# BAYES AVERAGING OR OVER-SMOOTHING? EXCESSIVE CONTRACTION IN TIME SERIES FORECASTING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Time series forecasting models often produce predictions that vary much less than observed futures, a behavior commonly interpreted as over-smoothing. Yet under squared loss, even the Bayes-optimal predictor averages over plausible futures, reducing variation under target uncertainty. Reduced variation alone therefore does not establish a modeling deficiency. We develop a Bayes-referenced framework for variation *across* predictions for similar inputs and *within* individual forecast trajectories. For both perspectives, we derive an observable criterion that relates predictive amplitude to target alignment and identifies excessive contraction along the model's current prediction direction, without directly estimating the Bayes-optimal predictor. Using the proposed contraction measures, we examine eight forecasting models across six benchmark datasets and confirm that predictive contraction is widespread and that stronger contraction is consistently associated with larger forecasting error. The Bayes-referenced analysis, however, reveals two reversals in this interpretation. First, observed contraction does not imply excessive contraction: sample-wise contrasts generally favor further attenuation, while temporal contrasts span both excessive- and insufficient-contraction regimes. Second, these aggregate tendencies themselves conceal heterogeneous window-level regimes. Within the same model–dataset pair, some windows favor attenuation while others favor amplification, and, in both regimes, larger departures from unit scaling tend to accompany higher error. A contrast-error decomposition further shows that optimal uniform rescaling captures only part of the pooled contrast error, motivating a finer analysis of alignment and window-level heterogeneity. Overall, our results establish a Bayes-referenced view of predictive contraction that distinguishes over-smoothing from uncertainty-induced attenuation and reveals heterogeneous scaling behavior across individual forecasts.

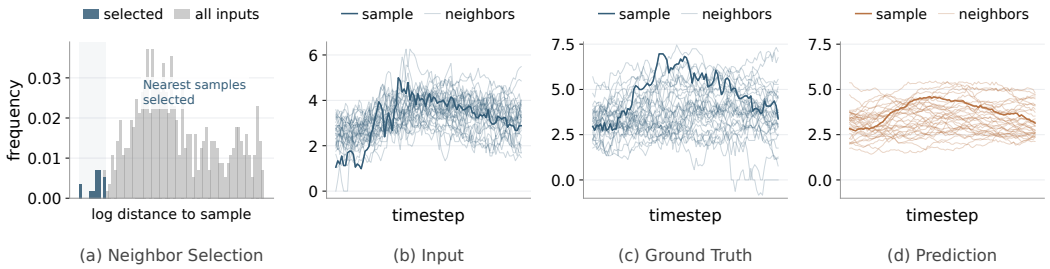
## 1 INTRODUCTION

Time series forecasting models often produce predictions that are smoother and more homogeneous than the observed futures. This behavior has been described as regression to the mean or over-smoothing, where forecasts suppress variation present in realized trajectories (Sun et al., 2025; Yu et al., 2026; Zhang et al., 2026). A common concern is that forecasts suppress meaningful future structure and fail to reflect the variation in the data. This has motivated methods seeking sharper or more structure-preserving predictions through alternative training objectives and model designs (Le Guen & Thome, 2019; Sun et al., 2025; Zhang et al., 2026).

Figure 1 illustrates this behavior for PatchTST (Nie et al., 2023) on ETTm1 (Zhou et al., 2021). Windows with similar observed histories can have substantially different futures, while their predictions remain more alike and exhibit weaker variation within individual trajectories. We refer to this reduction relative to realized targets as predictive contraction and distinguish two complementary perspectives: sample-wise contraction concerns variation across nearby inputs, whereas temporal contraction concerns variation within individual forecast trajectories. Using these measures, we examine eight forecasting models across six benchmark datasets and confirm two empirical patterns: predictive contraction is widespread, and stronger contraction is consistently associated with higher window-level forecasting error.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




(a) Neighbor Selection (b) Input (c) Ground Truth (d) Prediction

Figure 1: Predictive contraction in PatchTST forecasts on ETTm1. **(a)** Neighboring windows are selected by input similarity. **(b)** Their observed histories are highly similar, while **(c)** their realized futures differ substantially. **(d)** Predictions exhibit less variation both across samples and within individual trajectories.

These findings support the prevailing empirical intuition that stronger contraction accompanies poorer forecasts. Their interpretation, however, depends on the statistical target represented by a point forecast (Gneiting, 2011). Under squared loss, the Bayes-optimal point predictor is the conditional mean (Banerjee et al., 2005; Gneiting, 2011). It averages over uncertainty about the future, leaving some variation in realized outcomes absent even from an optimal point forecast (Gneiting & Katzfuss, 2014; Green et al., 2026). Greater conditional uncertainty can also simultaneously increase forecasting error and strengthen contraction relative to realized targets (Green et al., 2026). Contraction relative to realized futures therefore reflects both the attenuation induced by Bayes-optimal averaging and any additional contraction introduced by the forecasting model. Assessing over-smoothing requires distinguishing these two effects.

The Bayes-optimal predictor, however, is unobserved, making direct comparison impossible. We address this by studying sample-wise and temporal variation through contrast operators. Under expected squared loss, we show that optimal rescaling against the observed target contrast is equivalent to rescaling against its Bayes-optimal counterpart. This result yields an observable criterion for excessive contraction that depends jointly on predictive amplitude and target alignment. We apply the criterion both to model–dataset pairs as a whole and to individual forecast windows.

At the aggregate level, this Bayes-referenced assessment changes the interpretation of widespread contraction. Stronger contraction is consistently associated with higher forecasting error, yet sample-wise predictions generally favor further attenuation relative to their current amplitude. Temporal predictions, in contrast, span both excessive- and insufficient-contraction regimes. Observed contraction and excessive contraction therefore capture distinct properties of a forecast.

The window-level analysis shows substantial variation within each model–dataset pair. The aggregate scaling coefficient summarizes window-level requirements through a prediction-energy-weighted average, even when different windows favor opposite adjustments. Attenuation and amplification can therefore coexist within the same set of predictions. Moreover, larger departures from unit scaling tend to accompany higher forecasting error in both regimes. An aggregate contraction regime can thus conceal substantial differences among individual forecasts.

This heterogeneity also helps explain the limited effect of amplitude correction. We decompose aggregate contrast error into the part removable by a shared rescaling and the residual remaining after that adjustment. Across the evaluated settings, most of the error remains after optimal amplitude correction, and temporal calibration with a shared coefficient produces only modest forecasting gains even under test-oracle fitting. These results show that aggregate scaling captures only part of the variation in the adjustments preferred by individual forecasts.

Our contributions are threefold:

**A Bayes-referenced criterion for excessive contraction.** We define excessive contraction relative to the variation of the Bayes-optimal predictor rather than the realized target. For both sample-wise and temporal variation, we derive an observable criterion that compares predictive amplitude with target alignment without estimating the Bayes-optimal predictor.

**A different interpretation of widespread predictive contraction.** Across eight forecasting models and six benchmark datasets, predictions generally exhibit less variation than the observed futures, and stronger contraction is consistently associated with higher forecasting error. However, relative to the Bayes-optimal reference, sample-wise predictions are typically insufficiently contracted, while temporal predictions exhibit both excessive and insufficient contraction.

**Substantial heterogeneity behind aggregate contraction.** The appropriate amount of scaling varies considerably across individual forecast windows, with attenuation and amplification often required within the same model–dataset pair. Further analysis shows that amplitude correction alone accounts for only a limited part of prediction error, helping explain the modest gains from global rescaling.

## 2 Measurement of Predictive Contraction

We measure predictive contraction by comparing target and prediction contrasts, which isolate variation either across nearby samples or within forecast trajectories.

### 2.1 Forecasting Setup and Contrasts

Consider a univariate time series $\mathbf{u} = (u_t)_{t=1}^{T}$. Given a lookback length $L$ and forecast horizon $F$, stride-one windowing yields $N = T - L - F + 1$ pairs:

$$\mathbf{x}_i := (u_i, \dots, u_{i+L-1}), \qquad \mathbf{y}_i := (u_{i+L}, \dots, u_{i+L+F-1}), \qquad i = 1, \dots, N.$$

For a predictor $f_\theta : \mathbb{R}^L \to \mathbb{R}^F$, let $\widehat{\mathbf{y}}_i := f_\theta(\mathbf{x}_i)$.

Sample-wise and temporal contraction concern different modes of variation. We use a contrast operator $\mathcal{C}_k$ to isolate the variation relevant to each perspective, with $k = \mathrm{samp}$ for variation across nearby samples and $k = \mathrm{temp}$ for variation within a forecast trajectory. The corresponding target and prediction contrasts are

$$\mathbf{t}_i^k := \mathcal{C}_k(\mathbf{y}_i), \qquad \mathbf{p}_i^k := \mathcal{C}_k(\widehat{\mathbf{y}}_i). \tag{1}$$

The two contrast operators are defined below.

### 2.2 Sample-wise and Temporal Contrasts

**Sample-wise contrasts.** Sample-wise variation is defined locally among windows sharing similar observed histories. For each input window $\mathbf{x}_i$, we therefore construct a neighborhood $\mathcal{N}_i$ consisting of other windows that are close to $\mathbf{x}_i$ in input space. Using the root-mean-square distance,

$$d(\mathbf{x}_i, \mathbf{x}_j) := \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\sqrt{L}}, \qquad \mathcal{N}_i := \{j \neq i : \delta < d(\mathbf{x}_i, \mathbf{x}_j) \leq \varepsilon\}, \tag{2}$$

where $\varepsilon > 0$ controls the neighborhood radius and $0 \leq \delta < \varepsilon$ excludes identical or near-identical inputs. Threshold choices are specified in Section 3.1.

For $\mathcal{N}_i \neq \varnothing$, the sample-wise contrast operator is

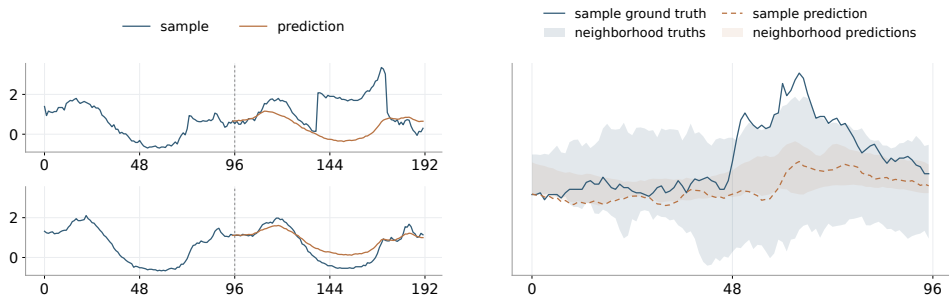
$$\mathcal{C}_{\mathrm{samp}}(\mathbf{z}_i) := \mathbf{z}_i - \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathbf{z}_j. \tag{3}$$

The resulting contrasts capture local deviations after neighborhood centering.

**Temporal contrasts.** To characterize variation within forecast trajectories (Figure 2(b)), define

$$\mathcal{C}_{\mathrm{temp}}(\mathbf{z}_i) := \mathbf{z}_i - \overline{z}_i \mathbf{1}_F, \qquad \overline{z}_i := \frac{1}{F} \sum_{h=1}^{F} z_{i,h}. \tag{4}$$

Here $\mathbf{1}_F$ is the all-ones vector; temporal centering removes the trajectory mean.




(a) Sample-wise contraction (b) Temporal contraction

Figure 2: Examples of predictive contraction. (a) Sample-wise: nearby inputs have substantially different targets but similar predictions. (b) Temporal: a prediction exhibits less within-window variation than its target.

### 2.3 Contraction Ratios

We measure contraction at two levels: $R_k$ summarizes aggregate contraction, while $R_{k,i}$ characterizes contraction for individual windows. For either perspective $k \in \{\mathrm{samp}, \mathrm{temp}\}$, we define

$$R_k := \left( \frac{\sum_i \|\mathbf{p}_i^k\|_2^2}{\sum_i \|\mathbf{t}_i^k\|_2^2} \right)^{1/2}, \qquad R_{k,i} := \frac{\|\mathbf{p}_i^k\|_2}{\|\mathbf{t}_i^k\|_2}. \tag{5}$$

For multivariate data, contrasts and neighborhoods are computed per channel; $\mathbf{p}_i^k$ and $\mathbf{t}_i^k$ concatenate valid channel contrasts. Empty sample-wise neighborhoods are excluded. Ratios are defined only when their denominators are positive; energies are pooled before division.

For both measures, $R < 1$ indicates contraction, $R = 1$ equal contrast amplitude, and $R > 1$ expansion relative to the observed targets. Smaller values therefore correspond to stronger contraction. Whether such contraction is excessive is examined in Section 4.

## 3 Empirical Study of Predictive Contraction

In this section, we turn to an empirical study of predictive contraction. Our goal is to assess whether predictive contraction (i) arises at scale in real-world datasets, (ii) persists across different datasets and models, and (iii) has impact on forecasting accuracy. We begin by analyzing predictive contraction on a single dataset–model pair, then extend the analysis to six benchmark datasets and eight representative models, and finally examine the relationship between predictive contraction and prediction error.

### 3.1 Contraction Quantified — Experimental Setting

**Datasets and models.** We evaluate eight forecasting models: DUET (Qiu et al., 2025), PDF (Dai et al., 2024), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024), DLinear (Zeng et al., 2023), FITS (Xu et al., 2024), Crossformer (Zhang & Yan, 2023), and Informer (Zhou et al., 2021), on six benchmarks: ETTm1, ETTm2, ETTh1, ETTh2, Traffic, and Electricity. Both the lookback length and forecast horizon are fixed to 96. We use prediction artifacts obtained under the Time Series Forecasting Benchmark (TFB) protocol (Qiu et al., 2024). All metrics are evaluated on the standardized scale, using channel-wise normalization parameters estimated from the training split.

**Instantiating the contraction metrics.** We compute the sample-wise and temporal contraction ratios defined in Section 2. For the sample-wise contraction analysis, test windows are sampled at a stride of 20, with the same sampled windows used across models on each dataset. This subsampling reduces the computational cost of pairwise distance computation, which scales quadratically with the number of windows.

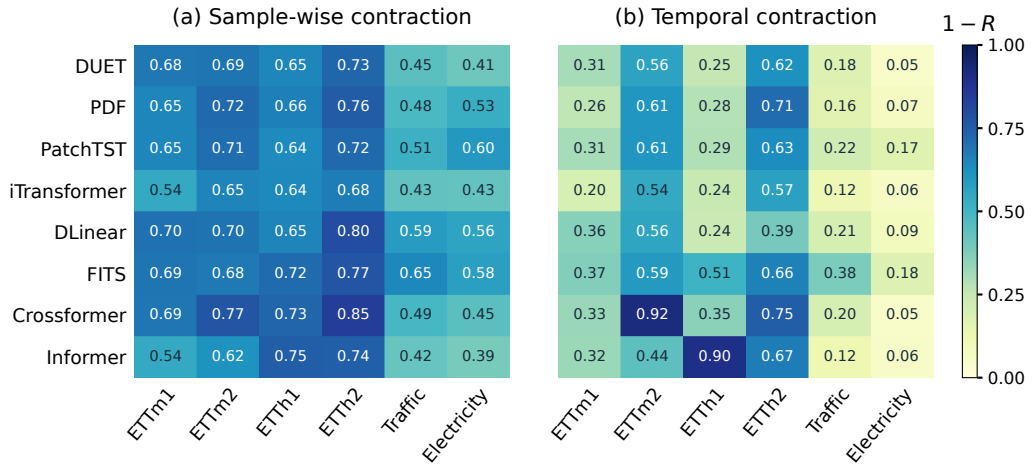



Figure 3: Contraction magnitude across datasets and models: (a) $1 - R_{\text{samp}}$ and (b) $1 - R_{\text{temp}}$. Higher values indicate stronger contraction; both panels share the same color scale.

Input neighborhoods are constructed using the distance metric defined in Section 2.2. We set the minimum distance to $\delta = 10^{-3}$ and determine the neighborhood radius $\varepsilon_c$ as the empirical $0.05$ quantile of pairwise input distances exceeding $\delta$ in channel $c$. The finding of sample-wise contraction remains consistent across the tested neighborhood quantiles, although the magnitudes of $R$ and $\beta$ vary with the neighborhood scale. Appendix A.1 provides detailed results.

### 3.2 Contraction Is Widespread

We first quantify the contraction illustrated in Figure 1. For PatchTST on ETTm1, the sample-wise contraction ratio is $R_{\text{samp}} = 0.349$, indicating substantially reduced variation across locally similar predictions relative to their corresponding targets. The temporal contraction ratio is $R_{\text{temp}} = 0.692$, showing that reduced variation also occurs within individual forecast trajectories.

We then extend the analysis to six benchmark datasets and eight forecasting models. Figure 3 summarizes the contraction magnitudes across all $6 \times 8$ dataset–model pairs.

Contraction is widespread across architectures and datasets, although its magnitude varies substantially. The two perspectives also exhibit distinct patterns: a model that preserves temporal variation relatively well can still show pronounced sample-wise contraction, and no single architecture consistently avoids contraction from both perspectives.

These results establish predictive contraction as a widespread empirical phenomenon. We next examine its relation to forecasting error within each dataset–model pair.

### 3.3 Contraction Matters — Relation to Error

Overly smooth or weakly varying forecasts are commonly viewed as an undesirable forecasting behavior, motivating methods that seek to preserve richer predictive dynamics. We therefore examine the association between the observed contraction and forecasting error.

Across all 48 dataset–model pairs, the correlation coefficients between contraction magnitude and forecasting error are positive under both the sample-wise and temporal perspectives. Thus, windows with stronger contraction tend to incur larger forecasting errors, although the strength of this association varies across datasets and models. Detailed results are provided in Appendix A.2.

The observed pattern is consistent with the prevailing view that overly contracted forecasts reflect undesirable predictive behavior. Yet the same association can also arise from Bayes-optimal averaging under heterogeneous conditional uncertainty, which can simultaneously increase forecasting er-

ror and reduce predictive variation. This motivates distinguishing contraction due to Bayes-optimal averaging from excessive contraction that can be corrected.

# 4 IDENTIFYING EXCESSIVE CONTRACTION

The preceding analysis establishes that predictive contraction is widespread and that stronger contraction is consistently associated with larger forecasting error. We revisit these findings through a Bayes-referenced criterion that relates predictive amplitude to target alignment. The analysis proceeds from aggregate contraction regimes to window-level scaling heterogeneity. We then use the same amplitude–alignment framework to quantify the contrast error removable by uniform rescaling and interpret model differences and the effects of amplitude correction.

## 4.1 A BAYES-REFERENCED CRITERION FOR EXCESSIVE CONTRACTION

The contraction ratio $R_k$ introduced in Equation 5 measures retained contrast amplitude relative to the realized target. Reduced amplitude alone does not indicate excessive contraction; its interpretation also depends on alignment with predictable target structure.

Let $\mathbf{T}$ and $\mathbf{P}$ denote target and prediction contrasts for either perspective, with the trained predictor fixed. Let $\mathcal{G}$ be the $\sigma$-field generated by all inputs used to construct $\mathbf{P}$, including those used for sample-wise neighborhood selection. Then $\mathbf{P}$ is $\mathcal{G}$-measurable, and

$$\mathbf{T}^\star = \mathbb{E}[\mathbf{T} \mid \mathcal{G}] \tag{6}$$

defines the contrast-level Bayes reference under squared loss.

For square-integrable contrasts with $\mathbb{E}\|\mathbf{P}\|_2^2 > 0$, the optimal scalar rescaling is

$$\beta^\star := \arg\min_{b\in\mathbb{R}} \mathbb{E}\left[\|b\mathbf{P} - \mathbf{T}^\star\|_2^2\right] = \frac{\mathbb{E}\langle \mathbf{P}, \mathbf{T}^\star\rangle}{\mathbb{E}\|\mathbf{P}\|_2^2}. \tag{7}$$

When $\beta^\star > 1$, amplification toward this optimum reduces the expected squared distance to $\mathbf{T}^\star$, indicating excessive contraction along the current prediction direction. Values $0 \le \beta^\star < 1$ favor attenuation, while $\beta^\star < 0$ requires a sign reversal.

Although $\mathbf{T}^\star$ is unobserved,

$$\mathbb{E}\langle \mathbf{P}, \mathbf{T}^\star\rangle = \mathbb{E}\langle \mathbf{P}, \mathbf{T}\rangle, \tag{8}$$

because $\mathbf{P}$ is $\mathcal{G}$-measurable and $\mathbb{E}[\mathbf{T} - \mathbf{T}^\star \mid \mathcal{G}] = 0$. Therefore,

$$\beta = \frac{\mathbb{E}\langle \mathbf{P}, \mathbf{T}\rangle}{\mathbb{E}\|\mathbf{P}\|_2^2} = \beta^\star, \tag{9}$$

identifying the population-optimal scaling through moments of observable contrasts.

We connect this population criterion to the empirical contraction ratio $R_k$. For each perspective $k \in \{\mathrm{samp}, \mathrm{temp}\}$, define the empirical target alignment as

$$A_k = \frac{\sum_i \langle \mathbf{p}_i^k, \mathbf{t}_i^k\rangle}{\sqrt{\left(\sum_i \|\mathbf{p}_i^k\|_2^2\right)\left(\sum_i \|\mathbf{t}_i^k\|_2^2\right)}}. \tag{10}$$

The sums use the same valid contrasts as $R_k$, with both pooled energies assumed positive. The corresponding empirical estimate of $\beta$ is

$$\beta_k = \frac{\sum_i \langle \mathbf{p}_i^k, \mathbf{t}_i^k\rangle}{\sum_i \|\mathbf{p}_i^k\|_2^2} = \frac{A_k}{R_k}. \tag{11}$$

This decomposition separates retained contrast amplitude, measured by $R_k$, from target alignment, measured by $A_k$. Thus, $R_k < 1$ indicates observed contraction, whereas $\beta_k > 1$, equivalently $A_k > R_k$, provides the aggregate empirical criterion for excessive contraction along the model's current prediction direction.

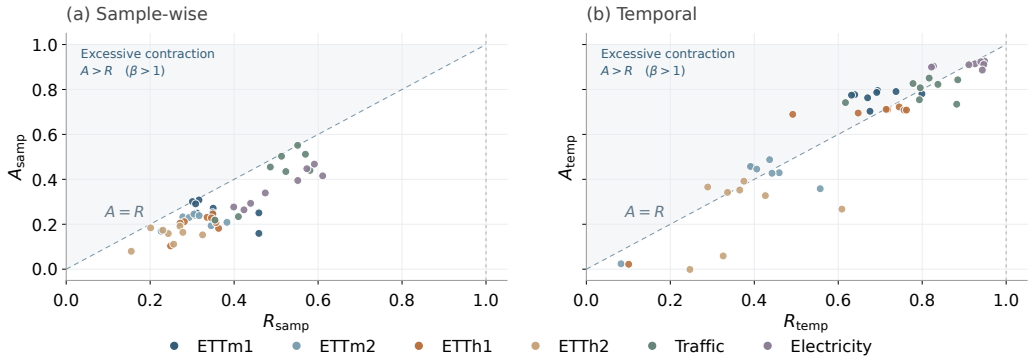



Figure 4: Alignment $A$ versus contraction ratio $R$: (a) sample-wise and (b) temporal contrasts. Each point represents a dataset–model pair. The dashed diagonal marks $A = R$; the shaded region corresponds to $A > R$, or equivalently $\beta > 1$.

### 4.2 Reassessing Widespread Contraction

Section 3 establishes that predictive contraction is widespread across datasets and models. We now revisit this finding using the Bayes-referenced criterion. Figure 4 first confirms the earlier observation: all dataset–model pairs lie to the left of $R = 1$, showing widespread contraction relative to the realized targets. Incorporating target alignment changes the interpretation. For sample-wise contrasts, nearly all settings lie below the $A = R$ boundary, favoring further attenuation at the aggregate level. Temporal contrasts are more heterogeneous, with settings appearing on both sides of the boundary.

Taken together, these results refine the first empirical finding from Section 3.2. Predictive contraction is widespread relative to realized futures, while excessive contraction is not universal. The distinction depends jointly on retained contrast amplitude and its alignment with the target structure.

### 4.3 Reassessing the Contraction–Error Association

The second empirical finding from Section 3 is that stronger observed contraction is consistently associated with larger forecasting error. We revisit this association at the window level, where opposing scaling preferences can coexist within the same model–dataset pair. For each perspective $k \in \{\mathrm{samp}, \mathrm{temp}\}$, define

$$\beta_{k,i} = \frac{\langle \mathbf{p}_i^k, \mathbf{t}_i^k \rangle}{\|\mathbf{p}_i^k\|_2^2}. \tag{12}$$

For nonzero prediction energy, this label-dependent coefficient retrospectively minimizes the contrast error for window $i$. Values $0 \le \beta_{k,i} < 1$ favor attenuation, $\beta_{k,i} > 1$ amplification, and $\beta_{k,i} < 0$ a sign reversal. The pooled coefficient is the prediction-energy-weighted aggregate

$$\beta_k = \frac{\sum_i \|\mathbf{p}_i^k\|_2^2 \beta_{k,i}}{\sum_i \|\mathbf{p}_i^k\|_2^2}. \tag{13}$$

An aggregate tendency toward attenuation can therefore coexist with individual windows that favor amplification, and conversely.

Figure 5 reveals a consistent two-sided pattern. Within the $\beta_{k,i} < 1$ regime, forecasting error tends to decrease as $\beta_{k,i}$ increases, whereas within the $\beta_{k,i} > 1$ regime, it tends to increase with $\beta_{k,i}$. Thus, larger deviations from $\beta_{k,i} = 1$ in either direction are generally associated with higher forecasting error.

This result refines the positive contraction–error association observed in Section 3. High-error forecasts are not uniformly characterized by insufficient predictive variation. In particular, strongly contracted sample-wise predictions can still lie in the $\beta_{k,i} < 1$ regime, where further attenuation improves their fit to the observed target contrasts. The association between contraction and fore-

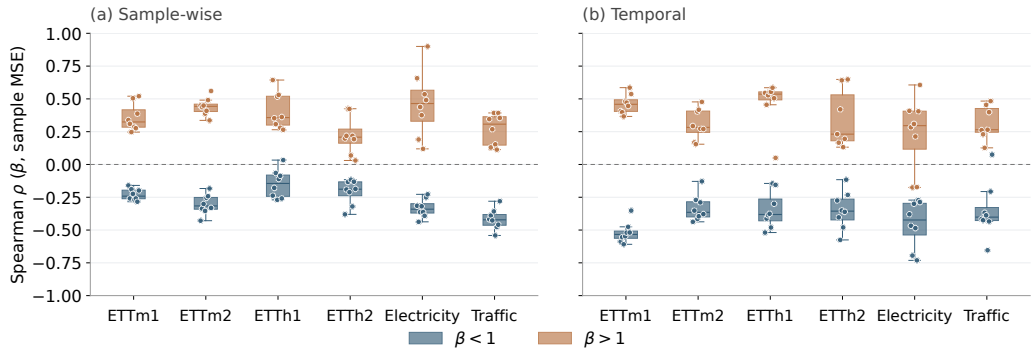



Figure 5: Spearman correlations between $\beta_{k,i}$ and window-level MSE across six datasets and eight forecasting models, computed separately for windows with $\beta_{k,i} < 1$ and $\beta_{k,i} > 1$. Panels (a) and (b) show sample-wise and temporal contrasts, respectively. Each point denotes one model, and boxes summarize the distribution across models. The dashed line marks zero correlation.

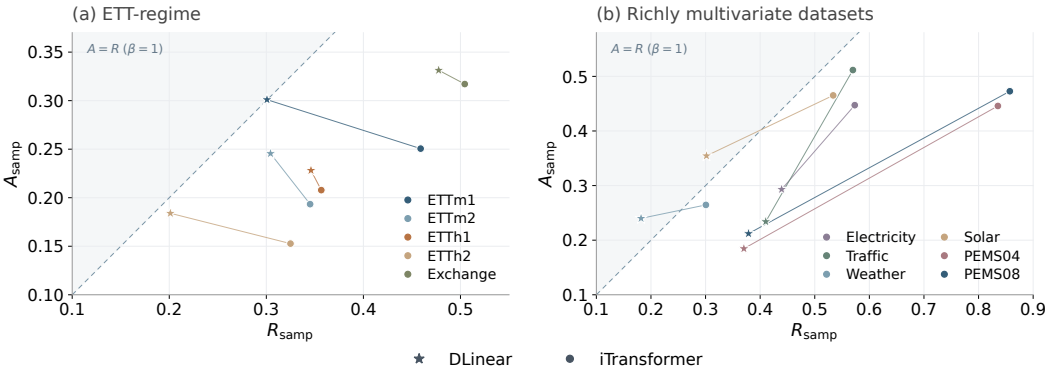



Figure 6: Sample-wise amplitude–alignment comparison between DLinear and iTransformer across datasets. Panel (a) shows the ETT datasets and Exchange, where iTransformer's higher $R_{\text{samp}}$ is accompanied by lower $A_{\text{samp}}$; panel (b) shows Electricity, Traffic, Weather, Solar, PEMS04, and PEMS08, where higher $R_{\text{samp}}$ is accompanied by higher $A_{\text{samp}}$. The dashed line marks $\beta = 1$.

casting error therefore does not reduce to a monotonic relation between predictive variation and accuracy.

### 4.4 Contrast Error under Optimal Rescaling

The coefficient $\beta_k$ identifies the optimal change in predictive amplitude. We next quantify how much contrast error can be removed by this adjustment. For either perspective, omitting the index $k$, the normalized contrast error satisfies

$$C := \frac{\sum_i \|\mathbf{t}_i - \mathbf{p}_i\|_2^2}{\sum_i \|\mathbf{t}_i\|_2^2} = \underbrace{(R - A)^2}_{\text{amplitude mismatch}} + \underbrace{(1 - A^2)}_{\text{residual after rescaling}} . \tag{14}$$

Since $\beta = A/R$, the first term equals $R^2(1 - \beta)^2$ and is exactly the error removed by optimal uniform rescaling. The second term is the error that remains.

The preceding results show that retaining more sample-wise variation does not provide a uniform advantage. We therefore compare DLinear and iTransformer to examine how differences in retained sample-wise variation translate into contrast quality across datasets.

For this model pair, the residual term accounts for more than 80% of the sample-wise contrast error across datasets (Appendix A.3), indicating that overall amplitude explains only a small part of the difference in contrast quality. Figure 6 shows that iTransformer consistently has a higher sample-wise $R$, while the accompanying change in $A$ is strongly dataset-dependent.

On the ETT datasets and Exchange, higher $R$ is accompanied by lower $A$ and higher contrast error than DLinear. Exchange provides an intuitive example, as exchange rates are known for their near-random-walk behavior and limited predictability from historical observations (Engel & Wu, 2023). The additional variation produced by iTransformer therefore does not translate into stronger target alignment.

On Electricity, Traffic, Weather, Solar, PEMS04, and PEMS08, higher $R$ is instead accompanied by higher $A$ and lower contrast error. The increase in alignment is particularly large on the PEMS datasets, where it outweighs the larger amplitude-mismatch term. The effect of additional predictive variation therefore depends on the accompanying change in target alignment.

### 4.5 Discussion: the Heterogeneous Nature of $\beta$

An intuitive use of $\beta$ is to adjust the temporal variation expressed by the model and examine its effect on forecasting accuracy. We therefore apply $\beta_{\mathrm{temp}}$ as a temporal rescaling coefficient and measure the resulting change in forecasting error. Even with test-oracle fitting, the resulting MSE gains are small in most settings (Appendix A.4).

This limited gain is consistent with the window-level heterogeneity observed in Section 4.3. Within the same model–dataset pair, windows favoring attenuation and amplification can coexist, and their preferred scaling factors vary substantially across windows. A single coefficient cannot adapt to these distinct window-level requirements. Instance-dependent calibration could capture this heterogeneity more directly, but lies beyond the scope of this work.

## 5 Conclusion

This work revisits predictive contraction in time-series forecasting through a Bayes-referenced distinction between observed contraction and excessive contraction. We characterize variation across nearby samples and within individual forecast trajectories, and find that both forms of contraction are widespread across models and datasets. Stronger observed contraction is also consistently associated with larger window-level forecasting error. This association, however, does not imply that contracted forecasts should simply be made more variable, since Bayes-optimal point prediction itself suppresses unpredictable variation.

We derive an observable criterion that combines predictive amplitude with target alignment and identifies the optimal rescaling of the variation already captured by a model. The resulting analysis reveals a clear distinction between observed and excessive contraction. Sample-wise predictions are strongly contracted relative to realized futures, yet generally favor further attenuation, while temporal predictions span both excessive- and insufficient-contraction regimes. At the instance level, larger departures from the Bayes-referenced scaling level in either direction tend to accompany higher forecasting error. Predictive contraction alone therefore provides an incomplete account of forecasting quality.

Overall, our results show that contraction relative to realized targets is insufficient evidence of over-smoothing. A Bayes-referenced assessment instead distinguishes excessive contraction from the contraction induced by uncertainty, providing a more principled interpretation of smooth or homogeneous forecasts.

### AI use statement

In this work, we used generative AI tools to aid and polish writing, retrieve and discover related literature, support research ideation and execution, draft sections of the manuscript, and assist in proving mathematical claims. AI assistance was also used for LaTeX formatting and preparation of the manuscript. The authors manually reviewed and verified the AI-assisted work. The authors take

responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### Reproducibility statement

Section 2 defines the forecasting setup, the sample-wise and temporal contrasts, and the contraction ratios. Section 3.1 specifies the benchmark datasets and models, the 96-step lookback and forecast horizon, normalization, test-window subsampling, and neighborhood construction. Section 4.1 gives the Bayes-referenced criterion and its mathematical derivation. Appendix A.1 reports neighborhood sensitivity analyses; Appendices A.2, and A.3 provide the detailed results supporting the main analyses.

## References


Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005. doi: 10.1109/TIT.2005.850145.

Tao Dai, Beiliang Wu, Peiyuan Liu, Naiqi Li, Jigang Bao, Yong Jiang, and Shu-Tao Xia. Periodicity decoupling framework for long-term series forecasting. In *International Conference on Learning Representations*, 2024.

Charles Engel and Steve Pak Yeung Wu. Forecasting the u.s. dollar in the 21st century. *Journal of International Economics*, 141:103715, 2023. doi: 10.1016/j.jinteco.2023.103715.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014. doi: 10.1146/annurev-statistics-062713-085831.

Riku Green, Zahraa S. Abdallah, and Telmo M. Silva Filho. Expectations vs. realities: The cost of mse-optimal forecasting under conditional uncertainty. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1298–1309, 2026. doi: 10.1145/3770855.3818087.

Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. In *Advances in Neural Information Processing Systems*, pp. 4191–4203. 2019.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

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

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: towards comprehensive and fair benchmarking of time series forecasting methods. *Proceedings of the VLDB Endowment*, 17 (9):2363–2377, 2024.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2025*, pp. 1185–1196, 2025.

Yanru Sun, Zongxia Xie, Dongyue Chen, Emadeldeen Eldele, and Qinghua Hu. Hierarchical classification auxiliary network for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20743–20751, 2025. doi: 10.1609/aaai.v39i19.34286.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations*, 2024.

Annan Yu, Danielle C. Maddix, Boran Han, Xiyuan Zhang, Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos, Andrew Gordon Wilson, Michael W. Mahoney, and Yuyang Wang. Understanding the implicit biases of design choices for time series foundation models. In *International Conference on Learning Representations*, 2026.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023.

Xingyu Zhang, Jingyao Wang, Xin Yu, Zeen Song, Jianqi Zhang, Changwen Zheng, and Wenwen Qiang. Dirichlet-guided group forecasting for alleviating over-smoothing in time series forecasting, 2026. URL https://arxiv.org/abs/2606.10592.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

# A Appendix

## A.1 Sensitivity to Neighborhood Definition and Sampling

We examine the sensitivity of the sample-wise analysis to three choices: the neighborhood quantile, the sampling stride, and the input distance used to define neighboring windows.

**Neighborhood quantile.** We vary the neighborhood quantile over $q \in \{0.01, 0.025, 0.05, 0.10, 0.20\}$, with $q = 0.05$ as the default. For each channel, the neighborhood radius $\varepsilon_c(q)$ is recomputed from RMS input distances after excluding near-duplicate pairs with distance at most $\delta = 10^{-3}$. Predictions, training-set standardization, and test-window sampling remain fixed. For each setting, $R$ and the dataset-level coefficient $\beta$ are computed by pooling contrast energies and inner products across valid windows and channels before taking ratios.

Across all six datasets and eight models, $R < 1$ holds throughout the tested quantile range. The same conclusion holds when evaluation is restricted to window–channel pairs that are valid under all five quantiles. Figure 7 shows the results on ETTm1. Although the magnitudes of $R$ and $\beta$ vary with neighborhood scale, the contraction finding is unchanged.

**Alternative input distances.** We compare RMS distance with cosine distance and DTW on ETTm1. Cosine distance is computed from the training-standardized input windows without additional window-wise centering. DTW uses squared pointwise differences as its local matching cost. All three distances use the same RMS-based near-duplicate exclusion rule, after which neighbors are selected according to the corresponding $q$-quantile. Predictions and sampled windows are unchanged.

Figure 8 shows that $R < 1$ persists across all three distances and tested quantiles. The magnitude of $\beta$, however, is more sensitive to the neighborhood definition. In particular, DTW produces systematically larger coefficients than RMS around the default quantile.

This difference is consistent with the interaction between DTW neighbor selection and the pointwise contrast construction used in our analysis. DTW can select input trajectories that are similar under temporal displacement, whereas the corresponding future targets and predictions are still averaged at their original forecast indices. This discrepancy is substantially reduced under the stride-96 setting in Figure 9, where input windows and target windows no longer overlap and DTW produces $R$ and $\beta$ values much closer to those obtained with RMS. We therefore use the RMS-based Euclidean construction for the Bayes-referenced analysis in the main text and treat DTW primarily as a sensitivity check for the contraction finding.

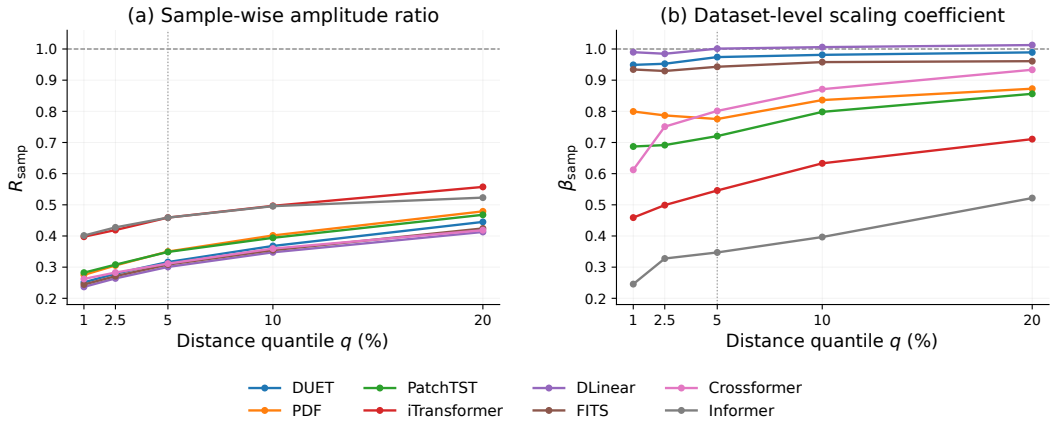



Figure 7: Sensitivity to the neighborhood quantile on ETTm1 under the RMS-distance definition. Left: sample-wise contraction ratio $R$; right: dataset-level coefficient $\beta$. Horizontal dashed lines indicate unity, and the vertical dotted lines mark the default $q = 0.05$.

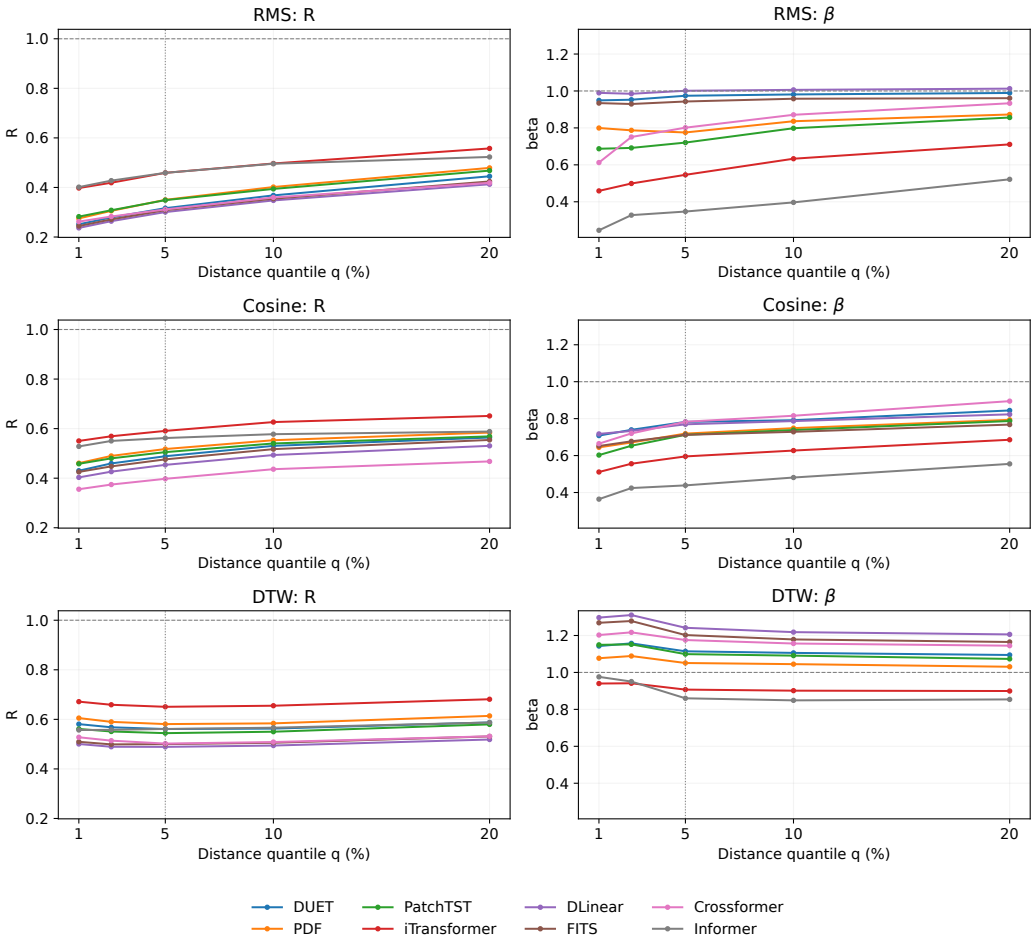



Figure 8: Sensitivity to the input distance on ETTm1. Rows correspond to RMS, cosine, and DTW distances; columns show the sample-wise contraction ratio $R$ and dataset-level coefficient $\beta$. All settings use the same RMS-based near-duplicate exclusion rule. Horizontal dashed lines indicate unity, and the vertical dotted lines mark $q = 0.05$.

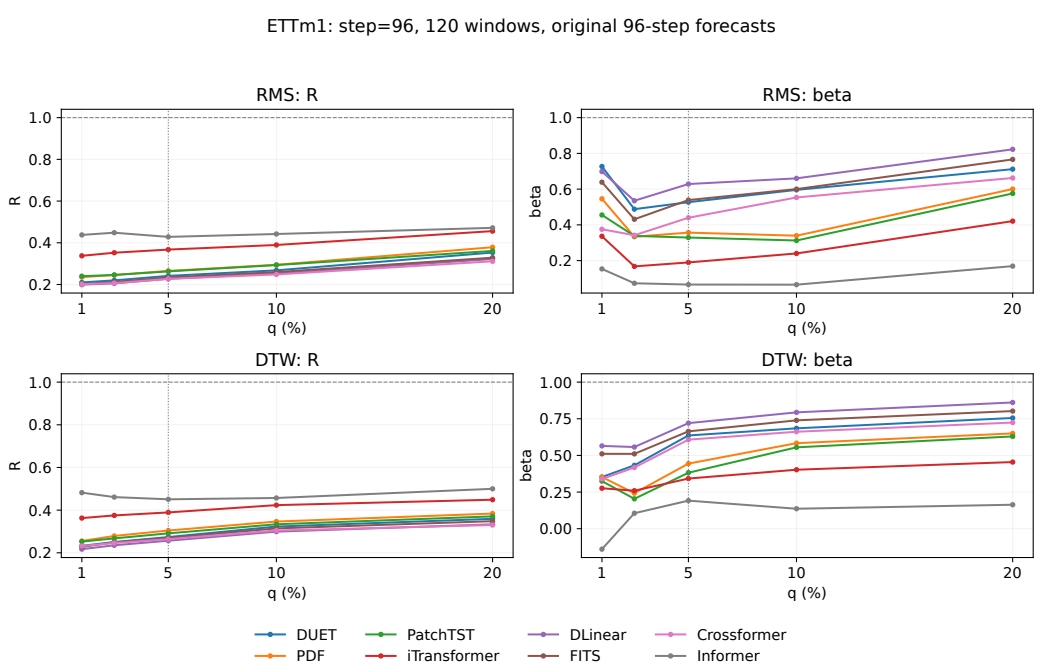



Figure 9: ETTm1 results with sampling stride 96. Rows correspond to RMS and DTW input distances; columns show the sample-wise contraction ratio $R$ and dataset-level coefficient $\beta$. Horizontal dashed lines indicate unity, and the vertical dotted lines mark $q = 0.05$.

**Sampling stride.** We additionally repeat the ETTm1 analysis with a sampling stride of 96, equal to both the input length and forecast horizon. This produces 120 test windows for which input windows do not overlap with one another, nor do target windows. Neighborhoods are recomputed from the resulting sampled set, while the predictions, standardization, and neighborhood quantiles remain unchanged.

As shown in Figure 9, $R < 1$ persists for every model and tested quantile under both RMS and DTW distances. The corresponding scaling coefficients are lower overall than under the default stride of 20 and remain below unity in this experiment. Because increasing the stride also changes the sampled windows and neighborhood density, we treat this experiment as a coarser-sampling sensitivity check rather than an isolated test of window overlap.

**Input and forecast length.** We finally vary the input length and forecast horizon jointly on ETTm1 using DLinear, with $P = F \in \{96, 192, 336, 720\}$. All other evaluation settings remain unchanged.

Figure 10 shows that contraction persists across all four settings from both perspectives, with $R < 1$ throughout. The Bayes-referenced coefficient varies more substantially with sequence length: sample-wise $\beta$ remains close to or below unity, while temporal $\beta$ decreases from above to below unity as the length increases. These results further separate the persistence of predictive contraction from its classification relative to the Bayes reference.

### A.2 Contraction and Forecasting Error

We assess the association between contraction and forecasting error separately within each of the 48 dataset–model pairs. Tables 1 and 2 report Spearman correlations between window-level $1 - R$ and MSE for the sample-wise and temporal perspectives, respectively. MSE is computed over all forecast steps and channels in the same window, using training-set standardization. We evaluate the correlation on ETTm1, ETTm2, ETTh1, ETTh2, Traffic, and Electricity.

For each window, channel energies are pooled before computing $R$. Sample-wise contrasts use the saved input-defined neighborhoods, while temporal contrasts subtract each channel's temporal

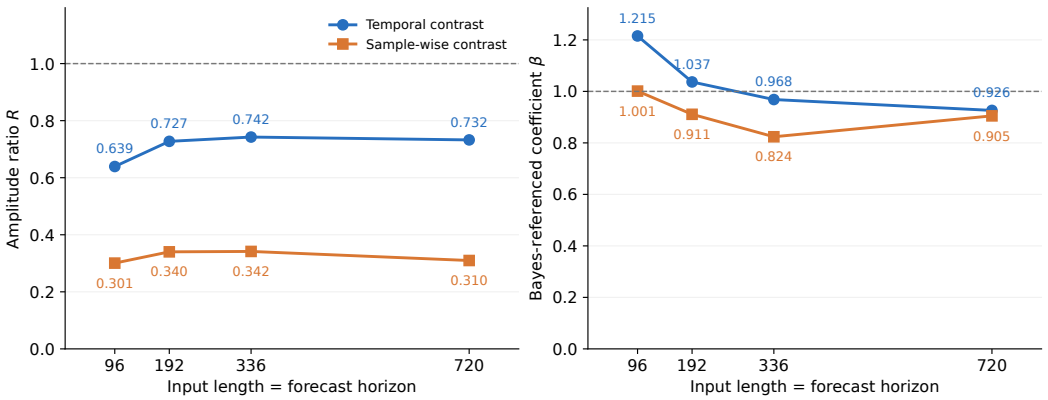



Figure 10: Sensitivity to input and forecast length on ETTm1 with DLinear. Input length and forecast horizon are varied jointly as $P = F \in \{96, 192, 336, 720\}$. Left: contraction ratio $R$; right: Bayes-referenced coefficient $\beta$. Results are shown for temporal and sample-wise contrasts. The dashed horizontal line indicates unity.

Table 1: Sample-wise Spearman correlations between window-level contraction magnitude $1 - R$ and MSE. Each entry uses windows within one dataset–model pair. Values are rounded to three decimal places.

| Model | ETTm1 | ETTm2 | ETTh1 | ETTh2 | Traffic | Electricity |
|---|---|---|---|---|---|---|
| DUET | 0.550 | 0.469 | 0.396 | 0.302 | 0.743 | 0.318 |
| PDF | 0.526 | 0.406 | 0.323 | 0.131 | 0.675 | 0.651 |
| PatchTST | 0.523 | 0.368 | 0.347 | 0.243 | 0.639 | 0.618 |
| iTransformer | 0.384 | 0.383 | 0.401 | 0.337 | 0.682 | 0.373 |
| DLinear | 0.498 | 0.401 | 0.378 | 0.071 | 0.414 | 0.627 |
| FITS | 0.478 | 0.440 | 0.334 | 0.243 | 0.413 | 0.592 |
| Crossformer | 0.529 | 0.317 | 0.367 | 0.079 | 0.738 | 0.540 |
| Informer | 0.314 | 0.372 | 0.046 | 0.151 | 0.348 | 0.096 |

mean. Channels without a valid neighborhood are excluded from the sample-wise energy aggregation; MSE still uses all channels. The correlations use $1 - R$ without clipping or restricting the analysis to $R < 1$.

All 48 estimated coefficients are positive under both perspectives, although their magnitudes vary. The direction of the association is therefore consistent across settings, while its strength is highly dataset- and model-dependent.

Table 2: Temporal Spearman correlations between window-level contraction magnitude $1 - R$ and MSE. Each entry uses windows within one dataset–model pair. Values are rounded to three decimal places.

| Model | ETTm1 | ETTm2 | ETTh1 | ETTh2 | Traffic | Electricity |
|---|---|---|---|---|---|---|
| DUET | 0.156 | 0.510 | 0.246 | 0.653 | 0.491 | 0.265 |
| PDF | 0.143 | 0.390 | 0.281 | 0.536 | 0.212 | 0.212 |
| PatchTST | 0.137 | 0.386 | 0.280 | 0.630 | 0.295 | 0.148 |
| iTransformer | 0.093 | 0.428 | 0.244 | 0.619 | 0.149 | 0.321 |
| DLinear | 0.297 | 0.500 | 0.258 | 0.308 | 0.037 | 0.140 |
| FITS | 0.277 | 0.368 | 0.510 | 0.542 | 0.315 | 0.105 |
| Crossformer | 0.265 | 0.629 | 0.373 | 0.461 | 0.562 | 0.288 |
| Informer | 0.215 | 0.251 | 0.485 | 0.205 | 0.316 | 0.236 |

Table 3: Sample-wise contrast-error decomposition for DLinear and iTransformer. The normalized contrast error satisfies $C = (R-A)^2+(1-A^2)$. Percentages in parentheses indicate the contribution of each term to $C$. Lower $C$ indicates a closer match between predicted and target contrasts; the lower value for each dataset is bold. DLinear uses shared weights on Weather, Solar, PEMS04, and PEMS08, and individual weights on the remaining datasets.

| Dataset | Model | $(R-A)^2$ | $1-A^2$ | $C$ |
|---|---|---|---|---|
| ETTm1 | DLinear | 0.0000 (0.0%) | 0.9094 (100.0%) | **0.9094** |
| | iTransformer | 0.0434 (4.4%) | 0.9372 (95.6%) | 0.9806 |
| ETTm2 | DLinear | 0.0035 (0.4%) | 0.9397 (99.6%) | **0.9432** |
| | iTransformer | 0.0230 (2.3%) | 0.9626 (97.7%) | 0.9856 |
| ETTh1 | DLinear | 0.0139 (1.4%) | 0.9479 (98.6%) | **0.9618** |
| | iTransformer | 0.0221 (2.3%) | 0.9568 (97.7%) | 0.9789 |
| ETTh2 | DLinear | 0.0003 (0.0%) | 0.9662 (100.0%) | **0.9665** |
| | iTransformer | 0.0296 (2.9%) | 0.9767 (97.1%) | 1.0063 |
| Exchange | DLinear | 0.0214 (2.3%) | 0.8902 (97.7%) | **0.9116** |
| | iTransformer | 0.0351 (3.8%) | 0.8994 (96.2%) | 0.9345 |
| Electricity | DLinear | 0.0214 (2.3%) | 0.9141 (97.7%) | 0.9354 |
| | iTransformer | 0.0159 (1.9%) | 0.7999 (98.1%) | **0.8158** |
| Traffic | DLinear | 0.0311 (3.2%) | 0.9453 (96.8%) | 0.9763 |
| | iTransformer | 0.0034 (0.5%) | 0.7383 (99.5%) | **0.7417** |
| Weather | DLinear | 0.0034 (0.4%) | 0.9424 (99.6%) | 0.9458 |
| | iTransformer | 0.0013 (0.1%) | 0.9300 (99.9%) | **0.9313** |
| Solar | DLinear | 0.0028 (0.3%) | 0.8743 (99.7%) | 0.8771 |
| | iTransformer | 0.0047 (0.6%) | 0.7836 (99.4%) | **0.7883** |
| PEMS04 | DLinear | 0.0345 (3.4%) | 0.9659 (96.6%) | 1.0004 |
| | iTransformer | 0.1518 (15.9%) | 0.8012 (84.1%) | **0.9530** |
| PEMS08 | DLinear | 0.0278 (2.8%) | 0.9550 (97.2%) | 0.9828 |
| | iTransformer | 0.1478 (16.0%) | 0.7765 (84.0%) | **0.9243** |

### A.3 Contrast-Error Decomposition for the Model-Pair Case Study

Table 3 reports the full sample-wise contrast-error decomposition for DLinear and iTransformer used in Section 4.4. Across the 22 model–dataset pairs in this table, the residual term $1 - A^2$ accounts for at least $84.0\%$ of the normalized contrast error. For iTransformer on PEMS04 and PEMS08, the shares are approximately $84.07\%$ and $84.01\%$, respectively; all other rows exceed $95\%$. This case study extends the six-dataset survey with Exchange, Weather, Solar, PEMS04, and PEMS08; it is a separate two-model cohort.

### A.4 Global Test-Oracle Temporal Rescaling

To quantify the improvement attainable from a single global rescaling of temporal variation, we fit one coefficient for each model–dataset pair using the test targets. Let $\overline{\widehat{y}}_{i,c}$ denote the temporal mean of the prediction for window $i$ and channel $c$. The rescaled prediction is

$$\widehat{\mathbf{y}}_{i,c}^{\,\mathrm{oracle}} = \overline{\widehat{y}}_{i,c}\mathbf{1}_F + \beta_{\mathrm{temp}}\big(\widehat{\mathbf{y}}_{i,c} - \overline{\widehat{y}}_{i,c}\mathbf{1}_F\big), \tag{15}$$

where a single $\beta_{\mathrm{temp}}$ is chosen to minimize MSE over all test predictions for that model–dataset pair. This transformation preserves the temporal mean of each predicted window and changes only its variation around that mean.

Table 4 reports the resulting forecasting gains. Despite using the test targets to choose the optimal global coefficient, the improvement is limited in most settings: the median MSE gain across the 48 model–dataset pairs is $0.361\%$, and 39 of the 48 gains are below $2\%$. The largest improvement is $5.711\%$. These results indicate that correcting global temporal amplitude alone leaves most of the forecasting error unchanged.

Table 4: Global test-oracle temporal rescaling gains (%) for 96-to-96 forecasting. Each entry reports MSE gain / MAE gain. One coefficient $\beta_{\mathrm{temp}}$ is fitted per model–dataset pair using the test targets while preserving each window's per-channel temporal mean.

| Model | ETTm1 | ETTm2 | ETTh1 | ETTh2 | Electricity | Traffic |
|---|---|---|---|---|---|---|
| Crossformer | 1.296 / 0.835 | 0.135 / 0.128 | 0.336 / 0.599 | 1.541 / 1.429 | 0.512 / 0.316 | 0.041 / −0.684 |
| DLinear | 0.184 / 0.377 | 0.677 / −0.476 | 0.523 / −0.275 | 3.377 / −1.157 | 0.001 / −0.024 | 0.313 / −0.338 |
| DUET | 1.816 / 2.533 | 0.134 / 0.511 | 0.095 / −0.150 | 0.028 / 0.047 | 0.267 / −0.159 | 0.379 / −1.991 |
| FITS | 3.223 / 3.270 | 0.080 / 0.249 | 5.711 / 4.380 | 0.004 / 0.025 | 2.213 / 4.456 | 2.956 / 8.810 |
| Informer | 0.092 / 0.034 | 2.157 / 0.509 | 0.553 / 0.316 | 3.816 / 2.751 | 1.196 / 0.728 | 4.322 / −0.159 |
| PDF | 0.434 / 0.481 | 0.212 / 0.627 | 0.012 / −0.042 | 0.342 / 0.354 | 0.062 / −0.166 | 0.058 / 0.477 |
| PatchTST | 1.400 / 1.787 | 0.235 / 0.649 | 0.004 / −0.038 | 0.005 / −0.004 | 2.278 / 4.202 | 0.675 / 1.563 |
| iTransformer | 0.096 / −0.037 | 0.063 / −0.092 | 0.385 / 0.138 | 0.632 / 0.553 | 0.130 / −0.052 | 0.513 / 1.313 |