When Pointwise Forecast Errors Are Not Enough: An Empirical Study of Temporal Alignment Metrics for Time Series Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Mean squared error (MSE) and mean absolute error (MAE) are the standard metrics used to evaluate time series forecasting models. Although these metrics are useful, they compare predictions and ground truth at fixed timestamps and can miss important failures on rapidly varying series. In particular, a model may obtain a strong MSE or MAE while smoothing sharp peaks, missing deep troughs, shifting ridges in time, or delaying abrupt changes. This paper studies this issue empirically by evaluating five forecasting models: DLinear, PatchTST, TimeMixer, iTransformer, and Chronos-2; using MSE, MAE, Dynamic Time Warping (DTW), and the Temporal Distortion Index (TDI). We compare these metrics on standard forecasting benchmarks and scientific network telemetry from ESnet, with emphasis on cases where local extrema and short-term temporal structure are important. Our results show that pointwise errors can give an incomplete view of model behavior: some forecasts score well under MSE and MAE while visibly smoothing or shifting peaks and troughs, whereas other forecasts better preserve local structure but receive worse pointwise scores. DTW and TDI help expose these differences by measuring shape similarity and temporal misalignment, respectively. We do not argue that DTW and TDI should replace MSE and MAE or that they are sufficient for every forecasting task. Rather, we show that they are useful diagnostic metrics when the timing and shape of peaks, troughs, and ridges matter.

1 Introduction

Time series forecasting models are commonly evaluated using pointwise error metrics such as mean squared error (MSE) and mean absolute error (MAE). These metrics are simple, interpretable, and useful for comparing predictions against ground truth values at fixed timestamps. They also align naturally with the standard supervised learning formulation of forecasting, where the objective is to predict a sequence of future values from a history window. For these reasons, MSE and MAE remain the dominant metrics in many long-term time series forecasting benchmarks.

However, pointwise errors do not fully describe forecast quality when the series contains sharp peaks, deep troughs, abrupt transitions, sustained elevated segments, or other short-lived structures. In such cases, a forecast can be close on average while still missing behavior that is important for interpretation or downstream use. For example, a model may smooth a peak, delay a trough, understate a short burst, or shift an abrupt transition in time. These errors can be difficult to distinguish from MSE and MAE alone because both metrics compare values only at matching timestamps. A prediction that has the correct shape but is slightly shifted can be penalized heavily, while a smoother prediction that suppresses local variation can sometimes obtain a favorable pointwise error.

This issue matters in domains where local temporal structure carries useful information. Particle-physics experiments such as NOvA (Adamson et al., 2016) motivate this concern. NOvA detectors are sensitive to neutrinos from a galactic core-collapse supernova, which would appear as a short burst of candidate events over a small time window rather than as a smooth long-term trend. In such a setting, smoothing or

shifting a burst could obscure the event timing and shape that the analysis is meant to preserve. Although the ESNet (Carder et al., 2022) data from NovA used in our experiments are network telemetry rather than detector-event counts, they come from a scientific-data setting where abrupt bandwidth changes and high-amplitude spikes can correspond to operationally important behavior. Similar concerns arise in financial volatility forecasting, where the timing of volatility bursts can affect risk management, trading decisions, and portfolio exposure. In these settings, the question is not only whether the forecast is numerically close at each timestamp, but also whether it preserves the timing and shape of important temporal structures.

This paper studies the extent to which temporal-alignment metrics provide additional information beyond MSE and MAE. We focus on Dynamic Time Warping (DTW) (Berndt & Clifford, 1994) and the Temporal Distortion Index (TDI) (Frías-Paredes et al., 2016). DTW measures how closely two sequences can be aligned when local stretching and compression of time are allowed. TDI summarizes how far the resulting alignment deviates from matching timestamps directly. Together, these metrics separate two aspects of forecast quality that pointwise errors conflate: whether the predicted trajectory has a similar shape to the ground truth, and whether the relevant structures occur at the right time.

Our goal is not to argue that DTW and TDI are universally better than MSE and MAE. They are not replacements for pointwise error metrics, and they do not solve all evaluation problems in forecasting. In particular, they do not by themselves address probabilistic calibration, rare-event accuracy, or application-specific decision costs. Instead, we treat DTW and TDI as diagnostic metrics. They are useful when the timing and shape of peaks, troughs, bursts, and abrupt transitions matter, and they should be interpreted alongside MSE, MAE, visual inspection, and domain-specific criteria where available.

We conduct a controlled empirical study using five forecasting models: DLinear (Zeng et al., 2022), PatchTST (Nie et al., 2023), TimeMixer (Wang et al., 2024), iTransformer (Liu et al., 2024), and Chronos-2 (Ansari et al., 2025). These models cover a range of commonly used forecasting approaches, including linear decomposition, patch-based transformers, multiscale mixing models, inverted transformers, and pretrained time series foundation models. We evaluate them on standard forecasting benchmarks, including ETT (Zhou et al., 2021) and Weather (Kolle) datasets, as well as ESNet network telemetry associated with scientific data movement. For each setting, we report MSE, MAE, DTW, and TDI and compare the numerical results with last-horizon forecast plots.

The results show that the four metrics can lead to different interpretations of the same forecast. In some cases, a model with strong MSE and MAE produces a visibly smoothed prediction that suppresses peaks and troughs or shifts abrupt transitions in time. In other cases, a model better preserves local variation but receives a worse pointwise score because of amplitude errors or small phase shifts. DTW and TDI help identify these differences by measuring shape similarity and timing distortion explicitly. The ESNet case further shows that alignment-aware metrics are not sufficient on their own for highly intermittent, high-dynamic-range series: when rare spikes dominate the visual behavior of the signal, additional event-aware or tail-sensitive criteria may be needed.

This paper makes three contributions. First, we provide a self-contained discussion of pointwise error, shape alignment, and timing distortion in time series forecast evaluation. Second, we empirically compare MSE, MAE, DTW, and TDI across multiple forecasting models, datasets, and prediction horizons, showing cases where these metrics emphasize different forecast failures. Third, we discuss the scope and limitations of alignment-aware evaluation, including when DTW and TDI are informative, when pointwise errors remain central, and why rare-event settings may require additional task-specific metrics.

2 Background

2.1 Error metrics for time series forecasting

Most time series forecasting studies report pointwise errors such as mean squared error (MSE) and mean absolute error (MAE). Let $y = (y_1, \ldots, y_T)$ denote the ground-truth sequence over a prediction horizon of

length T, and let $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T)$ denote the corresponding forecast. The two standard pointwise metrics are

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2, \qquad \mathrm{MAE}(y, \hat{y}) = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|. \tag{1}$$

These metrics compare the prediction and ground truth at the same timestamp. MSE gives larger weight to large errors because the residual is squared, while MAE grows linearly with the absolute residual. Both are useful measures of pointwise accuracy, but neither metric distinguishes between an error caused by incorrect amplitude and an error caused by a small temporal shift.

To measure shape similarity under temporal shifts, we also consider Dynamic Time Warping (DTW). Define the pairwise cost matrix $C \in \mathbb{R}^{T\times T}$ by

$$C_{i,j} = d(y_i, \hat{y}_j), \tag{2}$$

where $d(\cdot,\cdot)$ is a local discrepancy, which we take to be the absolute difference or squared difference depending on the implementation. A warping path is a sequence $\pi = ((i_1, j_1), \ldots, (i_L, j_L))$ satisfying boundary, monotonicity, and step-size constraints:

$$(i_1, j_1) = (1, 1), \qquad (i_L, j_L) = (T, T), \tag{3}$$

and

$$(i_{\ell+1} - i_\ell,\ j_{\ell+1} - j_\ell) \in \{(1,0), (0,1), (1,1)\} \quad \text{for } \ell = 1, \ldots, L-1. \tag{4}$$

Let $\mathcal{A}(T, T)$ denote the set of all valid warping paths between two length-T sequences. The DTW distance is the minimum accumulated alignment cost:

$$\mathrm{DTW}(y, \hat{y}) = \min_{\pi\in\mathcal{A}(T,T)} \sum_{(i,j)\in\pi} C_{i,j}. \tag{5}$$

Equivalently, it can be computed by dynamic programming. If $D_{i,j}$ is the optimal cost for aligning prefixes $(y_1, \ldots, y_i)$ and $(\hat{y}_1, \ldots, \hat{y}_j)$, then

$$D_{i,j} = C_{i,j} + \min\{D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1}\}, \tag{6}$$

with the usual boundary initialization. The DTW value is $D_{T,T}$, possibly normalized by the path length or horizon length. A low DTW score indicates that the two sequences can be matched with small cumulative cost after allowing local stretching or compression of the time axis.

DTW measures the cost of the best alignment, but it does not by itself indicate how far that alignment moves away from matching equal timestamps. For this reason, we also use the Temporal Distortion Index (TDI), following the idea of measuring the temporal displacement induced by the DTW alignment. Let $\pi^\star$ be an optimal DTW path:

$$\pi^\star \in \arg\min_{\pi\in\mathcal{A}(T,T)} \sum_{(i,j)\in\pi} C_{i,j}. \tag{7}$$

The TDI is then defined as the average squared deviation of the alignment path from the diagonal:

$$\mathrm{TDI}(y, \hat{y}) = \frac{1}{|\pi^\star|\, T^2} \sum_{(i,j)\in\pi^\star} (i - j)^2. \tag{8}$$

The factor T^2 makes the score comparable across horizons, while $|\pi^\star|$ averages over the number of matched pairs in the path. A forecast whose events occur at the correct times will have an alignment path near the diagonal and therefore a small TDI. A forecast that has a similar shape but is delayed or advanced in time can have a larger TDI even when its DTW score is relatively small.

The four metrics therefore emphasize different aspects of forecast quality. MSE and MAE measure pointwise error at fixed timestamps. DTW measures how well the forecast can be aligned to the target sequence when local time shifts are allowed. TDI measures the amount of temporal displacement required by that alignment.

In rapidly varying series, these distinctions are important: a smoothed forecast can have good pointwise error while suppressing peaks and troughs, whereas a forecast with the right local shape but a small phase shift can be penalized strongly by MSE and MAE.

A practical limitation of standard DTW is its quadratic time and memory cost in the sequence length. This can be burdensome for long horizons or for large benchmark suites. Several alternatives have been proposed to reduce this cost. FastDTW (Salvador & Chan, 2007) approximates DTW through a multiresolution procedure, although later work cautions that it is not always faster or preferable to optimized exact DTW in realistic settings (Wu & Keogh, 2021). Other variants, such as Segmental DTW (Tsai, 2021) and Tralie & Dempsey (2020), exploit alternative decompositions or parallel computation to accelerate sequence alignment. These methods are relevant options when alignment metrics must be computed at larger scale. In this study, we use FastDTW because even though it is slower than standard DTW in edge cases, it is faster on average and has a well-supported Python package.

2.2 Forecasting models considered

Our empirical study covers five modern forecasting models that represent a range of architectural choices, from simple linear baselines to pretrained foundation models.

DLinear. DLinear is a decomposition-based linear model derived from the LTSF-Linear family proposed by Zeng et al. (2022). The method splits each input window into a slowly varying trend component and a residual component, and then applies separate linear mappings to each part to generate multi-step forecasts. Despite its simplicity, this approach achieves competitive results on standard long-term forecasting benchmarks with low computational cost.

PatchTST. PatchTST is a transformer-based model that treats patches of each univariate series as tokens rather than individual time points (Nie et al., 2023). For each channel, the model segments the history into fixed-length patches, embeds these patches, and applies a shared transformer backbone in a channel-independent manner. This patching strategy allows attention over longer effective contexts while preserving local structure within each patch and reducing the number of tokens compared with pointwise encodings.

TimeMixer. TimeMixer is an MLP-based architecture designed for long-term time series forecasting (Wang et al., 2024). It uses structured mixing blocks that decompose the signal into multiple temporal scales and then combine information across these scales. The Past-Decomposable-Mixing blocks focus on extracting multiscale representations from the history, while the Future-Multipredictor-Mixing blocks integrate several predictors for future steps. By relying on carefully designed mixing operations instead of self-attention, TimeMixer aims to balance accuracy, efficiency, and scalability.

iTransformer. iTransformer reinterprets the transformer architecture for multivariate time series by inverting the usual roles of time and variables (Liu et al., 2024). Instead of using time steps as tokens, iTransformer treats each variate as a token whose feature vector consists of its full historical segment. Self-attention is then applied across variates, which emphasizes cross-series relationships, while per-token feed-forward networks process the temporal dimension implicitly. This design improves the handling of long lookback windows and has been shown to perform strongly on standard benchmarks.

Chronos-2. Chronos-2 is a pretrained time series foundation model that casts forecasting as sequence modeling over discretized time series tokens (Ansari et al., 2025). It uses an encoder-only transformer architecture trained at scale on synthetic and real time series data, and it outputs multi-step probabilistic forecasts, such as quantiles, in a zero-shot setting. A single Chronos-2 model can be applied to a variety of tasks, including univariate and multivariate forecasting, without task-specific retraining, which makes it a representative example of the emerging class of time series foundation models.

Taken together, these five models cover linear decompositions (DLinear), patch-based transformers (PatchTST), multiscale MLP mixing (TimeMixer), inverted transformers (iTransformer), and large pretrained encoders

(Chronos-2). Evaluating all of them under both pointwise and alignment-aware metrics allows us to study whether conclusions based on MSE and MAE alone agree with conclusions based on DTW and TDI.

2.3 CONTIME and evaluation with DTW and TDI

The CONTIME framework directly targets prediction delay in time series forecasting (Jhin et al., 2024). The authors propose a continuous-time recurrent architecture based on neural ordinary differential equations, and they introduce training objectives that encourage forecasts to respond promptly to changes in the input series. A key aspect of their evaluation protocol is the use of MSE, DTW, and TDI together. MSE captures traditional pointwise accuracy, DTW measures shape similarity under flexible time warping, and TDI summarizes temporal distortion along the alignment path.

The CONTIME results show that models with similar MSE can have markedly different DTW and TDI scores, and that lower TDI correlates with smaller prediction delays and better tracking of peaks and transitions. This provides concrete evidence that alignment-aware metrics reveal aspects of performance that are not visible from MSE alone, and it motivates our broader argument that DTW and TDI should become routine components of evaluation for modern time series forecasting models.

3 Experimental Setup

We test five state-of-the-art time series forecasting models on four benchmark datasets: ETTh1, ETTm1, ETTm2, and Weather; and on real-world network telemetry data from the NOvA neutrino experiments. We fix the sequence length, or lookback window, to 336 for all models and evaluate their performance on prediction lengths of 96, 336, and 720. We omit evaluation on prediction length 192, which is commonly tested, as we aim to show results for small, mid-range, and long-range prediction lengths specifically. Our experiments are run on an Nvidia T4 GPU. We use the NeuralForecast (Olivares et al., 2022) library to run all models apart from Chronos-2. Our code is publicly available at [1].

We compute MSE, MAE, DTW, and TDI scores for a given prediction length and dataset from each model. It should be noted that our score computation is univariate, i.e., the scores are computed only for the target column. In addition, score computation is performed on selected forecast horizons rather than being aggregated across all variates and all possible prediction windows in the test set. The state-of-the-art papers generally report MSE and MAE aggregated across all variates and multiple prediction horizons in the test dataset. However, since we are interested in showcasing visual comparisons alongside the performance metrics, we do not perform score computation in this way. Considering this, the MSE and MAE scores computed here do not exactly match the formal MSE and MAE scores reported in the papers of the state-of-the-art models, but we do perform an apples-to-apples comparison between models in the local context of our experiments, since the experimental setup is the same for all models. All metrics are reported after applying standard scaling for consistency across datasets.

To address the concern that reporting scores from a single forecast horizon may not adequately reflect uncertainty in the evaluation, we additionally compute each metric over multiple forecast origins. For each dataset, model, and prediction length, we evaluate up to five non-overlapping forecast windows from the test region. The final reported value is the mean score across these windows, and the uncertainty is reported as the corresponding standard deviation. This uncertainty should be interpreted as variability across forecast windows rather than as stochastic model uncertainty or confidence intervals over model parameters. For Chronos-2, we use the median forecast as the point prediction for MSE, MAE, DTW, and TDI, while its available quantile forecasts are used only for visualizing predictive intervals where applicable.

For computational efficiency, DTW is computed using FastDTW rather than exact dynamic programming DTW. This choice is especially important for longer prediction lengths and for repeated evaluation across multiple forecast windows. The DTW score is normalized by the alignment path length. TDI is computed from the same alignment path and measures the average temporal displacement between matched points,

[1] `https://anonymous.4open.science/r/Rethink_time_series_forecasting_metrics-262E`

Table 1: ETTh1 results for prediction length 96 on the specific forecast horizon shown in Figure 1 (scaled metrics; lower is better). Scores are computed only for the univariate target column.

Model	MSE	MAE	DTW	TDI
Chronos-2	0.0214	0.1142	0.0720	0.0375
DLinear	0.0841	0.2354	0.1548	0.0020
PatchTST	0.0775	0.2186	0.1476	0.0011
TimeMixer	0.0726	0.2237	0.1428	0.0012
iTransformer	0.0356	0.1548	0.0964	0.0214

normalized by the squared prediction length. Thus, MSE and MAE measure pointwise amplitude error, while DTW and TDI provide additional information about shape similarity and temporal alignment.

4 Results

Before beginning the discussion of results, it should be noted that since we ran experiments across datasets and prediction lengths, there is a lot of result data to report. Considering this, we report results only for selected datasets of interest for each prediction length in this section. Our full results can be found in our code repository [1].

4.1 Prediction Length 96

Figure 1 shows one representative test horizon on the ETTh1 dataset for a prediction length of 96. All models see the same history, but they differ markedly in how they extrapolate the next four days. Chronos-2 tracks the broad level and daily periodicity, but its forecast is smoother than the ground truth and does not reproduce several smaller local variations. iTransformer is also visually smooth and misses many of the short-term fluctuations. DLinear, by contrast, follows some of the local up-and-down structure more closely on this plotted horizon, although it underestimates the level in the later part of the forecast window.

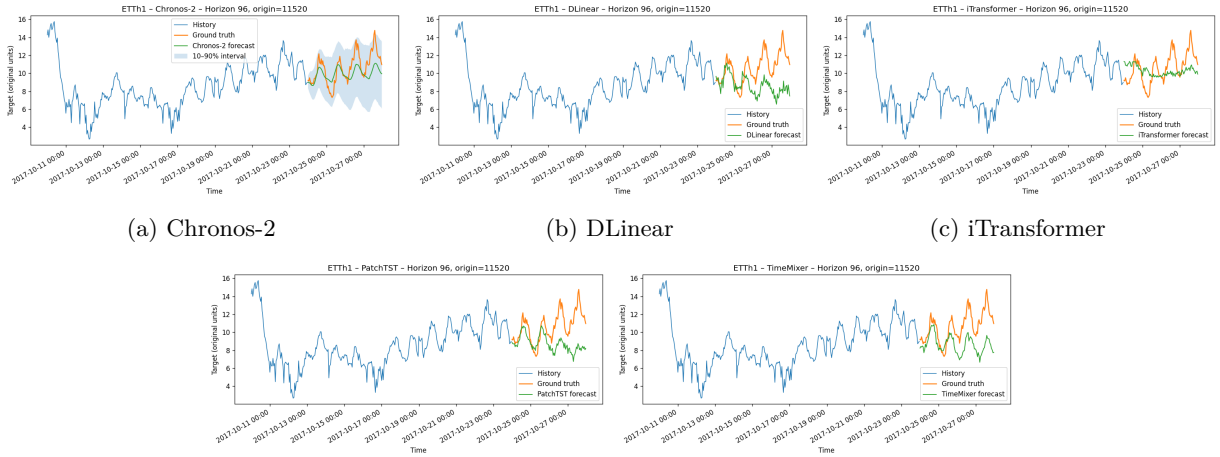
Table 1 reports the scaled metrics for the exact forecast horizon shown in Figure 1. On this specific horizon, Chronos-2 obtains the lowest MSE, MAE, and DTW. iTransformer is second-best under these three metrics. However, neither Chronos-2 nor iTransformer is the most visually faithful to the short-term fluctuations in the plotted target series. PatchTST obtains the lowest TDI on this horizon, followed closely by TimeMixer and DLinear. This shows that the metric conclusions depend on which aspect of forecast quality is emphasized: Chronos-2 is strongest under pointwise errors and DTW for this horizon, while the models with lower TDI keep the warping path closer to the diagonal.

Table 2 reports the updated rolling-origin evaluation. Each score is reported as the mean and standard deviation over five non-overlapping forecast windows from the test region. Under this evaluation, iTransformer obtains the lowest average MSE, MAE, and DTW, while DLinear obtains the lowest average TDI.

The visual comparison and the tables therefore support a more careful interpretation. Chronos-2 is best under MSE, MAE, and DTW on the plotted horizon, and iTransformer is best under those metrics on average across windows. However, both forecasts are visually smoother than the target and miss several short-term fluctuations. DLinear has worse MSE, MAE, and DTW on the plotted horizon, but visually follows some of the local variation more closely, even though it underestimates the later values. This is the type of discrepancy our evaluation is meant to highlight: pointwise errors, DTW, TDI, and visual behavior can emphasize different aspects of forecast quality. Reporting both the single-horizon metrics and the rolling-origin mean $\pm$ standard deviation helps separate the behavior of the plotted example from the average behavior across test windows.

4.2 Prediction length 336

Figure 2 shows a representative test horizon on the ETTm1 dataset with a prediction length of 336. The behavior in this case is different from the ETTh1 horizon-96 example because the visual comparison and



(a) Chronos-2 (b) DLinear (c) iTransformer

(d) PatchTST (e) TimeMixer

Figure 1: Forecasts on ETTh1 for prediction length 96 on a representative test horizon. In each panel, the blue curve shows the input history, the orange curve shows the ground truth over the next 96 steps, and the green curve shows the model forecast. For Chronos-2, the shaded band indicates the 10–90% predictive interval. Chronos-2 and iTransformer obtain strong pointwise scores, but both forecasts are visually smooth and miss several short-term fluctuations. DLinear follows some of the local up-and-down structure more closely on this horizon, although it underestimates the level in the later part of the forecast window.

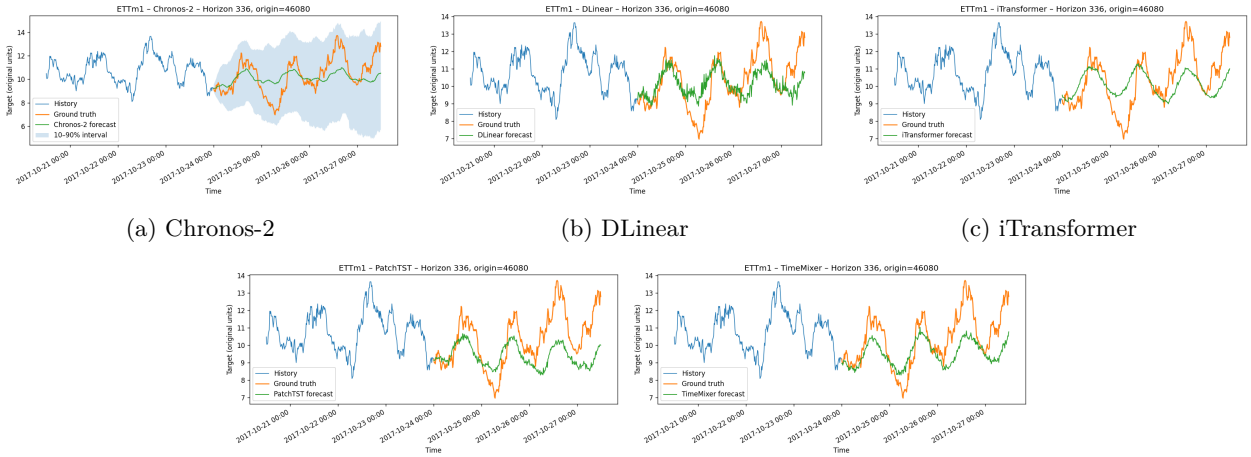
Table 2: ETTh1 results for prediction length 96 using the rolling-origin evaluation (scaled metrics; lower is better). Scores are computed only for the univariate target column and are reported as mean ± standard deviation over five non-overlapping forecast windows from the test region. Thus, these values do not match the aggregate MSE/MAE reported in prior work over all variates and all test horizons. However, the evaluation setup is identical for all models, so the comparison here is apples to apples.

Model	MSE	MAE	DTW	TDI
Chronos-2	0.0408 ± 0.0326	0.1508 ± 0.0768	0.1036 ± 0.0695	0.0268 ± 0.0351
DLinear	0.0404 ± 0.0282	0.1532 ± 0.0575	0.0807 ± 0.0490	0.0108 ± 0.0076
PatchTST	0.0465 ± 0.0330	0.1671 ± 0.0631	0.0897 ± 0.0472	0.0218 ± 0.0329
TimeMixer	0.0511 ± 0.0303	0.1745 ± 0.0700	0.0997 ± 0.0554	0.0243 ± 0.0327
iTransformer	0.0269 ± 0.0088	0.1307 ± 0.0208	0.0606 ± 0.0230	0.0260 ± 0.0214

the metrics are more consistent. DLinear is the closest model visually on this horizon: it follows the main local increases and decreases better than the other models, including the drop near the middle of the forecast window and the subsequent recovery. It still underestimates some of the larger peaks, but it captures the short-term structure more accurately than Chronos-2, iTransformer, PatchTST, or TimeMixer.

Table 3 reports the scaled metrics for the exact horizon shown in Figure 2. DLinear obtains the lowest MSE, MAE, DTW, and TDI on this plotted horizon. Chronos-2 is close to DLinear in MSE and MAE, and it has the second-lowest MAE, but it has a worse DTW score than DLinear, TimeMixer, and iTransformer, and it has the largest TDI among the five models. This matches the visual behavior: Chronos-2 captures the broad level but smooths over several local changes, whereas DLinear better follows the local structure of the target series.

Table 4 reports the updated rolling-origin evaluation. Each score is reported as the mean and standard deviation over five non-overlapping forecast windows from the test region. DLinear again obtains the lowest average MSE, MAE, DTW, and TDI. TimeMixer is second-best on average MSE, MAE, and DTW, but it has the largest average TDI. Chronos-2 remains relatively close to the best models in average MSE and



(a) Chronos-2 (b) DLinear (c) iTransformer

(d) PatchTST (e) TimeMixer

Figure 2: Forecasts on ETTm1 for prediction length 336 on a representative test horizon. The blue curve shows the input history, the orange curve shows the ground truth, and the green curve shows the model forecast. For Chronos-2, the shaded band indicates the 10–90% predictive interval. Chronos-2 tracks the broad level of the series but produces a smooth forecast that misses several short-term changes, including the deep drop around the middle of the forecast window and the larger peaks near the end. DLinear follows the local up-and-down structure most closely on this horizon, although it still underestimates some of the larger peaks. TimeMixer and iTransformer capture parts of the periodic structure but are smoother than the target, while PatchTST underestimates several later peaks.

Table 3: ETTm1 results for prediction length 336 on the specific forecast horizon shown in Figure 2 (scaled metrics; lower is better). Scores are computed only for the univariate target column.

Model	MSE	MAE	DTW	TDI
Chronos-2	0.0183	0.1073	0.0555	0.0024
DLinear	0.0165	0.0993	0.0379	0.0015
PatchTST	0.0285	0.1387	0.0592	0.0023
TimeMixer	0.0205	0.1133	0.0410	0.0018
iTransformer	0.0183	0.1074	0.0437	0.0016

MAE, but it has the worst average DTW. This is therefore not a case where the best pointwise model is visually poor. Instead, it shows why the additional metrics are still useful: MSE and MAE alone would make Chronos-2 appear close to the strongest baselines, while DTW shows that its smoothed trajectory is less faithful to the temporal structure of the sequence.

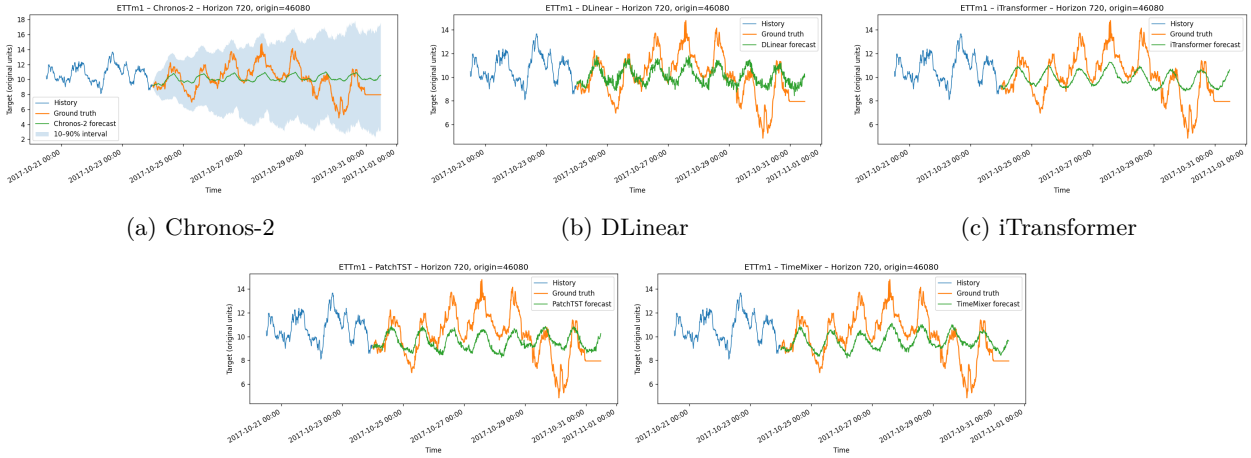
4.3 Prediction length 720

Figure 3 shows a representative test horizon on the ETTm1 dataset with a prediction length of 720. This longer horizon is challenging for all models, especially in the second half of the forecast window where the target contains larger peaks and a sharp drop. Chronos-2 captures the broad level of the series, but its forecast is much smoother than the ground truth. iTransformer is also smooth and does not reproduce several large local changes. DLinear is visually the closest on this horizon because it follows more of the local up-and-down structure, although it still underestimates some of the largest peaks and the final drop.

Table 5 reports the scaled metrics for the exact horizon shown in Figure 3. On this plotted horizon, DLinear obtains the lowest MSE, MAE, and DTW. iTransformer obtains the lowest TDI, meaning that its FastDTW alignment path stays closest to the diagonal, even though the forecast is visually smoother than the target.

Table 4: ETTm1 results for prediction length 336 using the rolling-origin evaluation (scaled metrics; lower is better). Scores are computed only for the univariate target column and are reported as mean ± standard deviation over five non-overlapping forecast windows from the test region. Thus, these values do not match the aggregate MSE/MAE reported in prior work over all variates and all test horizons. However, the evaluation setup is identical for all models, so the comparison here is apples to apples.

Model	MSE	MAE	DTW	TDI
Chronos-2	0.0265 ± 0.0135	0.1265 ± 0.0432	0.0761 ± 0.0441	0.0330 ± 0.0565
DLinear	0.0214 ± 0.0085	0.1172 ± 0.0288	0.0483 ± 0.0114	0.0221 ± 0.0229
PatchTST	0.0288 ± 0.0086	0.1375 ± 0.0285	0.0582 ± 0.0165	0.0366 ± 0.0482
TimeMixer	0.0257 ± 0.0208	0.1214 ± 0.0533	0.0531 ± 0.0318	0.0483 ± 0.0612
iTransformer	0.0266 ± 0.0140	0.1284 ± 0.0331	0.0590 ± 0.0237	0.0307 ± 0.0405



(a) Chronos-2 (b) DLinear (c) iTransformer

(d) PatchTST (e) TimeMixer

Figure 3: Forecasts on ETTm1 for prediction length 720 on a representative test horizon. History is shown in blue, ground truth in orange, and model forecasts in green; the Chronos-2 plot also includes a 10–90% predictive interval. Chronos-2 stays close to the broad level of the series but is substantially smoother than the ground truth and misses the larger peaks and the sharp drop near the end of the horizon. iTransformer is also comparatively smooth and underestimates several large excursions. DLinear follows the local up-and-down structure most closely on this horizon, although it still underestimates some of the largest peaks and the final drop. PatchTST and TimeMixer capture parts of the periodic structure but also damp several short-term variations.

Chronos-2 has the worst DTW on this horizon, while PatchTST has the worst TDI. These values match the visual picture more closely than in the horizon-96 case: DLinear is both visually strong and best under MSE, MAE, and DTW for the plotted horizon.

Table 6 reports the updated rolling-origin evaluation over five non-overlapping forecast windows. Under this evaluation, iTransformer obtains the lowest average MSE, MAE, and TDI. TimeMixer obtains the lowest average DTW, although it is very close to DLinear. DLinear is second-best on average MSE, MAE, and DTW. Chronos-2 has the worst average DTW, while PatchTST has the worst average MSE, MAE, and TDI. Thus, the average-window results differ from the plotted-horizon results: DLinear is strongest on the representative horizon under MSE, MAE, and DTW, but iTransformer has the best average pointwise errors across the five windows.

This case is useful because the metrics do not all point to the same model. On the plotted horizon, DLinear best matches the visible local structure and also has the best MSE, MAE, and DTW. Across five windows,

Table 5: ETTm1 results for prediction length 720 on the specific forecast horizon shown in Figure 3 (scaled metrics; lower is better). Scores are computed only for the univariate target column.

Model	MSE	MAE	DTW	TDI
Chronos-2	0.0346	0.1457	0.0827	0.0104
DLinear	0.0272	0.1317	0.0536	0.0058
PatchTST	0.0344	0.1535	0.0707	0.0771
TimeMixer	0.0307	0.1429	0.0589	0.0087
iTransformer	0.0299	0.1418	0.0610	0.0006

Table 6: ETTm1 results for prediction length 720 using the rolling-origin evaluation (scaled metrics; lower is better). Scores are computed only for the univariate target column and are reported as mean ± standard deviation over five non-overlapping forecast windows from the test region. Thus, these values do not match the aggregate MSE/MAE reported in prior work over all variates and all test horizons. However, the evaluation setup is identical for all models, so the comparison here is apples to apples.

Model	MSE	MAE	DTW	TDI
Chronos-2	0.0440 ± 0.0127	0.1725 ± 0.0332	0.0923 ± 0.0367	0.0089 ± 0.0085
DLinear	0.0379 ± 0.0119	0.1599 ± 0.0315	0.0663 ± 0.0152	0.0174 ± 0.0172
PatchTST	0.0625 ± 0.0391	0.2022 ± 0.0720	0.0985 ± 0.0312	0.0379 ± 0.0339
TimeMixer	0.0371 ± 0.0115	0.1588 ± 0.0292	0.0660 ± 0.0244	0.0360 ± 0.0526
iTransformer	0.0333 ± 0.0056	0.1511 ± 0.0171	0.0756 ± 0.0235	0.0082 ± 0.0067

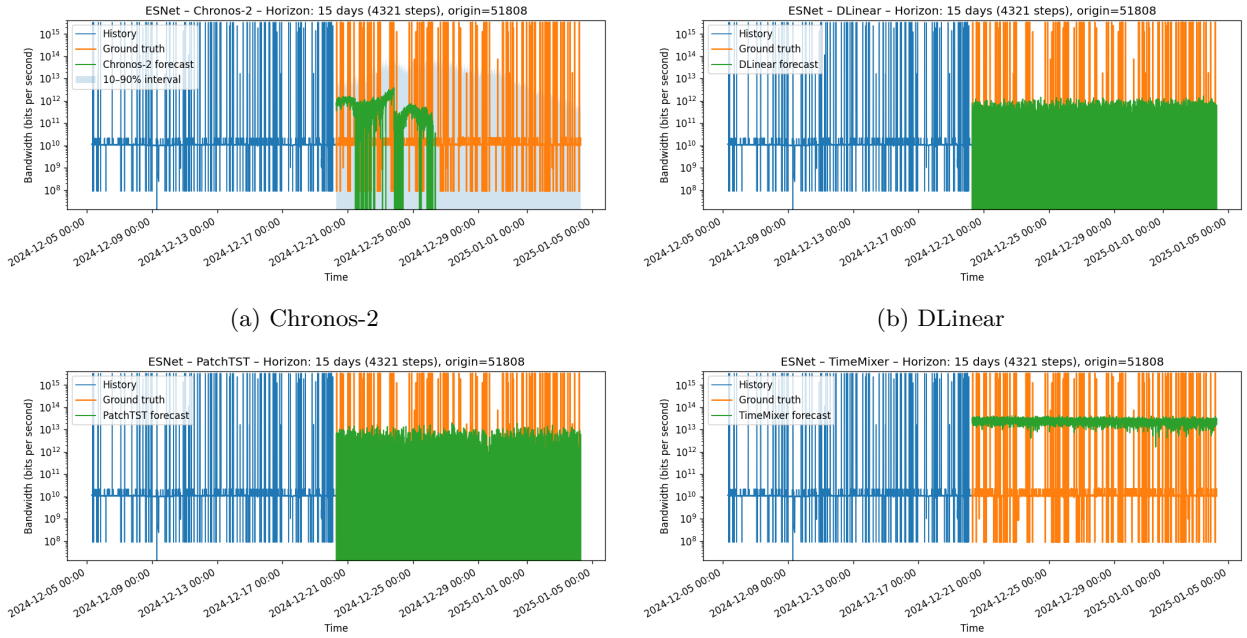
iTransformer has the best average MSE and MAE despite producing a smoother forecast in the representative plot. Its low TDI shows that the forecast is not strongly shifted in time, but TDI alone does not penalize the fact that several large fluctuations are damped. DTW, in contrast, gives stronger penalties to the models that fail to reproduce the shape of the sequence. This example reinforces the need to report multiple metrics: MSE and MAE summarize pointwise accuracy, DTW is more sensitive to the overall matched shape, and TDI separates timing displacement from amplitude error.

4.4 ESNet: 15-day horizons

ESNet (Carder et al., 2022) is a high-performance network that interconnects U.S. Department of Energy research sites and supports large-scale scientific data movement. In this case study, we evaluate forecasting behavior on ESNet bandwidth telemetry associated with data movement between Fermilab and ALCF. This experiment is not intended to claim that any evaluated model is operationally sufficient for network forecasting. Rather, it is included as a stress test for the metrics on a highly intermittent, heavy-tailed time series where visual forecast quality and aggregate scores can diverge.

Figure 4 shows the representative 15-day forecast horizon used for the single-horizon comparison in Table 7. The target contains many abrupt spikes and drops on a log scale. None of the models captures these dynamics reliably. Chronos-2 produces a comparatively smooth forecast and misses most of the large excursions. DLinear and PatchTST show substantially more short-term variation, but their fluctuations do not consistently match the timing or magnitude of the observed spikes. TimeMixer produces a high-amplitude band over much of the horizon and also does not track the ground truth spikes accurately.

Table 7 reports the scaled metrics for the exact horizon shown in Figure 4. On this horizon, TimeMixer obtains the lowest MSE, DLinear obtains the lowest MAE and DTW, and PatchTST obtains the lowest TDI. The differences in MSE are extremely small: all four models lie between 1.3054 and 1.3254. This is important because the plots show that all models fail to reproduce the most salient behavior of the series. The low TDI values for PatchTST and TimeMixer should also be interpreted carefully. They indicate that



(a) Chronos-2

(b) DLinear

(c) PatchTST

(d) TimeMixer

Figure 4: Forecasts on ESNet telemetry for a 15-day horizon, corresponding to 4321 time steps, shown on a log y-axis in bandwidth measured in bits per second. We visualize a 15-day pre-horizon context window followed by the 15-day forecast horizon. History is shown in blue, ground truth in orange, and model forecasts in green; the Chronos-2 plot also includes a 10–90% predictive interval. The target series contains abrupt excursions spanning several orders of magnitude. None of the models accurately reproduces these extreme changes. Chronos-2 is comparatively smooth and only partially covers the observed range. DLinear and PatchTST produce more rapid variation, while TimeMixer stays in a higher-amplitude band for much of the horizon.

Table 7: ESNet results for the specific 15-day horizon shown in Figure 4 (4321 steps; scaled metrics; lower is better). Scores are computed only for the univariate target column.

Model	MSE	MAE	DTW	TDI
Chronos-2	1.3222	0.1878	0.1649	0.0013
DLinear	1.3216	0.1863	0.1358	0.0028
PatchTST	1.3254	0.1974	0.1805	0.0000
TimeMixer	1.3054	0.2337	0.2251	0.0001

the FastDTW alignment path stays close to the diagonal, but this does not mean that the forecasts capture the spike amplitudes or the rare extreme events.

Table 8 reports the rolling-origin evaluation over three non-overlapping 15-day windows. TimeMixer obtains the lowest average MSE, DLinear obtains the lowest average MAE and DTW, and TimeMixer obtains the lowest average TDI. Chronos-2 has the largest average TDI and the largest standard deviation in DTW and TDI. However, the broader conclusion is not that one model solves the ESNet task. The average MSE values remain very close across models, while the visual forecasts remain poor for the extreme events. This supports a more cautious interpretation: MSE, MAE, DTW, and TDI expose different aspects of the forecast, but even this expanded metric set is incomplete for heavy-tailed network telemetry. For such data, additional event- or tail-sensitive measures may be needed to evaluate whether a model captures operationally important spikes.

Table 8: ESNet results for 15-day horizons using the rolling-origin evaluation (4321 steps; scaled metrics; lower is better). Scores are computed only for the univariate target column and are reported as mean ± standard deviation over three non-overlapping forecast windows. These metrics summarize pointwise error and alignment behavior, but they do not fully capture the visual failure to reproduce abrupt, multi-order-of-magnitude spikes.

Model	MSE	MAE	DTW	TDI
Chronos-2	1.4789 ± 0.1468	0.2066 ± 0.0182	0.1751 ± 0.0438	0.0151 ± 0.0251
DLinear	1.4777 ± 0.1460	0.2040 ± 0.0170	0.1524 ± 0.0145	0.0025 ± 0.0004
PatchTST	1.4794 ± 0.1433	0.2144 ± 0.0153	0.1891 ± 0.0075	0.0003 ± 0.0003
TimeMixer	1.4672 ± 0.1509	0.2314 ± 0.0055	0.2195 ± 0.0062	0.0001 ± 0.0000

5 Discussion and Limitations

Our evaluation is designed to compare metrics against visible forecast behavior, not to reproduce standard benchmark ranking settings. Metrics are computed only on the univariate target column, and the plotted examples use specific forecast horizons. This makes the relationship between tables and figures easier to inspect, but it also means that our MSE and MAE values are not directly comparable to aggregate multivariate benchmark scores reported in prior work.

The uncertainty values should also be interpreted narrowly. We report mean and standard deviation across non-overlapping forecast windows, so the uncertainty reflects sensitivity to the selected test horizon. It does not measure stochastic model uncertainty, parameter uncertainty, or full predictive calibration. Chronos-2 provides quantile forecasts, which we visualize, but the other baselines are deterministic. A fuller probabilistic evaluation could include prediction interval coverage probability (Sluijterman et al., 2024), pinball loss (Koenker & Bassett, 1978), or continuous ranked probability score (CRPS) (Matheson & Winkler, 1976).

DTW and TDI are useful diagnostics, but they do not remove the need for application-specific evaluation. DTW can favor forecasts that match after temporal warping, while TDI can be low when a forecast is well aligned in time but too smooth in amplitude. Other alignment- or shape-sensitive alternatives include Soft-DTW (Cuturi & Blondel, 2018), derivative DTW (Keogh & Pazzani), ERP (Chen & Ng, 2004), TWED (Marteau, 2008), and correlation-based measures.

The ESNet case study shows this limitation most clearly. All evaluated models miss abrupt multi-order-of-magnitude spikes, even though their aggregate scaled errors can be close. For heavy-tailed telemetry, event- or tail-sensitive metrics may be necessary, such as spike precision and recall and threshold exceedance error.

6 Conclusion

The state of the art in time series forecasting is currently judged largely through MSE and MAE. These metrics are useful and should not be discarded, but our results show that they are not always indicative of better forecast performance. A model can obtain strong pointwise errors while producing forecasts that are too smooth, miss local structure, or fail to represent important temporal behavior in the target series.

We therefore argue that forecasting evaluation should report complementary metrics alongside MSE and MAE. DTW and TDI provide one such pair of diagnostics: DTW measures similarity after temporal alignment, while TDI measures the temporal distortion required by that alignment. These metrics do not replace pointwise errors, but they expose behavior that MSE and MAE alone can miss.

More broadly, metric choice should depend on the forecasting task. As forecasting models are applied to scientific and operational data, model comparisons based only on MSE and MAE risk overstating practical progress.

References

P. Adamson, C. Ader, M. Andrews, N. Anfimov, and Anghel. First measurement of electron neutrino appearance in nova. *Phys. Rev. Lett.*, 116:151806, Apr 2016. doi: 10.1103/PhysRevLett.116.151806. URL https://link.aps.org/doi/10.1103/PhysRevLett.116.151806.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting, 2025. URL https://arxiv.org/abs/2510.15821.

Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, pp. 359–370. AAAI Press, 1994.

California Department of Transportation. California performance measurement system (pems) traffic data, san francisco bay area freeways. http://pems.dot.ca.gov. Hourly road occupancy rates from 862 sensors, 2015–2016. Dataset configuration as in Lai et al. (2018).

D. Carder, M. Dart, E. Graf, et al. Basic energy sciences network requirements review (final report). Technical report, Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA (United States), 11 2022. URL https://www.osti.gov/biblio/1899590.

Lei Chen and Raymond Ng. On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pp. 792–803. VLDB Endowment, 2004. ISBN 0120884690.

Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series, 2018. URL https://arxiv.org/abs/1703.01541.

Laura Frías-Paredes, Fermín Mallor, Teresa León, and Martín Gastón-Romeo. Introducing the temporal distortion index to perform a bidimensional analysis of renewable energy forecast. *Energy*, 94:180–194, 2016. ISSN 0360-5442. doi: https://doi.org/10.1016/j.energy.2015.10.093. URL https://www.sciencedirect.com/science/article/pii/S0360544215014619.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, 03 2007. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2007.00587.x. URL https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. doi: 10.1016/j.ijforecast.2006.03.001.

Sheo Yon Jhin, Seojin Kim, and Noseong Park. Addressing prediction delays in time series forecasting: A continuous gru approach with derivative regularization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 1234–1245. ACM, August 2024. doi: 10.1145/3637528.3671969. URL http://dx.doi.org/10.1145/3637528.3671969.

Eamonn J. Keogh and Michael J. Pazzani. *Derivative Dynamic Time Warping*, pp. 1–11. doi: 10.1137/1.9781611972719.1. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611972719.1.

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

Olaf Kolle. Weather station data, max planck institute for biogeochemistry, jena. https://www.bgc-jena.mpg.de/wetter/. Accessed: 2025-12-03.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting, 2024. URL https://arxiv.org/abs/2310.06625.

Pierre-François Marteau. Time warp edit distance, 2008. URL https://arxiv.org/abs/0802.3522.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2629907.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers, 2023. URL https://arxiv.org/abs/2211.14730.

Kin G. Olivares, Cristian Challú, Azul 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.

OpenAI. Chatgpt. https://chat.openai.com, 2026. Large language model accessed in May 2026.

Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5):561–580, October 2007. ISSN 1088-467X.

Laurens Sluijterman, Eric Cator, and Tom Heskes. How to evaluate uncertainty estimates in machine learning for regression? *Neural Networks*, 173:106203, May 2024. ISSN 0893-6080. doi: 10.1016/j.neunet.2024.106203. URL http://dx.doi.org/10.1016/j.neunet.2024.106203.

Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1), February 2011. ISSN 1350-7265. doi: 10.3150/10-bej267. URL http://dx.doi.org/10.3150/10-BEJ267.

Christopher Tralie and Elizabeth Dempsey. Exact, parallelizable dynamic time warping alignment with linear memory, 2020. URL https://arxiv.org/abs/2008.02734.

TJ Tsai. Segmental dtw: A parallelizable alternative to dynamic time warping. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 106–110, 2021. doi: 10.1109/ICASSP39728.2021.9413827.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting, 2024. URL https://arxiv.org/abs/2405.14616.

Renjie Wu and Eamonn J. Keogh. FastDTW is approximate and Generally Slower than the Algorithm it Approximates (Extended Abstract) . In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 2327–2328, Los Alamitos, CA, USA, April 2021. IEEE Computer Society. doi: 10.1109/ICDE51399.2021.00249. URL https://doi.ieeecomputersociety.org/10.1109/ICDE51399.2021.00249.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting?, 2022. URL https://arxiv.org/abs/2205.13504.

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 *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, volume 35, pp. 11106–11115. AAAI Press, 2021.