h_T} \prod_{j=0}^{T-1} \frac{\partial h_{T-j}}{\partial h_{T-j-1}}. \tag{19}$$

Taking norms and applying the bound $\left\| \frac{\partial h_j}{\partial h_{j-1}} \right\| \leq \gamma$, the gradient magnitude is bounded by

$$\left\| \frac{\partial \mathcal{L}}{\partial h_0} \right\| \leq \gamma^T \left\| \frac{\partial \mathcal{L}}{\partial h_T} \right\|. \tag{20}$$

For forking-sequences, the total loss is

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t, \tag{21}$$

where $\mathcal{L}_t$ is the loss from decoding at FCD t. We omit standard batch and hidden dimension inverse scaling factors (see Equation 10) as they are shared across forking-sequences and window-sampling.

The gradient to h_0 is computed via the chain rule through each decoding path:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial h_0} = \frac{1}{T} \Bigg(& \frac{\partial \mathcal{L}_T}{\partial h_T} \frac{\partial h_T}{\partial h_{T-1}} \frac{\partial h_{T-1}}{\partial h_{T-2}} \cdots \frac{\partial h_1}{\partial h_0} \\ & + \frac{\partial \mathcal{L}_{T-1}}{\partial h_{T-1}} \frac{\partial h_{T-1}}{\partial h_{T-2}} \frac{\partial h_{T-2}}{\partial h_{T-3}} \cdots \frac{\partial h_1}{\partial h_0} \\ & \vdots \\ & + \frac{\partial \mathcal{L}_1}{\partial h_1} \frac{\partial h_1}{\partial h_0} \Bigg) \end{aligned} \tag{22}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t}{\partial h_t} \prod_{j=0}^{t-1} \frac{\partial h_{t-j}}{\partial h_{t-j-1}}. \tag{23}$$

Taking norms and applying the bound $\left\|\frac{\partial h_j}{\partial h_{j-1}}\right\| \leq \gamma$, the gradient magnitude is bounded by

$$\left\|\frac{\partial \mathcal{L}}{\partial h_0}\right\| \leq \frac{1}{T}\sum_{t=1}^{T}\gamma^t \left\|\frac{\partial \mathcal{L}_t}{\partial h_t}\right\|. \tag{24}$$

□

Proposition 2. (Transformer gradient scaling with forking-sequences.) Unlike LSTMs, Transformers do not suffer from vanishing gradients, so forking-sequences provides only a linear scaling factor of the number of FCDs T rather than the exponential improvement observed in LSTMs. When losses are averaged by a factor of $\frac{1}{T}$, forking-sequences yields gradient magnitudes comparable to window-sampling.

Proof. In Transformer, each token attends to all tokens with self-attention. Let us assume each timestep t is an individual token. The hidden state at t is computed as

$$h_t = \text{Attention}(h_0, h_1, \ldots, h_T). \tag{25}$$

This attention mechanism ensures that $\frac{\partial h_i}{\partial h_j}$ does not decay with distance $|i-j|$.

For window-sampling that decodes at a single FCD T,

$$\frac{\partial \mathcal{L}}{\partial h_0} = \frac{\partial \mathcal{L}}{\partial h_T}\frac{\partial h_T}{\partial h_0}, \tag{26}$$

$$\tag{27}$$

and so the gradient magnitude is bounded by

$$\left\|\frac{\partial \mathcal{L}}{\partial h_0}\right\| \leq \left\|\frac{\partial \mathcal{L}}{\partial h_T}\right\| \left\|\frac{\partial h_T}{\partial h_0}\right\|. \tag{28}$$

For forking-sequences, the total loss is

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t, \tag{29}$$

where $\mathcal{L}_t$ is the loss from decoding at FCD t. We omit standard batch and hidden dimension inverse scaling factors (see Equation 10) as they are shared across forking-sequences and window-sampling methods.

The gradient to h_0 receives contributions from all decoded FCDs:

$$\frac{\partial \mathcal{L}}{\partial h_0} = \frac{1}{T}\left(\frac{\partial \mathcal{L}_T}{\partial h_T}\frac{\partial h_T}{\partial h_0} + \frac{\partial \mathcal{L}_{T-1}}{\partial h_{T-1}}\frac{\partial h_{T-1}}{\partial h_0} + \cdots + \frac{\partial \mathcal{L}_1}{\partial h_1}\frac{\partial h_1}{\partial h_0}\right) \tag{30}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathcal{L}_t}{\partial h_t}\frac{\partial h_t}{\partial h_0}. \tag{31}$$

$$\tag{32}$$

The gradient magnitude is bounded by

$$\left\|\frac{\partial \mathcal{L}}{\partial h_0}\right\| \leq \frac{1}{T}\sum_{t=1}^{T}\left\|\frac{\partial \mathcal{L}_t}{\partial h_t}\right\| \left\|\frac{\partial h_t}{\partial h_0}\right\|. \tag{33}$$

□

Unlike LSTMs, where forking-sequences provides exponential improvement over window-sampling by mitigating vanishing gradients, Transformers do not suffer from vanishing gradients. Averaging the losses by $\frac{1}{T}$ for forking-sequences compensates for the linear accumulation from multiple gradient paths, resulting in similar gradient magnitudes at early timesteps for both Transformers with forking-sequences and window-sampling. This result is demonstrated empirically in Fig. 10.

C Forking-Sequences Convergence Speed Ablation Study

C.1 Linear Model Experiment

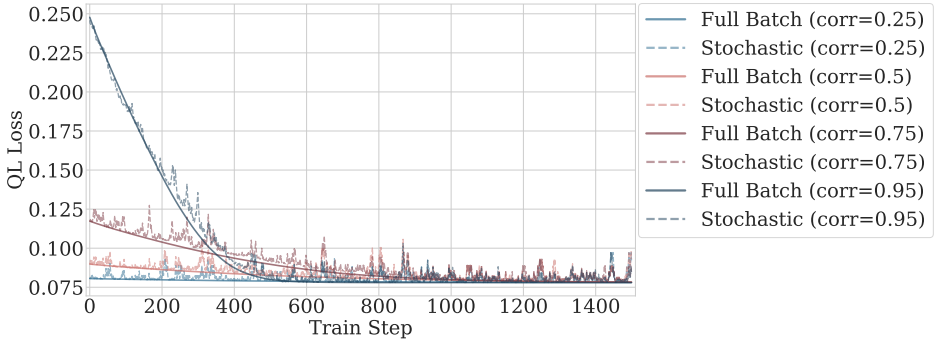


Figure 11: Training dynamics comparison between full-batch and SGD approaches across different autocorrelation levels. Full-batch training exhibits faster and more stable convergence compared to SGD across all scenarios. Models trained on highly autocorrelated time series ($\rho = 0.95$) show higher initial training loss but demonstrate steeper convergence rates, compared to those with lower autocorrelation ($\rho = 0.25$). This behavior illustrates how temporal dependencies influence optimization dynamics, with implications for batch processing strategies in time series forecasting.

We conducted an ablation study comparing full-batch gradient descent (GD) versus stochastic gradient descent (SGD) training approaches in time series forecasting. This comparison provides insights relevant to forking sequences implementations, as the key distinction between batch sizes during training (all samples vs. single sample) parallels the training schemes of forking-sequences and window-sampling, respectively.

The experimental setup involved four scenarios with varying temporal dependency levels (autocorrelation coefficients $\rho \in 0.25, 0.5, 0.75, 0.95$) following the data-generating process:

$$X_t = \rho X_{t-1} + \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{34}$$

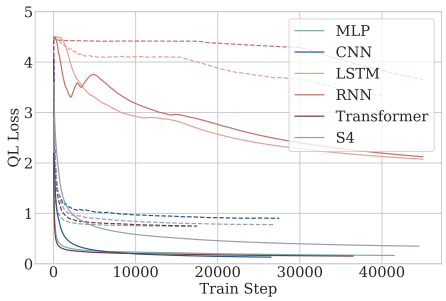
Higher autocorrelation values (ρ) indicate stronger temporal dependencies between consecutive time steps. Training was conducted over 1,500 iterations, with full-batch gradient descent processing all 1,000 samples simultaneously, while SGD processed one sample per iteration. This experimental design helps understand the impact of batch size on training dynamics across different temporal dependency structures. Experimental parameters are included in Table 4.

Table 4: Parameters for Linear Autoregressive Model.

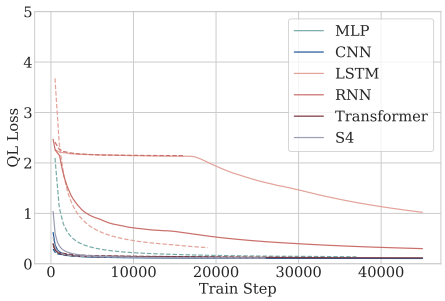
Horizon	1
Input size (lag)	1
Batch size (number of time series)	1
Random seed	42
Maximum Train Steps	1,500
Loss quantiles	0.5
Learning rate	0.01
Sample Size	T (Full) or 1 (SGD)
Number of samples (T)	1,000
Noise level	0.1
Autocorrelation levels (ρ)	{0.25, 0.5, 0.75, 0.95}
Loss function	MAE

C.2 Deep Learning Model Experiments

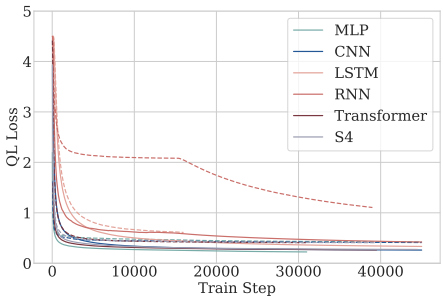
The use of forking-sequences leads to faster training convergence across all models and datasets compared to window-sampling, as observed in Figure 12.



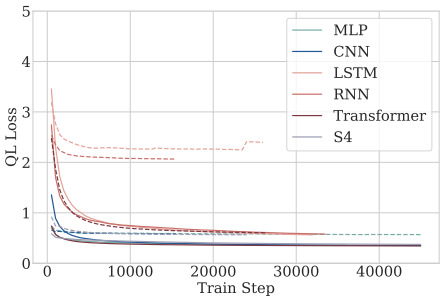
(a) Hourly Frequency Data



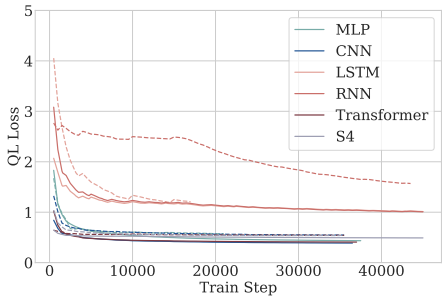
(b) Daily Frequency Data



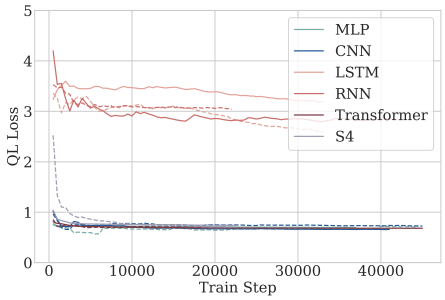
(c) Weekly Frequency Data



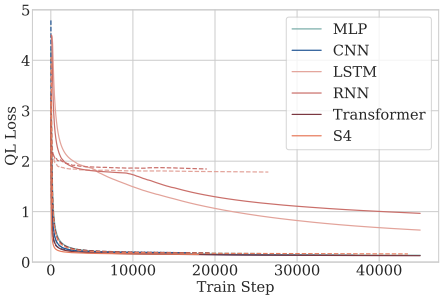
(d) Monthly Frequency Data



(e) Quarterly Frequency Data



(f) Yearly Frequency Data



(g) Other Frequency Data

Figure 12: Deep Learning optimization convergence using either *forking-sequences* (solid) or *window-sampling* (dashed) techniques. Figures show training quantile loss for the train set as a function of train step.

D Theoretical Encoder Computational Complexity

We provide propositions and supporting proofs for various encoder types to supplement the results presented in Table 1. We begin with the following assumptions:

1. The input is a univariate time series $\mathbf{y} \in \mathbb{R}$.
2. Hidden sizes (d for RNNs, d for CNN output channels, d_1, d_2 for MLP layers) and kernel sizes k are treated as fixed constants, independent of T.
3. Window-sampling (WS)-restricted uses a fixed window size $L < T$. WS-full uses T timesteps of history.
4. For forking-sequences (FS), the encoder processes the entire sequence of input length T.

D.1 Multi-Layer Perceptrons (MLPs)

Proposition 3 (Complexity of MLP Encoders)**.** For an `MLP` encoder with fixed hidden dimensions and window size L, the computational complexity scales as follows: FS: $\mathcal{O}(T)$, WS-restricted: $\mathcal{O}(TL)$, and WS-full: $\mathcal{O}(T^2)$.

Proof. Given a 2-layer `MLP`,

$$\mathbf{h}_{t,[h]} = \sigma(W_2\sigma(W_1\mathbf{y}_{t-L+1:t} + b_1) + b_2) \tag{35}$$

where $\mathbf{y}_{t-L+1:t} \in \mathbb{R}^L$ is a window of L consecutive observations, $W_1 \in \mathbb{R}^{d_1 \times L}$ flattens and projects the window to d_1 dimensions, $W_2 \in \mathbb{R}^{d_2 \times d_1}$ projects to d_2 dimensions, and $\mathbf{h}_{t,[h]} \in \mathbb{R}$ is the encoder output (hidden state) at forecast creation date t for forecast horizon h.

For a single encoder computation on a window of size L, the complexity is $\mathcal{O}(d_1L + d_2d_1)$. Processing T FCDs yields total complexity $\mathcal{O}(T(d_1L + d_2d_1))$.

Treating L, d_1, and d_2 as fixed architectural constants, the complexity for FS is $\mathcal{O}(T)$. Similarly, for WS-restricted, since each encoder computation processes a window of size L, the complexity is $\mathcal{O}(TL)$.

For WS-full, each encoder computation processes a window of size up to T, where $W_1 \in \mathbb{R}^{d_1 \times T}$. The cost per computation becomes $\mathcal{O}(d_1T)$, and with T encoder computations, the total complexity is $\mathcal{O}(T \cdot d_1T) = \mathcal{O}(T^2)$. □

D.2 Convolutional Encoders

Proposition 4 (Complexity of Convolutional Encoders)**.** For a convolutional encoder with fixed kernel size k and channel dimension, the computational complexity scales as follows: FS: $\mathcal{O}(T)$, WS-restricted: $\mathcal{O}(TL)$, WS-full: $\mathcal{O}(T^2)$

Proof. Given a 1D `CNN` encoder with kernel size k and a single output channel, a single convolution operation at timestep t produces the encoder output,

$$\mathbf{h}_{t,[h]} = \sigma\Big(W * \mathbf{y}_{t-k+1} + b\Big), \tag{36}$$

where $W \in \mathbb{R}^{1 \times k}$ is the convolutional kernel, $b \in \mathbb{R}$ is the bias, $*$ denotes 1D convolution, $\sigma(\cdot)$ is a nonlinearity, and $\mathbf{h}_{t,[h]} \in \mathbb{R}$ is the encoder output (hidden state) at forecast creation date t for forecast horizon h.

Processing the entire sequence of length T requires applying the convolution T times with kernel size k, yielding total complexity $\mathcal{O}(T \cdot k \cdot d) = \mathcal{O}(T)$, treating k and d as fixed constants.

For WS-restricted, each encoder evaluation sees a fixed window of length L, so the convolution is applied L times per evaluation. With T evaluations, the total complexity is $\mathcal{O}(T \cdot L \cdot k \cdot d) = \mathcal{O}(TL)$.

For WS-full, each encoder process processes a window of size up to T, giving complexity of $\mathcal{O}(T \cdot k \cdot d)$. With T processes of the encoder, the overall complexity is $\mathcal{O}(T \cdot T \cdot k \cdot d) = \mathcal{O}(T^2)$. □

D.3 Recurrent Encoders

Proposition 5 (Complexity of RNN Encoders)**.** For a recurrent encoder with hidden size d treated as constant, the complexity is FS: $\mathcal{O}(T)$, WS-restricted: $\mathcal{O}(TL)$, and WS-full: $\mathcal{O}(T^2)$.

Proof. Considering an `LSTM` encoder, which processes input sequentially via:

$$\mathbf{i}_t = \sigma(W_{ii}y_t + \mathbf{b}_{ii} + W_{hi}\mathbf{h}_{t-1} + \mathbf{b}_{hi}) \quad \text{(input gate)} \tag{37}$$

$$\mathbf{f}_t = \sigma(W_{if}y_t + \mathbf{b}_{if} + W_{hf}\mathbf{h}_{t-1} + \mathbf{b}_{hf}) \quad \text{(forget gate)} \tag{38}$$

$$\mathbf{g}_t = \tanh(W_{ig}y_t + \mathbf{b}_{ig} + W_{hg}\mathbf{h}_{t-1} + \mathbf{b}_{hg}) \quad \text{(candidate cell state)} \tag{39}$$

$$\mathbf{o}_t = \sigma(W_{io}y_t + \mathbf{b}_{io} + W_{ho}\mathbf{h}_{t-1} + \mathbf{b}_{ho}) \quad \text{(output gate)} \tag{40}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \quad \text{(cell state update)} \tag{41}$$

$$\mathbf{h}_{t,[h]} = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad \text{(hidden state)}, \tag{42}$$

where $y_t \in \mathbb{R}$, $\mathbf{h}_{t,[h]} \in \mathbb{R}^d$, $\mathbf{c}_t \in \mathbb{R}^d$, $\mathbf{h}_{t-1}$ is the hidden state of the layer at time t-1 or the initial hidden state at time 0, σ is the sigmoid function, and $\odot$ is the Hadamard product.

At each timestep, the LSTM computes four gates, each requiring matrix-vector multiplications costing $\mathcal{O}(d^2)$ operations. Treating hidden dimension d as constant, processing each timestep costs $\mathcal{O}(1)$. □

With forking-sequence, the recurrent encoder consumes the sequence sequentially in a single forward pass, resulting in $\mathcal{O}(T)$ total computation. However, WS-restricted requires recomputing encoder states over length-L sequences at each timestep, incurring $\mathcal{O}(TL)$ total complexity. Similarly, WS-full recomputes over the entire history T, leading to $\mathcal{O}(T^2)$

D.4 Self-Attention (Transformers)

Proposition 6 (Complexity of Transformer Encoders)**.** For a self-attention encoder with hidden dimension d, the computational complexity scales as FS: $\mathcal{O}(T^2)$, WS-restricted: $\mathcal{O}(TL^2)$, and WS-full: $\mathcal{O}(T^3)$.

Proof. Self-attention computes attention scores between all pairs of positions:

$$\mathbf{h}_{t,[h]} = \text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{43}$$

where $Q, K, V \in \mathbb{R}^{T\times d}$ are the query, key, and value matrices, and $\mathbf{h}_{t,[h]} \in \mathbb{R}^d$ is the encoder output (hidden state) at forecast creation date t for forecast horizon h. The dominant operation is computing QK^T, which costs $\mathcal{O}(T^2d)$. Since d is constant, the overall complexity of self-attention for FS is $\mathcal{O}(T^2)$.

For WS-restricted, self-attention is applied within each window of size L, costing $\mathcal{O}(L^2)$ per window. With T encoder processes, the total complexity is $\mathcal{O}(TL^2)$. The computation of attention matrices of $\mathcal{O}(T^2)$ or $\mathcal{O}(L^2)$ can be recycled with the use of causal masks without additional computation.

If implemented in a window-based sampling manner, each subsequence would require its own self-attention pass. In the optimistic case—where attention is applied only to each progressively shorter subsequence; the total cost is $T^2 + (T-1)^2 + \cdots + 1^2 = \mathcal{O}(T^3)$. In practice, shorter windows are typically padded to a common length, meaning each of the T windows incurs a full T^2 attention cost, again yielding an overall complexity of $\mathcal{O}(T^3)$. In other words, WS-full requires T encoder computations over windows of average size $\mathcal{O}(T)$, and because self-attention on a window of length T is $\mathcal{O}(T^2)$, the total encoder cost remains cubic in T.

□

D.5 Structured State Space (S4)

Proposition 7 (Complexity of S4 Encoders)**.** For a structured state space encoder with state dimension N and hidden dimension d, the computational complexity scales as FS: $\mathcal{O}(T \log T)$, WS-restricted: $\mathcal{O}(TL \log L)$, and WS-full: $\mathcal{O}(T^2 \log T)$.

Proof. The S4 model processes input through a continuous-time state space:

$$x'(t) = Ax(t) + Bu(t) \tag{44}$$

$$y(t) = Cx(t) + Du(t) \tag{45}$$

which is discretized with step size Δ.

S4 computes the output as a convolution $\mathbf{h}_{t,[h]} = (\bar{K} * u) + Du$, where $\mathbf{h}_{t,[h]} \in \mathbb{R}^d$ is the encoder output (hidden state) at forecast creation date t for forecast horizon h, and the kernel $\bar{K} \in \mathbb{R}^T$ is:

$$\bar{K}_t = C \cdot \frac{e^{\Delta A} - 1}{A} \cdot e^{t\Delta A} \tag{46}$$

Computing the kernel for T timesteps requires $\mathcal{O}(TN)$ operations. The convolution is computed via FFT in $\mathcal{O}(T \log T)$ time. Treating N as constant, the FFT dominates, giving $\mathcal{O}(T \log T)$ complexity.

For FS, the encoder processes the entire sequence of length T once, so the kernel computation $\mathcal{O}(T)$ plus FFT-based convolution $\mathcal{O}(T \log T)$ yields total complexity $\mathcal{O}(T \log T)$.

For WS-restricted, each encoder computation processes a fixed window of size L, costing $\mathcal{O}(L \log L)$ per window. With T encoder computations to generate forecasts at each FCD, the total complexity is $\mathcal{O}(TL \log L)$.

For WS-full, each encoder computation processes a window of size up to T, costing $\mathcal{O}(T \log T)$ per computation. With T encoder computations, the total complexity is $\mathcal{O}(T^2 \log T)$. □

E Excess Volatility Metric

Recall the *Excess Volatility* (EV) definition in Equation 12 from Section 3.6, here we provide the proof of its convenient properties.

$$\mathrm{EV}(y,\ \hat{\mathbf{y}}_1,\ \hat{\mathbf{y}}_2) = \mathrm{QL}(\hat{\mathbf{y}}_2, \hat{\mathbf{y}}_1) - (\mathrm{QL}(y, \hat{\mathbf{y}}_1) - \mathrm{QL}(y, \hat{\mathbf{y}}_2)). \tag{47}$$

Theorem 2. (*Excess Volatility Metric Properties*) Consider ground truth y, and two consecutive forecasts $\hat{\mathbf{y}}_1$, $\hat{\mathbf{y}}_2$, we prove in Appendix E the EV metric has the following properties:

- **Positivity:** $\mathrm{EV}(y,\ \hat{\mathbf{y}}_1,\ \hat{\mathbf{y}}_2) \geq 0$
- **Zero-Penalty for Improving Revisions:** When the revision from $\hat{\mathbf{y}}_1$ to $\hat{\mathbf{y}}_2$ represents a direct, proportional improvement, and $\hat{\mathbf{y}}_2$ lies on the linear path from y to $\hat{\mathbf{y}}_1$ then $\mathrm{EV}(y,\ \hat{\mathbf{y}}_1,\ \hat{\mathbf{y}}_2) = 0$. An improved revision example is shown in Fig. 6a.
- **Maximum penalty for Deteriorating Revisions:** When $\hat{\mathbf{y}}_1$ lies on the direct path between y and $\hat{\mathbf{y}}_2$; that is, when the revision from $\hat{\mathbf{y}}_1$ to $\hat{\mathbf{y}}_2$ moves linearly away from the truth, then $\mathrm{EV}(y,\ \hat{\mathbf{y}}_1,\ \hat{\mathbf{y}}_2) = \mathrm{QL}(y, \hat{\mathbf{y}}_2) - \mathrm{QL}(y, \hat{\mathbf{y}}_1)$. In this fully misdirected case, EV assigns the maximum penalty, equal to the entire accuracy degradation. A deteriorating revision example is shown in Fig. 6b.
- **Overshoot-Revision Penalty Property:** When y lies on the line segment between $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$—that is, when $\hat{\mathbf{y}}_2$ revises $\hat{\mathbf{y}}_1$ in the correct direction but overshoots—then $\mathrm{EV} = \mathrm{QL}(y, \hat{\mathbf{y}}_2)$. In this excess-revision case, EV penalizes only the overshoot. An overshoot revision example is shown in Fig. 6c.

Proof. All the above properties can be proved by cases, as shown in Table 5 below:

Case	Sub-case	$L_q(y_2, y_1)$	$L_q(y, y_1)$	$L_q(y, y_2)$	EV
$\lvert y - y_1\rvert = \lvert y - y_2\rvert + \lvert y_2 - y_1\rvert$	$y \leq y_2 \leq y_1$	$(1-q)(y_1 - y_2)$	$(1-q)(y_1 - y)$	$(1-q)(y_2 - y)$	0
	$y_1 \leq y_2 \leq y$	$q(y_2 - y_1)$	$q(y - y_1)$	$q(y - y_2)$	0
$\lvert y - y_2\rvert = \lvert y - y_1\rvert + \lvert y_1 - y_2\rvert$	$y \leq y_1 \leq y_2$	$q(y_2 - y_1)$	$(1-q)(y_1 - y)$	$(1-q)(y_2 - y)$	$y_2 - y_1$
	$y_2 \leq y_1 \leq y$	$(1-q)(y_1 - y_2)$	$q(y - y_1)$	$q(y - y_2)$	$y_1 - y_2$
$\lvert y_1 - y_2\rvert = \lvert y_1 - y\rvert + \lvert y - y_2\rvert$	$y_1 \leq y \leq y_2$	$q(y_2 - y_1)$	$q(y - y_1)$	$(1-q)(y_2 - y)$	$y_2 - y$
	$y_2 \leq y \leq y_1$	$(1-q)(y_1 - y_2)$	$(1-q)(y_1 - y)$	$q(y - y_2)$	$y - y_2$

Table 5: Proof by cases.

□

F Dataset Details

In this Appendix, we describe the datasets used in our analysis.

Table 6: Summary of forecasting datasets used in our empirical study.

	Frequency	Seasonality	Horizon	Series	Min Length	Max Length	% Erratic
M1	Monthly	12	18	617	48	150	0
	Quarterly	4	8	203	18	114	0
	Yearly	1	6	181	15	58	0
M3	Other	4	8	174	71	104	0
	Monthly	12	18	1428	66	144	2
	Quarterly	4	8	756	24	72	1
	Yearly	1	6	645	20	47	10
M4	Hourly	24	48	414	748	1008	17
	Daily	1	14	4,227	107	9933	2
	Weekly	1	13	359	93	2610	16
	Monthly	12	18	48,000	60	2812	6
	Quarterly	4	8	24,000	24	874	11
	Yearly	1	6	23,000	19	841	18
Tourism	Monthly	12	18	366	91	333	51
	Quarterly	4	8	427	30	130	39
	Yearly	1	6	518	11	47	23

M1 Dataset Details. The early M1competition (Makridakis et al., 1982) focused on 1,001 time series drawn from demography, industry, and economics, with lengths ranging from 9 to 132 observations and varying in frequency (monthly, quarterly, and yearly). A key empirical finding of this competition was that simple forecasting methods, such as ETS (Holt, 1957), often outperformed more complex approaches. These results had a lasting impact on the field, initiating a research legacy that emphasized accurate forecasting, model automation, and caution against overfitting. The competition also marked a conceptual shift, helping to distinguish time-series forecasting from traditional time series analysis.

M3 Dataset Details. The M3competition (Makridakis and Hibon, 2000), held two decades after the M1competition, featured a dataset of 3,003 time series spanning business, demography, finance and economics. These series ranged from 14 to 126 observations and included monthly, quarterly, and yearly frequencies. All series had positive values, with only a small proportion displaying erratic behavior and none exhibiting intermittency (Syntetos et al., 2005). The M3competition reinforced the trend of simple forecasting methods outperforming more complex alternatives, with the `Theta` method (Hyndman and Billah, 2003) emerging as the best performing approach.

M4 Dataset Details. The M4competition marked a substantial increase in both the size and diversity of the M competition datasets, comprising 100,000 time series across six frequencies: hourly, daily, weekly, monthly, quarterly, and annual. These series covered a wide range of domains, including demography, finance, industry, and both micro- and macroeconomic indicators. The competition also introduced the evaluation of prediction intervals in addition to point forecasts, broadening the assessment criteria. M4's proportion of non-smooth or erratic time series increased to 18 percent (Syntetos et al., 2005). For the first time, a neural forecasting model - ESRNN(Smyl, 2019) - outperformed traditional methods.

Tourism Dataset Details. The Tourism dataset (Athanasopoulos et al., 2011) was designed to evaluate forecasting methods applied to tourism demand data across multiple temporal frequencies. It comprises 1,311 time series at monthly, quarterly, and yearly frequencies. This competition introduced the Mean Absolute Scaled Error (MASE) as an alternative metric to evaluate scaled point forecasts, alongside the evaluation of forecast intervals. Notably, 36% of the series were classified as erratic or intermittent. Due to this high proportion of irregular data, the Naïve1 method proved particularly difficult to outperform at the yearly frequency.

G Training Methodology and Hyperparameters

In this Appendix, we expand on the training methodology outlined in Section 4. We train multi-quantile loss (MQ) models varying only the encoder type to include `MLP`, `RNN`, `LSTM`, `CNN`, `Transformer`, and `S4` modules. We train models with hyperparameter tuning using Optuna. Using 5 trials, we train individual models and select hyperparameters that minimize validation loss using the multi-quantile loss function defined in Eqn. 5. Tables 7 and 8–13 list the shared and encoder-specific hyperparameters, respectively. For our implementations of `NBEATS` (Oreshkin et al., 2020) and `PatchTST` (Nie et al., 2023) pre-trained on a synthetic dataset, model parameters are set to the defaults of the original implementations in the NeuralForecast library (Olivares et al., 2022). We conducted all neural network experiments using a single AWS p4d.24xlarge with 1152 GiB of RAM and 96 vCPUs. Training times mostly depend on the architecture; however, we restrict the SGD training steps to 45K per architecture.

Table 7: Shared hyperparameters across encoder types

Hyperparameter	Considered Values
Single GPU SGD Batch Size*	{4, 8, 16}
Initial learning rate	loguniform(1e-4, 5e-2)
Maximum Training steps S_{max}	45,000
Learning rate decay	0.1
Learning rate step size	15,000
Scaler type	Standard
H-Agnostic Decoder Dimension	100
H-Specific Decoder Dimension	20

Table 8: `MLP`

Hyperparameter	Considered Values
Main Activation Function	ReLU
Encoder Dimension	{128, 256}
Number of Layers	{3, 4}
Input size.	{1, 8, 32, 96}

Table 9: `RNN`

Hyperparameter	Considered Values
Main Activation Function	tanh
Encoder Dimension	{128, 256}
Dilations	{[[1,2], [4,8]], [[1,2], [4,8], [16,32]]}

Table 10: `LSTM`

Hyperparameter	Considered Values
Main Activation Functions	Sigmoid, tanh
Encoder Dimension	{128, 256}
Dilations	{[[1,2], [4,8]], [[1,2], [4,8], [16,32]]}

Table 11: `CNN`

Hyperparameter	Considered Values
Main Activation Function	ReLU
Encoder Dimension	{128, 256}
Temporal Convolution Kernel Size	{2, 5}
Temporal Convolution Dilations	{[1, 2, 4, 8], [1, 2, 4, 8, 16, 32]}

Table 12: `Transformer`

Hyperparameter	Considered Values
Main Activation Functions	Sigmoid, tanh
Encoder Dimension	{128, 256}
Projection Dimension	{128, 256}
Input Patch Lengths	{1, 8, 32, 96}
Attention Dilations	{[2, 6, 8], [2, 4, 8, 16]}
Attention Dropout	0.1
Number of Attention Heads	4

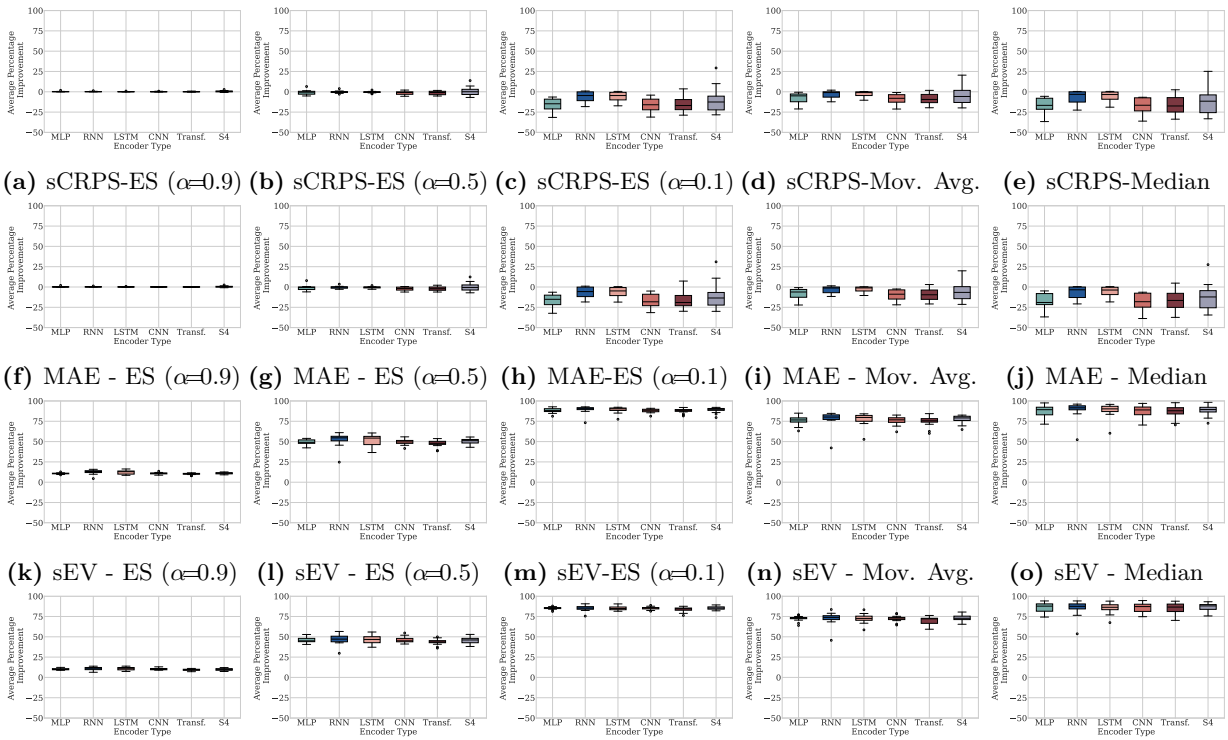
Table 13: State Space `S4`

Hyperparameter	Values
Main Activation Function	GELU
Encoder Dimension	{128, 256}
Projection Hidden Size	{128, 256}

H Ensembling Ablation Studies

We propose the use of ensembling techniques, which aggregate predictions for each target date across FCDs to reduce forecast variance, for model inference with the forking-sequences scheme.

H.1 Forecast Ensembling Impact on Trained Model Predictions



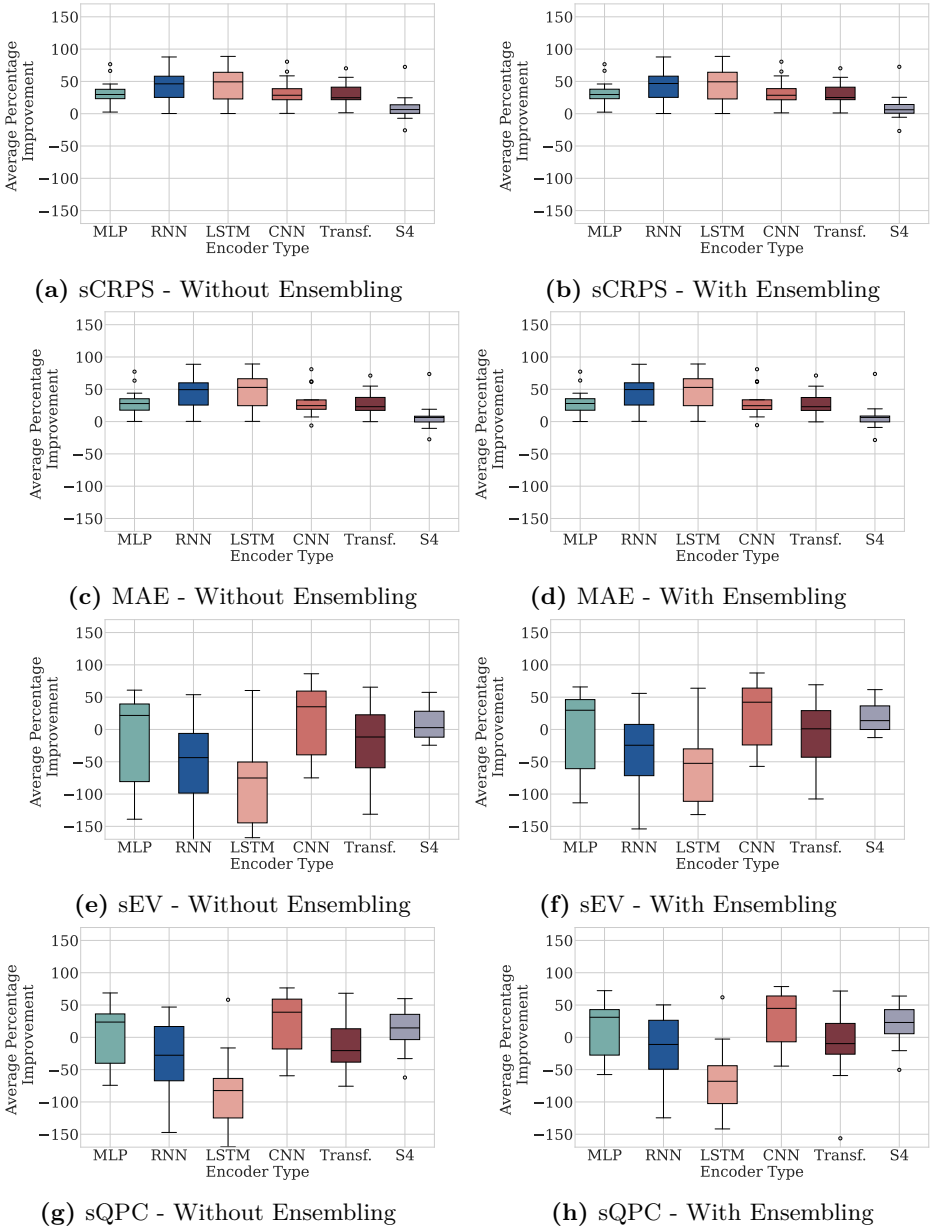
(a) sCRPS-ES (α=0.9) **(b)** sCRPS-ES (α=0.5) **(c)** sCRPS-ES (α=0.1) **(d)** sCRPS-Mov. Avg. **(e)** sCRPS-Median

(f) MAE - ES (α=0.9) **(g)** MAE - ES (α=0.5) **(h)** MAE-ES (α=0.1) **(i)** MAE - Mov. Avg. **(j)** MAE - Median

(k) sEV - ES (α=0.9) **(l)** sEV - ES (α=0.5) **(m)** sEV-ES (α=0.1) **(n)** sEV - Mov. Avg. **(o)** sEV - Median

(p) sQPC - ES (α=0.9) **(q)** sQPC - ES (α=0.5) **(r)** sQPC-ES (α=0.1) **(s)** sQPC - Mov. Avg. **(t)** sQPC - Median

Figure 13: Comparative performance gains achieved through diverse ensembling techniques applied to `MLP`, `RNN`, `LSTM`, `CNN`, `Transformer` (Transf.), and `State Space` (S4) models trained with forking-sequences, expressed as percentage improvements over baseline (no ensembling). Results are aggregated across all evaluated datasets. The ensembling strategies include moving average (Mov. Avg.), moving median functions, and exponential smoothing (ES) with varying smoothing parameters (α). Larger α parameter values assigns higher weights to predictions where the target date appears earlier in the forecast horizon.



(a) sCRPS - Without Ensembling

(b) sCRPS - With Ensembling

(c) MAE - Without Ensembling

(d) MAE - With Ensembling

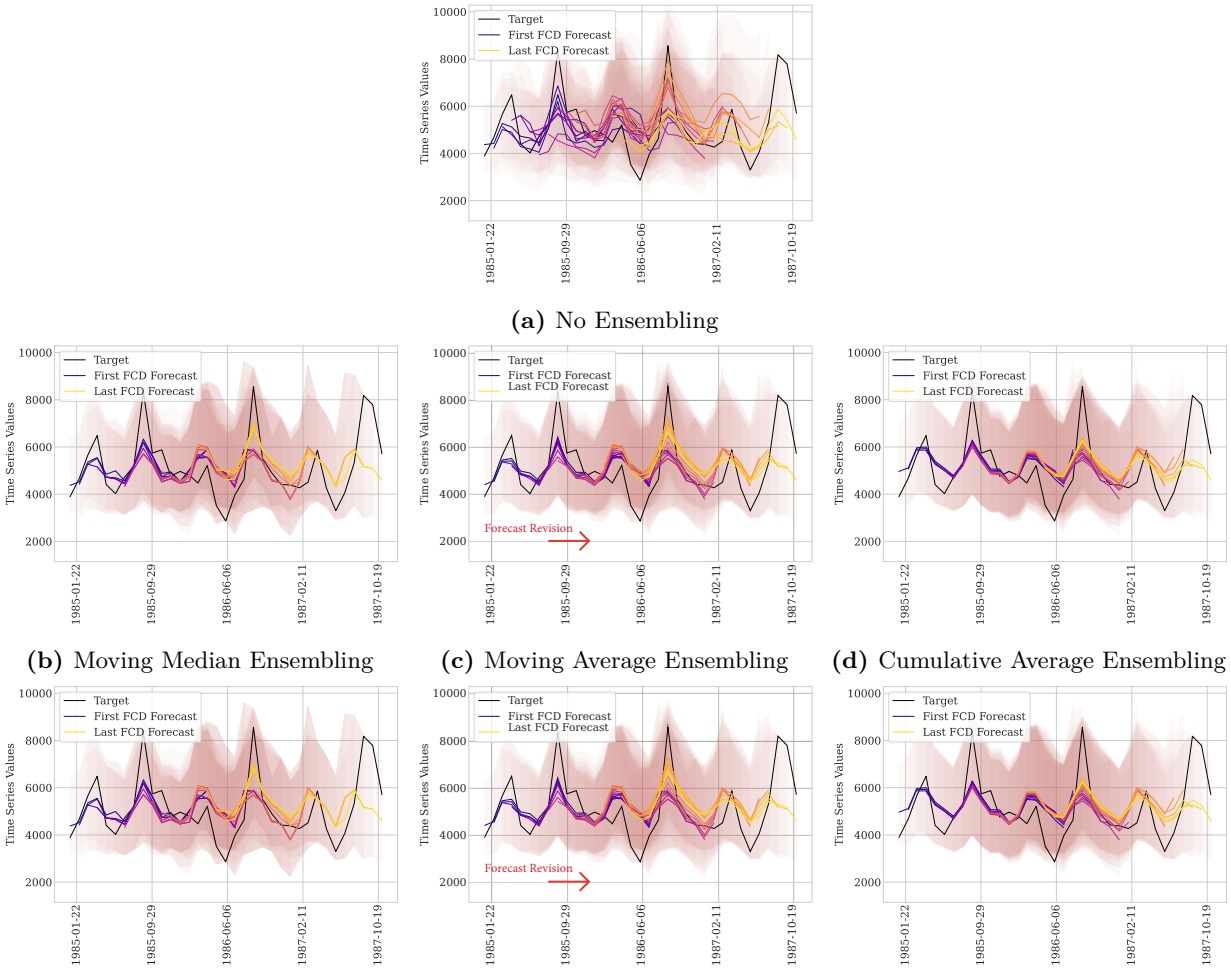
(e) sEV - Without Ensembling

(f) sEV - With Ensembling

(g) sQPC - Without Ensembling

(h) sQPC - With Ensembling

Figure 14: Percentage improvement in sCRPS, MAE, sEV, and sQPC metrics for models with forking-sequences compared with the *window-sampling* scheme both without ensembling during inference, averaged across datasets **(a, c, e, g)**. Percentage improvement in sCRPS, MAE, sEV, and sQPC metrics for models with the *forking-sequences* scheme and exponential smoothing ($\alpha = 0.9$) ensembling during inference compared with the *window-sampling* scheme, averaged across datasets **(b, d, f, h)**. Results greater than zero indicate lower metric values using forking-sequences. Forking-sequences with ensembling during inference demonstrates significant improvements in sCRPS and MAE values over windows-sampling across all encoders. While forecast ensembling during inference improves forecast stability compared to no ensembling for models with forking-sequences (see Fig. 13), it does not improve sEV over window-sampling for `RNN` and `LSTM` encoders **(f)**.



(a) No Ensembling

(b) Moving Median Ensembling **(c)** Moving Average Ensembling **(d)** Cumulative Average Ensembling

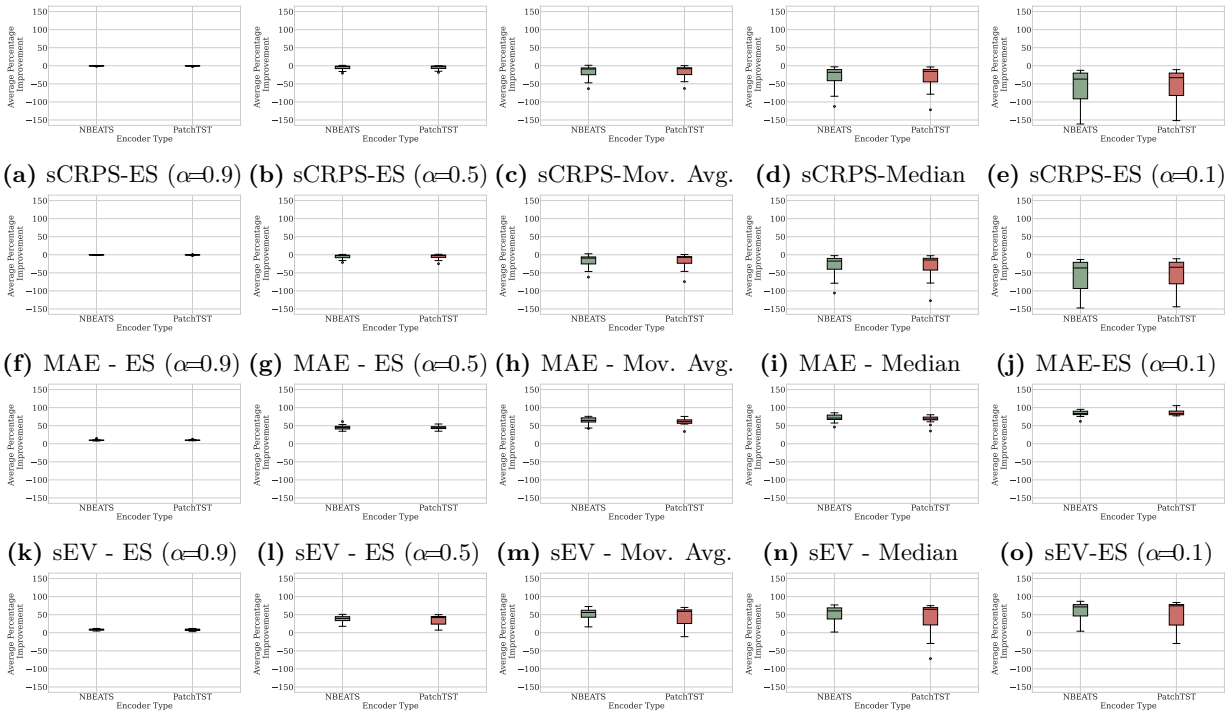
(e) Exponential Smoothing (α=0.1) **(f)** Exponential Smoothing (α=0.5) **(g)** Exponential Smoothing (α=0.9)

Figure 15: Forecast outputs for **a)** no ensembling technique and various ensembling techniques applied during inference, including **b)** moving median, **c)** moving average, and **d)** cumulative average, and **e,f,g)** exponential smoothing with various levels of smoothing $\alpha = \{0.1, 0.5, 0.9\}$.

I Forecast Ensembling Impact on Foundation Forecasting Models

To demonstrate the benefits of forecast ensembling during inference, we reimplement `PatchTST` (Nie et al., 2023) and `NBEATS` (Oreshkin et al., 2020) using the forking-sequences architectural design, while maintaining their original window-sampling for pre-training.

These models are trained on a synthetic dataset and evaluated in the cross-frequency transfer learning task, achieving performance comparable to their original implementations. We evaluate several ensembling functions as shown in Fig 16. Our analysis reveals a trade-off between forecast stability and accuracy when applying different ensembling functions. Notably, exponential smoothing with $\alpha \in \{0.5, 0.9\}$ emerges as an effective approach, successfully reducing forecast revisions and enhancing stability during inference while maintaining model accuracy. In contrast, functions like median and exponential smoothing with $\alpha = 0.1$, while achieving larger reductions in forecast revisions, do so at the cost of significant degradation in forecast accuracy. Exponential smoothing with a larger α parameter value assigns higher weights to predictions where the target date appears earlier in the forecast horizon. The smaller degradation in accuracy with larger α values aligns with the general expectation that models will make more accurate predictions for near-term versus distant forecast horizons.



(a) sCRPS-ES (α=0.9) **(b)** sCRPS-ES (α=0.5) **(c)** sCRPS-Mov. Avg. **(d)** sCRPS-Median **(e)** sCRPS-ES (α=0.1)

(f) MAE - ES (α=0.9) **(g)** MAE - ES (α=0.5) **(h)** MAE - Mov. Avg. **(i)** MAE - Median **(j)** MAE-ES (α=0.1)

(k) sEV - ES (α=0.9) **(l)** sEV - ES (α=0.5) **(m)** sEV - Mov. Avg. **(n)** sEV - Median **(o)** sEV-ES (α=0.1)

(p) sQPC - ES (α=0.9) **(q)** sQPC - ES (α=0.5) **(r)** sQPC - Mov. Avg. **(s)** sQPC - Median **(t)** sQPC-ES (α=0.1)

Figure 16: Comparative performance gains achieved through diverse ensembling techniques applied to pre-trained `NBEATS` and `PatchTST` models, expressed as percentage improvements over baseline (no ensembling). Results are aggregated across all evaluated datasets. The ensembling strategies include moving average (Mov. Avg.), moving median functions, and exponential smoothing (ES) with varying smoothing parameters (α). Larger α parameter values assigns higher weights to predictions where the target date appears earlier in the forecast horizon.

J Additional Point Forecasting Results

Table 14: Empirical evaluation of probabilistic forecasts. Mean Absolute Error (MAE) averaged over 5 runs. Lower measurements are preferred. The methods without standard deviation have deterministic solutions. For the `MQForecaster` architecture we vary the type of encoder, and the training scheme between *forking-sequences* (FS) and *window-sampling* (WS).

		MLP		RNN		LSTM		CNN		Transformer		S4		StatsForecast	
	Freq	FS	WS	FS	WS	FS	WS	FS	WS	FS	WS	FS	WS	ETS	ARIMA
M1	M	1746.8	2437.3	7156.5	9926.1	7208.6	9689.3	1768.5	2617.6	1951.4	2370.3	2911.1	2634.8	1898.7	1987.1
		(17.9)	(187.5)	(882.8)	(1204.7)	(780)	(1668.9)	(23.7)	(253.8)	(107.2)	(168.6)	(549.4)	(29.6)	(-)	(-)
	Q	2425.1	2552.4	11323.1	15452	11839.2	16863	2471	2661.3	2611.7	2821.4	2906.1	2687.2	2465.6	2745.7
		(55.8)	(98.4)	(854.3)	(591.6)	(1411.4)	(1811.4)	(121.4)	(151.2)	(44.1)	(364.2)	(219.4)	(47.3)	(-)	(-)
	Y	94477.5	140612.9	734929.5	737291	734928	737230.4	96092.8	90540.5	100019.8	116972.4	99182.7	107798.3	100188.5	96408.3
		(4321.9)	(75333.1)	(689.6)	(1190.1)	(874.5)	(1401.8)	(7279.3)	(10321.3)	(9379.2)	(17944.2)	(6952)	(11464)	(-)	(-)
M3	O	216	385.4	392.8	1381.4	699.6	2057.5	230.3	284.7	205.9	456.7	312.2	244.9	203	206.6
		(19.1)	(200.1)	(231.3)	(691.4)	(861.3)	(722.6)	(47.9)	(57.2)	(3.5)	(223.1)	(27.1)	(1.9)	(-)	(-)
	M	603	770	742.9	1565.6	709.7	1601.3	597.1	859.3	633.8	772.2	649.1	802.1	686.9	691.3
		(9.2)	(39.7)	(37.8)	(671.3)	(35.6)	(626.8)	(4.8)	(166.2)	(32.3)	(60.3)	(21.2)	(5.2)	(-)	(-)
	Q	485.3	594.4	567.4	934.9	583.3	1567.4	491.8	626.7	489.5	769.9	521.4	574.4	536.1	535.3
		(5.1)	(37.7)	(21.8)	(289.5)	(40.6)	(405)	(19.7)	(70.1)	(15)	(368.6)	(18.3)	(2.5)	(-)	(-)
	Y	972.1	1366.5	1184.5	2826.1	1256.3	2832.9	925.4	1195.2	962.9	1264.8	980.3	1053.2	1050.8	1066.8
		(16.5)	(421)	(172.2)	(1210.8)	(229.1)	(890.9)	(14.8)	(54.3)	(70.4)	(241.6)	(51.2)	(43.8)	(-)	(-)
M4	H	291.6	1283.6	2582	5233	2601	4803.5	294.3	1551.1	372.8	1292.5	409.1	1550.8	635.6	290.8
		(29.9)	(126.2)	(532.6)	(239.8)	(693.2)	(285.7)	(20.9)	(62.5)	(62.8)	(97.7)	(73.4)	(15.5)	(-)	(-)
	D	169	169.4	193.2	1711.8	196.2	1806.3	169.1	438.5	172.9	172.7	171.2	175.1	168.4	170
		(1.1)	(1.9)	(7.7)	(1426.1)	(18.4)	(1364.8)	(0.9)	(297.9)	(1.2)	(5.1)	(3.2)	(9.9)	(-)	(-)
	W	285.1	357.2	368.5	1232	394.3	1941.8	285.9	430.4	298.1	383.3	340.3	353.7	336.6	325.4
		(2.1)	(7.6)	(21.2)	(370.9)	(18.5)	(311.7)	(7)	(107.1)	(28.1)	(42.3)	(11.3)	(2.5)	(-)	(-)
	M	527.5	624.7	623.5	1814.3	602.8	1846.2	523.5	715.7	547.8	654.7	561	642.5	560.6	567.6
		(6.2)	(8.8)	(34.1)	(862.8)	(27.3)	(821.5)	(2.7)	(173.9)	(16.8)	(57.8)	(11.9)	(7.3)	(-)	(-)
	Q	558	652.3	627.3	1210.7	669.4	2505.5	565.3	695.4	572.8	862.9	597.2	634.6	568.4	597.1
		(9.4)	(39)	(14.9)	(265.7)	(54.8)	(864.9)	(16.8)	(79.3)	(18.5)	(446.9)	(16.9)	(2.3)	(-)	(-)
	Y	800.7	1236.6	1308.7	3105.4	1493.5	2997.4	774.4	967.5	806	1064	833.4	901.1	849	889.3
		(17.1)	(548.8)	(317.2)	(999.9)	(511.3)	(887.8)	(11.4)	(32)	(36)	(190.2)	(29)	(22.4)	(-)	(-)
Tourism	M	2023	5526.3	12647.1	16232.7	12492.2	15925.6	2225.7	5883	3143.6	5269.2	5514.1	5922.6	2624.5	3125.8
		(103.7)	(497.2)	(1314.7)	(1205.7)	(1234.7)	(1503.2)	(153.9)	(401.7)	(760.2)	(430.7)	(1034.8)	(15.6)	(-)	(-)
	Q	11234.5	17465.2	74238.3	84967.3	75149.3	86636.3	11642.5	17520.6	11137.1	23072.6	18446.6	17921.3	11375.6	12567.5
		(569.3)	(151.2)	(1966.1)	(774.4)	(4150.4)	(1703.6)	(1102.9)	(886.8)	(520.2)	(10670.8)	(956.5)	(271)	(-)	(-)
	Y	78055.4	107570.9	372644.4	376107.2	372762.5	375857.6	76514.3	88681.1	80648.5	98405.9	85851.7	86030.8	74574	86727.3
		(960.3)	(32072.7)	(970.1)	(1518.4)	(1302.6)	(1753.5)	(1269.9)	(1184.2)	(1368.7)	(14645.7)	(1919.1)	(1696.4)	(-)	(-)

To complement the probabilistic results in Section 4.2, we evaluate median forecasts denoted by $\hat{\mathbf{y}}_{[b][t][h]}$ through the *mean absolute error* (MAE), as described by

$$\text{MAE}\left(\mathbf{y}_{[b][t][h]},\ \hat{\mathbf{y}}_{[b][t][h]}\right) = \frac{1}{B \times T \times H} \sum_{b,t,h} |y_{b,t,h} - \hat{y}_{b,t,h}|. \tag{48}$$

The forking-sequences training scheme improved median MAE across datasets by 27.9%, 49.4%, 52.9%, 24.7%, 23.0%, and 6.4% for `MLP`, `RNN`, `LSTM`, `CNN`, `Transformer`, and `S4` encoders, respectively. These changes can be attributed to improved training convergence, induced by better gradient flow with the forking-sequences training scheme. The results are generally consistent with sCRPS outcomes in Table 2.

Table 15: Empirical evaluation of probabilistic forecasts. Mean *Symmetric Quantile Percentage Change* (sQPC) averaged over 5 runs. Lower measurements are preferred. The methods without standard deviation have deterministic solutions. For the `MQForecaster` architecture we vary the type of encoder, and the training scheme between *forking-sequences* (FS) and *window-sampling* (WS).

		MLP		RNN		LSTM		CNN		Transformer		S4		StatsForecast	
	Freq	FS	WS	FS	WS	FS	WS	FS	WS	FS	WS	FS	WS	ETS	ARIMA
M1	M	2.225	4.004	3.941	4.538	3.284	2.821	2.181	4.467	3.867	2.768	5.696	6.716	2.015	2.585
		(0.11)	(2.353)	(0.517)	(3.136)	(0.456)	(2.697)	(0.112)	(0.398)	(1.146)	(2.133)	(1.471)	(1.449)	(-)	(-)
	Q	3.335	4.181	7.262	5.198	3.615	1.792	3.276	3.779	4.604	4.829	7.689	5.784	3.636	3.898
		(0.461)	(1.59)	(6.04)	(1.876)	(0.642)	(1.146)	(0.692)	(0.922)	(0.872)	(2.259)	(1.587)	(1.569)	(-)	(-)
	Y	4.742	3.443	8.151	3.297	8.485	3.446	5.623	3.526	8.015	6.742	8.155	9.227	2.843	3.45
		(0.185)	(1.506)	(3.98)	(1.699)	(4.963)	(3.086)	(0.793)	(0.116)	(2.67)	(1.812)	(1.779)	(1.125)	(-)	(-)
M3	O	1.062	0.609	0.999	1.635	0.7	0.369	1.178	0.96	1.073	0.611	1.54	0.949	0.898	0.858
		(0.176)	(0.262)	(0.075)	(1.358)	(0.274)	(0.186)	(0.122)	(0.028)	(0.145)	(0.33)	(0.25)	(0.015)	(-)	(-)
	M	2.426	4.409	4.491	3.209	3.766	2.296	2.091	5.167	2.677	2.19	3.22	6.286	1.375	1.72
		(0.212)	(2.608)	(0.678)	(2.682)	(0.195)	(1.857)	(0.077)	(0.665)	(0.219)	(2.138)	(0.444)	(0.073)	(-)	(-)
	Q	2.12	3.189	2.732	4.103	2.503	1.542	2.135	3.164	1.87	3.042	2.559	3.708	1.859	2.194
		(0.086)	(1.195)	(0.219)	(0.606)	(0.384)	(1.412)	(0.169)	(0.479)	(0.196)	(1.318)	(0.209)	(0.052)	(-)	(-)
	Y	5.307	3.602	4.629	2.342	4.264	1.936	4.908	4.212	4.899	3.597	4.696	4.521	3.489	5.001
		(0.183)	(1.407)	(0.761)	(0.764)	(1.031)	(0.674)	(0.159)	(0.238)	(0.326)	(1.238)	(0.096)	(0.073)	(-)	(-)
M4	H	2.33	1.845	3.082	5.799	1.666	3.98	1.961	5.083	2.993	1.051	2.109	5.259	2.488	0.786
		(0.428)	(2.089)	(0.708)	(1.68)	(0.408)	(1.962)	(0.318)	(0.601)	(1.718)	(0.952)	(0.058)	(0.339)	(-)	(-)
	D	0.45	0.477	0.511	0.312	0.492	0.183	0.457	1.942	0.388	0.476	0.475	0.477	0.497	0.499
		(0.034)	(0.001)	(0.065)	(0.14)	(0.065)	(0.087)	(0.023)	(2.827)	(0.048)	(0.0004)	(0.001)	(0.001)	(-)	(-)
	W	1.387	2.083	1.746	2.359	1.594	0.947	1.329	2.575	1.469	1.66	2.11	2.452	1.494	1.14
		(0.079)	(0.743)	(0.174)	(0.451)	(0.076)	(0.315)	(0.1)	(0.505)	(0.491)	(0.997)	(0.238)	(0.003)	(-)	(-)
	M	1.934	2.767	2.705	1.834	2.249	1.323	1.776	3.245	1.992	1.465	2.362	3.806	1.858	2.232
		(0.081)	(1.491)	(0.33)	(1.521)	(0.092)	(0.964)	(0.033)	(0.374)	(0.106)	(1.233)	(0.205)	(0.049)	(-)	(-)
	Q	2.339	3.204	2.917	3.379	2.758	1.157	2.271	3.142	2.156	3.039	2.837	3.689	2.047	2.246
		(0.062)	(1.145)	(0.154)	(0.313)	(0.317)	(0.797)	(0.164)	(0.454)	(0.212)	(1.298)	(0.192)	(0.054)	(-)	(-)
	Y	4.493	2.937	3.77	2.11	3.287	1.873	4.194	3.423	4.127	2.949	3.832	3.703	3.082	4.532
		(0.16)	(1.017)	(0.411)	(0.492)	(0.412)	(0.648)	(0.044)	(0.214)	(0.355)	(0.913)	(0.095)	(0.053)	(-)	(-)
Tourism	M	3.871	11.02	8.355	7.219	6.7	4.92	3.802	12.85	5.216	5.268	8.255	15.41	2.146	1.775
		(0.374)	(6.234)	(0.888)	(5.779)	(0.325)	(4.418)	(0.369)	(1.251)	(1.29)	(4.995)	(1.144)	(0.628)	(-)	(-)
	Q	4.249	13.57	9.128	10.251	8.0001	2.994	5.133	12.519	4.199	13.216	10.692	16.399	3.155	2.774
		(0.169)	(6.146)	(0.723)	(1.629)	(0.904)	(2.164)	(2.205)	(2.981)	(0.312)	(6.298)	(1.772)	(0.046)	(-)	(-)
	Y	8.263	5.066	4.829	2.574	4.502	2.304	8.066	5.637	8.496	6.157	7.813	7.545	9.785	12.016
		(0.161)	(2.042)	(1.432)	(0.983)	(1.125)	(0.945)	(0.215)	(0.149)	(1.345)	(1.535)	(0.398)	(0.675)	(-)	(-)

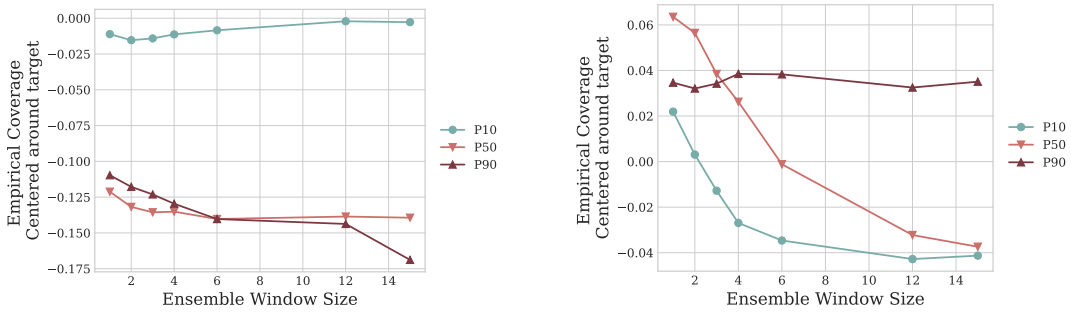
To complement the probabilistic forecast revision results, we evaluate median forecasts with the *Symmetric Quantile Percentage Change* (sQPC), that we define in Equation 49. sQPC measures the relative change in predicted quantiles across consecutive forecast creation dates, providing a quantitative view of temporal instability or forecast revision rates. Inspired by the sMAPE metric, sQPC uses a symmetric denominator, based on both current and previous forecasts, to mitigate issues of numerical instability (Hyndman and Koehler, 2006). This design ensures robustness when dealing with small predicted values and avoids the division-by-zero problems common in traditional percentage-based metrics.

$$\text{sQPC}_q\left(\hat{\mathbf{Y}}^{(q)}_{[b][t][h]}\right) = \frac{200}{B \times T \times H} \sum_{b,t,h} \frac{|\hat{Y}^{(q)}_{b,t+1,h} - \hat{Y}^{(q)}_{b,t,h+1}|}{|\hat{Y}^{(q)}_{b,t+1,h}| + |\hat{Y}^{(q)}_{b,t,h+1}|}. \tag{49}$$

Because uncertainty decreases across FCDs it is expected that forecasted quantiles above P50 decrease while below P50 increase, to avoid complications, we measure the revisions for the median (q=.50).

Table 3 presents sQPC values, averaged over five runs, for all dataset–frequency combinations and model architectures, including statistical baselines. For `LSTM` encoder models, the forking-sequences scheme reduces forecast revisions by %, on average across datasets, compared to window-sampling. Consistent gains are observed for `MLP` (28.8%), `RNN` (28.8%), and `CNN` (31.3%) encoders, while the `Transformer`-based encoder has relatively less improvement (8.8%). Table 2 is summarized in Fig. 14h which shows the percentage change of the sQPC metric for models with the *forking-sequences* training scheme over that of models with the *window-sampling* scheme, averaged across datasets. All encoder variants with forking-sequences have improved sQPC.

K Ensemble Effects on Quantile Coverage



(a) M1 Monthly **(b)** M4 Weekly

Figure 17: Difference between empirical and theoretical coverage for `NBEATS` model with and without forking-sequence moving-average ensembling during inference. (a) On the M1 monthly dataset, ensembling slightly improves the undercoverage of the lower quantile (P10), while negatively affecting P50 and P90. (b) On the M4 weekly dataset, ensembling modestly improves undercoverage of the upper quantile (P90), with limited or mixed effects on P10 and P50. Overall, the influence of the ensemble window on quantile coverage is not consistent across datasets or quantiles.

In this experiment, we apply the ensemble technique to a pre-trained `NBEATS` model using the benchmark M1 monthly and M4 weekly datasets. The `NBEATS` model is employed in a zero-shot capacity, with the only variation being the length of the ensemble considered for the forking sequences. In our controlled experiment, we vary the windows used in the forking-sequences ensemble and measure their impact on the coverage of the estimated quantiles:

$$\text{Coverage}\left(\mathbf{y}^{(q)}_{[b][t][h]},\ \hat{\mathbf{Y}}_{[b][t][h]}\right) = \frac{1}{B \times T \times H} \sum_{b,t,h} \left(\mathbb{1}_{\{y_{b,t,h} \leq \hat{y}^{(q)}_{b,t,h}\}} - q \right). \tag{50}$$

For the M1 dataset (Figure 17a), we observe a general under-coverage across quantiles. P50 and P90 start at roughly 10% under-coverage with a window size of 1 and gradually move toward about 14% under-coverage as the window increases. P10, in contrast, shows a mild improvement. While these trends suggest that the ensemble window can influence calibration, the direction and magnitude of the impact differ across quantiles. Results on the M4 dataset (Figure 17b) display a different pattern. The model begins with slight over-coverage—approximately 6% for P50, 2% for P10, and 4% for P90. Increasing the ensemble window reduces this over-coverage and eventually shifts P10 and P50 toward slight under-coverage, while the effect on P90 remains ambiguous.

The two experiments reveal no consistent calibration effect attributable to the ensemble window. The influence varies by dataset and quantile, and in several cases the shifts are small or in opposing directions. These mixed results indicate that the coverage impact of forking-sequence ensembling is inconclusive and point to the need for a deeper understanding of how ensembling interacts with quantile estimation.

Building on these observations, future research can explore more sophisticated methods for aggregating quantile forecasts, moving beyond simple averaging techniques. Specifically, inspiration can be drawn from the Quantile Regression Averaging (QRA) method (Liu et al., 2017). QRA uses quantile linear regression on the output of individual base models to generate more robust probabilistic forecasts, a method that has proven effective in various forecasting competitions. The varying impact of ensembling on different quantiles observed here suggests a potential benefit in employing the more flexible aggregation strategies described in the paper by Fakoor et al. (2023). Their approach considers weighted ensembles where weights can vary not only across individual models but also across quantile levels, which could help address the inconsistent coverage effects noted in our experiments.