CLOCKUP: CLOCK-STRUCTURED ADDITIVE UNCERTAINTY MODELING FOR NON-STATIONARY PROBABILISTIC TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable probabilistic time series forecasting requires adapting to local distribution shifts while capturing recurring temporal structure. Although instance normalization mitigates variation in local level and scale, similar normalized histories can occur at different cycle positions even when their future distributions differ. Consequently, latent uncertainty models conditioned solely on normalized history may struggle to distinguish phase-dependent predictive distributions, a limitation we term *phase blindness*. To address this limitation, we propose **ClockUp**, a **Clock**-structured framework that explicitly incorporates temporal phase into latent **U**ncertainty modeling and **p**robabilistic forecasting. ClockUp combines a history-conditioned Gaussian state component with a Fourier-phase-conditioned Gaussian clock component to form an additive latent prior, enabling direct reparameterized sampling with closed-form mean and variance. A push-forward decoder further combines lead-time-specific and phase-conditioned mappings to generate forecasts non-iteratively. We train ClockUp using empirical CRPS and additionally investigate P-CRPS, a phase-normalized training variant designed to reduce scale-driven loss imbalance across phases. Experiments on seven real-world benchmarks demonstrate improved probabilistic and point forecasting accuracy. Together with phase-wise analyses on selected datasets, these results support explicit phase conditioning as a useful inductive bias for probabilistic forecasting and highlight the value of phase-wise calibration diagnostics.

1 INTRODUCTION

Probabilistic time series forecasting supports decision-making in energy management Ying et al. (2023); Terrén-Serrano et al. (2026), transportation planning Ermagun & Levinson (2018); Cao et al. (2026), and resource allocation Teng et al. (2020); Salinas et al. (2020) by characterizing the distribution of future observations Gneiting & Katzfuss (2014) rather than providing point estimates alone. Reliable forecasting must account for both evolving local conditions Kim et al. (2021); Liu et al. (2022) and recurring temporal structure Lin et al. (2024; 2025). These sources of variation are related but distinct: adapting to local distribution shifts does not, by itself, ensure that a model captures recurring phase-dependent differences in future behavior. An effective probabilistic forecaster should therefore accommodate window-specific dynamics while exploiting temporal structure shared across forecasting windows.

Instance normalization Kim et al. (2021); Liu et al. (2023); Ye et al. (2024) is widely used to mitigate variation in local level and scale. By centering and rescaling each lookback window, it encourages models to learn from relative temporal patterns rather than sample-specific magnitudes. However, normalized history may not uniquely determine the temporal phase (the position within a recurring cycle), as similar lookbacks can occur at different positions with differing future distributions. Although normalization does not necessarily remove phase information Zhou et al. (2021), a latent prior conditioned solely on normalized history must infer it from potentially ambiguous patterns. Calendar covariates provide complementary information and are commonly incorporated into encoders or decoders Rangapuram et al. (2018); Salinas et al. (2020). Rather than assuming such conditioning is insufficient, we investigate whether assigning cycle position a dedicated pathway in the latent prior provides a useful inductive bias for phase-sensitive uncertainty modeling.

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.

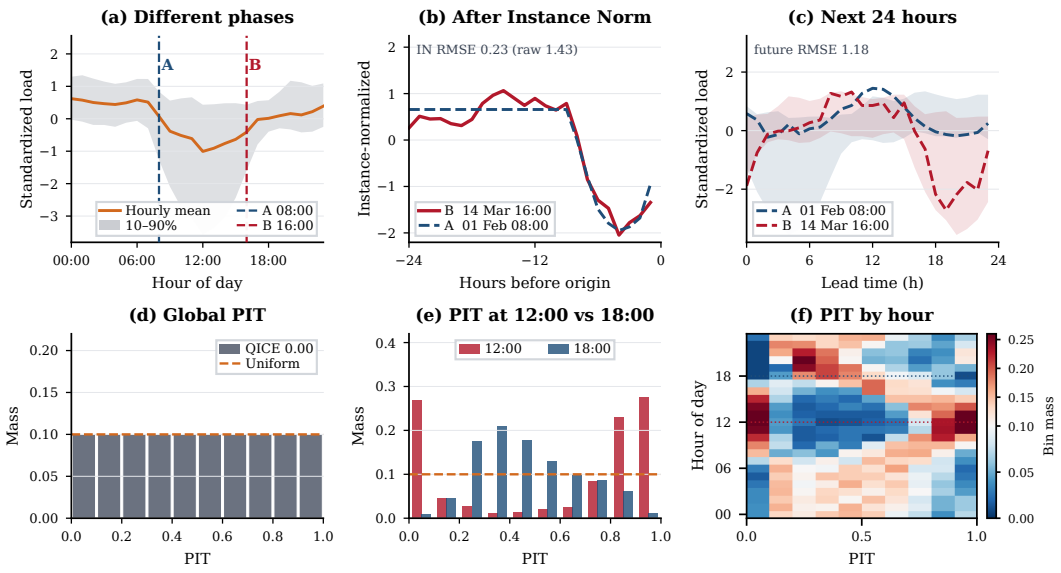


Figure 1: Illustration of phase blindness and phase-wise calibration on ETTh1. (a–b) Lookbacks at different daily phases become similar after IN (RMSE: 1.43 → 0.23), yet (c) their futures and phase-conditional 10%–90% quantile bands differ (RMSE: 1.18). (d–f) A pooled residual climatology (hourly means + a shared residual CDF) exhibits uniform pooled PIT bin occupancy (QICE = 0), masking midday underdispersion and evening overdispersion (mean phase-wise QICE = 0.053).

Figure 1 illustrates both this ambiguity and its implications for evaluation. In panels (a–c), two lookbacks from the HUFL channel of ETTh1 originate at different daily phases and become substantially more similar after instance normalization. Their subsequent trajectories nevertheless differ and traverse distinct phase-dependent quantile bands. This example motivates explicit phase conditioning without establishing that phase alone explains the differences. Panels (d–f) show that a reference forecast with a shared residual distribution exhibits uniform pooled probability integral transform (PIT) Gneiting & Katzfuss (2014) bin occupancy while displaying midday underdispersion and evening overdispersion. Thus, aggregate calibration can conceal opposing phase-specific errors. We refer to the difficulty of distinguishing phase-dependent predictive distributions when cycle position is inadequately represented as *phase blindness*.

To address this difficulty, we propose **ClockUp**, a **Clock**-structured framework for latent **U**ncertainty modeling and **p**robabilistic forecasting. ClockUp combines a Gaussian state component conditioned on normalized history with a Gaussian clock component parameterized by Fourier phase features. Conditional independence is imposed by construction, yielding an additive prior with closed-form mean and variance for direct reparameterized sampling. This prior provides a structured inductive bias rather than an identifiable decomposition of uncertainty. An additive push-forward decoder further combines lead-time-specific and phase-conditioned mappings, incorporating phase into both latent sampling and non-iterative forecast generation. ClockUp is trained with empirical CRPS; we additionally investigate ClockUp-V, which replaces this objective with a phase-normalized variant using fixed training-set dispersion estimates to reduce scale-driven loss imbalance across phases. Our main contributions are threefold:

- **Conceptual.** We articulate *phase blindness* as a potential limitation of uncertainty modeling conditioned solely on normalized history and illustrate how aggregate calibration can conceal phase-specific errors.
- **Technical.** We propose an additive state–clock Gaussian prior and a phase-conditioned push-forward decoder for non-iterative probabilistic forecasting, together with a phase-normalized CRPS training variant.
- **Empirical.** We evaluate ClockUp on seven real-world time series benchmarks, demonstrating its effectiveness in probabilistic and point forecasting relative to competing baselines. We further conduct phase-wise calibration diagnostics on selected datasets, complementing aggregate evaluation by examining calibration within temporal cycles.

2 Related Work

2.1 Probabilistic Time Series Forecasting

Probabilistic time series forecasting characterizes uncertainty through predictive distributions rather than point estimates. Likelihood-based methods, such as DeepAR Salinas et al. (2020), learn the parameters of prescribed distribution families, providing a direct way to model conditional uncertainty. Subsequent work Zheng et al. (2024); Zheng & Sun (2024) further incorporates temporal or cross-series dependencies to improve distributional modeling. Quantile-based methods Wen et al. (2017); Lim et al. (2021); Smyl et al. (2025) instead estimate selected predictive quantiles using pinball losses, avoiding the need to specify a full parametric output distribution. Generative approaches Rasul et al. (2021); Li et al. (2022); Shen & Kwok (2023); Li et al. (2024) offer an alternative by producing forecast samples through learned stochastic transformations, using variational Kingma & Welling (2013) or diffusion Sohl-Dickstein et al. (2015) mechanisms to capture complex future distributions. Recent work further considers non-stationarity and structural regularization in generative forecasting, for example through adaptive distributional statistics Ye et al. (2025) or parametric distributional structures Li et al. (2026). Within this sample-based paradigm, ClockUp jointly uses normalized historical information and known temporal phase to structure the latent distribution.

2.2 Temporal Structure and Prior Design

Temporal structure is commonly incorporated through calendar covariates, positional representations, and periodic embeddings Lim et al. (2021); Nie et al. (2023); Ma et al. (2026). Such mechanisms provide temporal information to forecasting models, while conditioning the latent prior on known cycle position offers a complementary design choice. A related line of work explores history-conditioned stochastic formulations and structured uncertainty modeling. For example, TMDM Li et al. (2024) and TSFlow Kollovieh et al. (2025) use data-dependent stochastic formulations in diffusion and flow-matching frameworks, while D^{3}U Li et al. (2025) decomposes deterministic and uncertain components within a conditional diffusion framework. PPM Li et al. (2026) constructs a history-conditioned Gaussian prior for one-pass generation, and TimeBridge Park et al. (2026) investigates diffusion-bridge formulations with soft and hard constraints. Together, these studies illustrate different ways to organize conditioning information and stochastic structure for forecasting. Complementing these approaches, ClockUp combines a history-conditioned Gaussian state with a Fourier-phase-conditioned Gaussian clock component in an additive latent prior. The clock component encodes recurring phase-dependent structure shared across windows, providing an explicit inductive bias for phase-sensitive uncertainty modeling.

3 Preliminaries

3.1 Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{V \times T}$ denote a historical window with V channels and T observations, ending at absolute time index a. The forecasting target is the subsequent trajectory $\mathbf{Y} \in \mathbb{R}^{V \times H}$, where H denotes the prediction horizon. Let $\boldsymbol{\tau} \in \mathbb{R}^{H}$ denote the known temporal phases associated with these future steps. The goal is to learn the conditional predictive distribution of $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\tau})$. A latent-variable formulation represents this distribution by marginalizing over a latent variable $\mathbf{z}$:

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\tau}) = \int p_\phi(\mathbf{Y} \mid \mathbf{z}; \mathbf{X}, \boldsymbol{\tau}) p_\theta(\mathbf{z} \mid \mathbf{X}, \boldsymbol{\tau})\, d\mathbf{z}. \tag{1}$$

Here, p_θ denotes the conditional prior and p_ϕ denotes the conditional distribution of future observations given the latent variable and forecasting context. A conditional prior allows latent uncertainty to adapt directly to the available context. ClockUp instantiates this formulation with an additive prior comprising a history-conditioned Gaussian state and a clock-conditioned Gaussian clock component. Forecast samples are generated through a deterministic push-forward decoder:

$$\hat{\mathbf{Y}} = g_\phi(\mathbf{z}, \mathbf{X}, \boldsymbol{\tau}), \quad \mathbf{z} \sim p_\theta(\mathbf{z} \mid \mathbf{X}, \boldsymbol{\tau}). \tag{2}$$

In this case, the conditional distribution given z is concentrated at the decoder output, and predictive uncertainty is induced by sampling from the latent prior. This construction supports sample-based forecasting without requiring an explicit predictive likelihood.

3.2 Probabilistic Scoring and Calibration

We assess probabilistic forecasts using a proper scoring rule and a complementary calibration diagnostic. For a scalar predictive cumulative distribution function (CDF) F and an observation y, the continuous ranked probability score (CRPS) Matheson & Winkler (1976) is defined as:

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(u) - \mathbb{I}(y \le u)\right)^2 du, \tag{3}$$

where $\mathbb{I}(y \le u)$ is an indicator function that equals one if $y \le u$ and zero otherwise. For the empirical distribution of K equally weighted forecasting samples, CRPS can be computed as:

$$\mathrm{CRPS}(F_K, y) = \frac{1}{K}\sum_{k=1}^{K} |\hat{y}_k - y| - \frac{1}{2K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} |\hat{y}_i - \hat{y}_j|. \tag{4}$$

Calibration is assessed using quantile interval coverage error (QICE) Han et al. (2022). For a continuous predictive CDF, the PIT value is $u = F(y)$. If F matches the true conditional distribution, the PIT is uniform conditional on the forecasting context. QICE provides a discretized diagnostic by measuring deviations from uniform occupancy across S equal-probability predictive intervals:

$$\mathrm{QICE} = \frac{1}{S}\sum_{s=1}^{S}\left| r_s - \frac{1}{S}\right|, \quad r_s = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}(y_n \in \mathcal{I}_{n,s}), \tag{5}$$

where $\mathcal{I}_{n,s}$ denotes the s-th quantile interval of the predictive distribution for the n-th data point, and N is the number of evaluated points. CRPS assesses overall distributional accuracy, whereas QICE diagnoses quantile-interval coverage discrepancies; computing these metrics within temporal phase groups can reveal calibration differences obscured by aggregate evaluation.

4 Proposed Method: ClockUp

ClockUp combines normalized local-state information with known temporal phase to construct a structured latent prior for probabilistic time series forecasting. The overall framework is illustrated in Figure 2. A state encoder f_s estimates a Gaussian component $\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s^2)$ from the normalized lookback, while a clock encoder f_c estimates a complementary Gaussian component $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$ from Fourier phase features. Their analytical composition supplies latent samples to an additive push-forward decoder g_ϕ with lead-time and clock-conditioned mappings. The entire network is optimized end-to-end using empirical CRPS and a phase-normalized variant that reduces scale-driven imbalance across temporal phases.

4.1 Additive State–Clock Prior Design

State Prior. Time series data $\mathbf{X}$ often exhibit substantial variation in local level and scale. To prevent sample-specific magnitudes from dominating the latent representation, we first apply instance

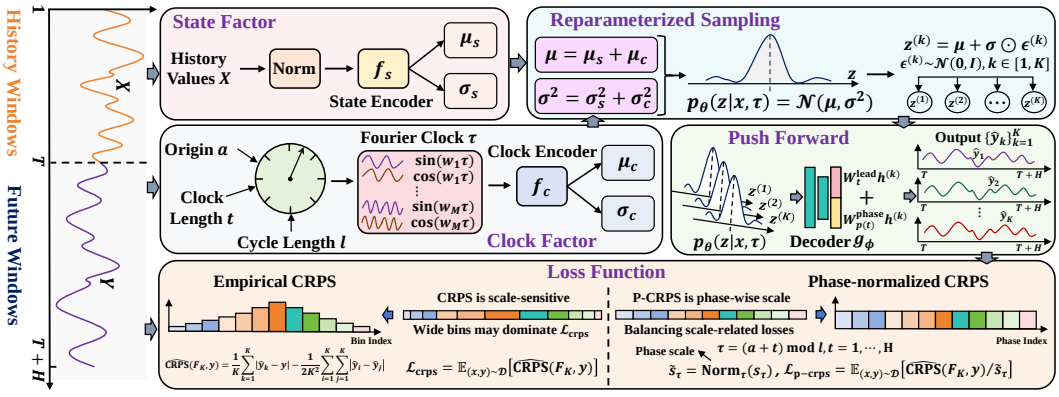


Figure 2: Overview of the proposed ClockUp framework.

normalization Kim et al. (2021) along the temporal dimension:

$$m_v = \frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_{i,v}, \quad s_v = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(\mathbf{x}_{i,v} - m_v)^2 + \epsilon}, \\ \bar{\mathbf{x}}_{:,v} = \frac{\mathbf{x}_{:,v} - m_v}{s_v}, \tag{6}$$

where $\epsilon > 0$ is a numerical stability constant, T is the lookback window length, and $v \in \{1, \ldots, V\}$ indexes the channel. A shared channel-wise MLP maps the normalized input $\bar{\mathbf{X}} \in \mathbb{R}^{V\times T}$ to $h \in \mathbb{R}^{V\times D}$, followed by separate moment heads. The sufficient statistics of the state prior are:

$$[\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s] = f_s(\bar{\mathbf{X}}), \quad \text{where } \boldsymbol{\mu}_s, \boldsymbol{\sigma}_s \in \mathbb{R}^{V\times D}. \tag{7}$$

Here, D denotes the latent dimension. The state prior is a diagonal multivariate Gaussian:

$$p_s(\mathbf{z}_s \mid \mathbf{X}) = \mathcal{N}(\mathbf{z}_s; \boldsymbol{\mu}_s, \mathrm{diag}(\boldsymbol{\sigma}_s^2)), \tag{8}$$

which captures history-dependent latent variability while remaining robust to sample magnitudes.

Clock Prior. The normalized lookback captures recent local dynamics, but it may not uniquely identify their position within a recurring temporal cycle, as illustrated in Figure 1. To address this ambiguity, ClockUp introduces a separate clock-conditioned component that parameterizes phase-dependent latent uncertainty from known cycle positions.

For regularly sampled time series with sampling interval e, we define the cycle length $l = P/e$, where P is the physical period (e.g., 24 hours for a daily cycle). Given a discrete time index t measured from a fixed temporal reference, its phase is $q(t) = t \bmod l$, taking values from 0 to $l-1$. Directly using integer phase indices does not reflect the circular relationship between cycle positions: phases 0 and $l-1$ are numerically distant but temporally adjacent. To encode this periodic structure, we represent each phase using harmonic Fourier features:

$$\mathcal{C}(q(t), l) = \left[\sin\left(\frac{2\pi m q(t)}{l}\right), \cos\left(\frac{2\pi m q(t)}{l}\right)\right]_{m=1}^{M}, \tag{9}$$

where M denotes the number of harmonics and the resulting feature vector has dimension $2M$. These features provide a periodic representation of cycle position at multiple harmonic frequencies.

For a historical window ending at time index a, the future phase representation comprises the Fourier features associated with the H forecasting steps:

$$\boldsymbol{\tau} = [\mathcal{C}(q(a+1), l), \cdots, \mathcal{C}(q(a+H), l)]. \tag{10}$$

A shared channel-wise MLP and two separate moment heads form the clock encoder f_c:

$$[\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c] = f_c(\boldsymbol{\tau}), \quad \text{where } \boldsymbol{\mu}_c, \boldsymbol{\sigma}_c \in \mathbb{R}^{V\times D}, \\ p_c(\mathbf{z}_c \mid \boldsymbol{\tau}) = \mathcal{N}(\mathbf{z}_c; \boldsymbol{\mu}_c, \mathrm{diag}(\boldsymbol{\sigma}_c^2)). \tag{11}$$

Since f_c depends on cycle position rather than observed values, it parameterizes a recurring latent distribution shared across windows with the same future phase sequence. This component complements the history-conditioned state component by providing an explicit pathway for phase-dependent latent means and variances. Unlike conditioning only the decoder on temporal covariates, ClockUp incorporates known cycle position directly into the latent prior, providing a structural inductive bias for phase-sensitive uncertainty modeling.

Analytical Composition. We combine the two components through an additive latent construction, $\mathbf{z} = \mathbf{z}_s + \mathbf{z}_c$, and define them to be conditionally independent given $\mathbf{X}$ and $\boldsymbol{\tau}$:

$$p(\mathbf{z}_s, \mathbf{z}_c \mid \mathbf{X}, \boldsymbol{\tau}) = p_s(\mathbf{z}_s \mid \mathbf{X}) p_c(\mathbf{z}_c \mid \boldsymbol{\tau}). \tag{12}$$

Under this construction, the sum of the two Gaussian components is also Gaussian, with mean and covariance given by the sums of the corresponding parameters:

$$p_\theta(\mathbf{z} \mid \mathbf{X}, \boldsymbol{\tau}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_s + \boldsymbol{\mu}_c, \mathrm{diag}(\boldsymbol{\sigma}_s^2 + \boldsymbol{\sigma}_c^2)). \tag{13}$$

This formulation combines history-dependent and phase-dependent latent variability without introducing cross-component covariance parameters, and it permits direct sampling from the composed prior. Conditional independence is imposed by construction given $\mathbf{X}$ and $\boldsymbol{\tau}$; it neither requires history and phase to be independent nor prevents the two components from encoding overlapping periodic information. The additive prior therefore provides a structured inductive bias, rather than an identifiable decomposition of predictive uncertainty into distinct underlying sources.

4.2 Additive Push-forward Decoder

The decoder maps samples from the composed latent prior to future trajectories via a shared nonlinear trunk and two additive output heads: a lead-time head and a phase-conditioned residual head.

Let $\boldsymbol{\mu} = \boldsymbol{\mu}_s + \boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}_s^2 + \boldsymbol{\sigma}_c^2$. We draw K independent latent samples using the reparameterization trick Kingma & Welling (2013):

$$\boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{z}^{(k)} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}^{(k)}, \quad k = 1, \cdots, K, \tag{14}$$

where $\boldsymbol{\sigma}$ is the elementwise square root of the composed variance. Each sample is combined with a linear embedding of the normalized history and passed through a shared nonlinear trunk:

$$\mathbf{h}^{(k)} = \mathrm{MLP}\big([\mathbf{z}^{(k)},\, E(\bar{\mathbf{X}})]\big), \tag{15}$$

where E denotes the history embedding operation and $[\cdot]$ denotes feature concatenation. The trunk receives phase information indirectly through the composed latent sample, without taking phase features as an additional explicit input.

Lead-time and Phase-conditioned Heads. We decode each hidden representation using two additive mappings, indexed by lead time and temporal phase. For channel v and forecast step t, the normalized prediction is:

$$\bar{y}_{v,t}^{(k)} = W_t^{\mathrm{lead}}\, \mathbf{h}_v^{(k)} + W_{q(t)}^{\mathrm{phase}} \mathbf{h}_v^{(k)}, \quad \text{where } q(t) = (a+t) \bmod l,\ t \in [1, H]. \tag{16}$$

The vector W_t^{lead} is the t-th row of a shared linear output head and depends only on the forecast lead time. In contrast, $W_{q(t)}^{\mathrm{phase}}$ is a phase-conditioned weight vector parameterized using the Fourier phase features introduced above. This residual head allows the mapping from latent representations to forecasts to vary with the known target-time phase. Since both heads act on the same stochastic hidden representation, the phase-conditioned residual can modify the predictive spread as well as the predictive mean, rather than merely adding a deterministic phase-dependent offset. Its output weights are initialized to zero, so decoding initially uses only the lead-time head while retaining the phase conditioning already present in the latent prior.

Predictive Samples. We restore each normalized trajectory to the original data scale using lookback statistics, $\hat{\mathbf{y}}_{v,t} = m_v + s_v \bar{\mathbf{y}}_{v,t}^{(k)}$. The future trajectories form the empirical predictive distribution:

$$\widehat{p}(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\tau}) = \frac{1}{K} \sum_{k=1}^{K} \delta_{\widehat{\mathbf{Y}}^{(k)}}(\cdot), \tag{17}$$

where $\delta_{\widehat{\mathbf{Y}}^{(k)}}(\cdot)$ denotes a Dirac point mass at the k-th sampled trajectory. This construction supports non-iterative generation, with all K trajectories evaluated in parallel. Together, the composed prior and the additive decoder incorporate temporal phase at two complementary stages: shaping latent variability and modulating its mapping to the forecasting horizon.

4.3 Optimization Objective

ClockUp is trained with empirical CRPS as the primary objective, using the estimator in equation 4, which directly encourages sharp and calibrated predictive distributions. However, CRPS is expressed in the same units as the target and scales linearly when the forecasting distribution and observation are rescaled together, so phases with larger characteristic target variation may contribute disproportionately to the aggregate objective, as illustrated in Figure 1. Therefore, we investigate phase-dependent scale normalization to reduce scale-driven differences in loss contribution, rather than to enforce identical losses across phases.

To mitigate this effect, we investigate a phase-normalized CRPS variant. For each temporal phase τ and channel c, we estimate a fixed scale factor from the training set:

$$s(\tau, c) = \mathrm{std}\big(y \mid \mathrm{phase} = \tau, \mathrm{channel} = c\big), \tag{18}$$

and normalize it within each channel to obtain a relative phase scale:

$$\tilde{s}(\tau, c) = \frac{s(\tau, c)}{\frac{1}{|\mathcal{P}|} \sum_{\tau'} s(\tau', c)}, \qquad \tilde{s}(\tau, c) \leftarrow \mathrm{clip}\big(\tilde{s}(\tau, c), \tfrac{1}{\kappa}, \kappa\big)\,, \tag{19}$$

where $|\mathcal{P}|$ is the set of phase bins and κ is a clipping threshold. The phase-normalized CRPS is then:

$$\mathcal{L}_{\mathrm{p-crps}} = \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\mathcal{D}}\left[\frac{\widehat{\mathrm{CRPS}}(F_K, \mathbf{Y})}{\tilde{s}(\tau, c)}\right], \tag{20}$$

where $\widehat{\mathrm{CRPS}}(F_K, \mathbf{Y})$ follows Eq. equation 4. Dividing by the relative scale $\tilde{s}(\tau, c)$ reweights wide and narrow phases *within* each channel, rather than handing the loss to low-scale channels. Phase weights are fixed functions of the conditioning context, strictly positive, and independent of the predictive distribution being evaluated. See Appendix A for theoretical analysis.

5 Experiments

In this section, we perform extensive experiments to compare the proposed ClockUp with 8 strong forecasting baselines across seven real-world datasets. All experiments are implemented using PyTorch Paszke et al. (2019) and executed on an NVIDIA A100.

5.1 Setup

Datasets. We evaluate on ETTh1, ETTh2, ETTm1, ETTm2, Electricity, Traffic, and Weather Wu et al. (2021); Nie et al. (2023); Liu et al. (2024), following the preprocessing of PPM Li et al. (2026). The lookback length is $T = 96$, and the forecasting horizon is $H = 192$. The details of statistical information about these datasets are provided in Table 4 of Appendix B.

Baselines. We compare ClockUp with several representative forecasting baselines, including DeepAR Salinas et al. (2020), TimeGrad Rasul et al. (2021), D3VAE Li et al. (2022), TimeDiff Shen & Kwok (2023), DiffusionTS Yuan & Qiao (2024), TMDM Li et al. (2024), NsDiff Ye et al. (2025), and PPM Li et al. (2026).

Metrics. To comprehensively evaluate forecasting performance, we adopt two sets of complementary metrics: CRPS Matheson & Winkler (1976) and QICE Han et al. (2022) for probabilistic forecasting, while mean squared error (MSE) and mean absolute error (MAE) for point forecasting. For all four metrics, lower values indicate superior performance.

Implementation. In our experiments, we train ClockUp with Adam Kingma & Ba (2014), using early stopping with a patience of 5 and a maximum of 30 epochs. For the phase-normalized CRPS, the scale $s(\tau, c)$ is the within-channel standard deviation of the training targets over the known cycle. It is fitted once on the training forecast windows and then frozen, with a clipping threshold of $\kappa = 4$. The number of samples in the training stage is set to 100, and 100 samples are used at inference to approximate the estimated distribution. More details can be found in Appendix B.

Table 1: Performance comparison on seven real-world datasets in terms of CRPS and QICE (%). The best and second-best results are highlighted in **bold** and underlined, respectively.

Dataset	ETTh1		ETTh2		ETTm1		ETTm2		Weather		Electricity		Traffic	
Method	CRPS	QICE	CRPS	QICE	CRPS	QICE	CRPS	QICE	CRPS	QICE	CRPS	QICE	CRPS	QICE
DeepAR	0.466	2.215	0.738	5.067	0.497	2.559	0.378	4.422	0.293	7.186	0.376	3.775	0.496	5.488
TimeGrad	0.636	5.031	0.990	7.921	1.001	8.097	0.644	2.744	0.362	2.954	0.405	5.788	0.694	6.546
TimeDiff	0.457	14.821	0.508	7.663	0.466	14.772	0.313	7.735	0.304	7.819	0.744	9.536	0.774	9.433
D3VAE	0.548	5.124	1.098	6.687	0.519	6.034	0.818	5.582	0.427	7.985	0.449	6.705	0.504	6.961
DiffusionTS	0.539	3.041	1.077	5.690	0.640	6.997	0.835	5.987	0.375	2.351	0.493	5.021	0.477	4.843
TMDM	0.468	2.354	0.382	1.955	0.364	2.024	0.322	1.843	0.240	3.674	0.462	10.995	0.562	10.873
NsDiff	0.417	1.564	0.349	2.125	0.336	**1.174**	0.270	2.343	0.253	5.134	0.286	7.595	0.367	8.366
PPM	0.337	1.481	0.306	2.011	0.314	1.782	0.237	1.580	0.215	2.684	0.206	2.435	0.252	2.744
ClockUp	**0.326**	**1.231**	**0.291**	**1.164**	**0.292**	1.332	**0.223**	**1.360**	**0.191**	**1.971**	**0.183**	**0.793**	**0.217**	**0.760**
ClockUp-V	**0.324**	**1.167**	**0.291**	**1.194**	**0.290**	**1.294**	**0.223**	**1.365**	**0.186**	**2.532**	**0.183**	**0.719**	**0.218**	**0.802**

5.2 Main Results

We evaluated ClockUp on seven real-world datasets for probabilistic MTS forecasting, comparing it against several competitive baselines. All experiments were repeated 5 times and averaged. We denote the model trained with the empirical CRPS loss as ClockUp, and the model trained with the phase-normalized CRPS loss as ClockUp-V. As shown in Tables 1 and 2, ClockUp consistently outperforms all eight external baselines across the evaluated benchmarks, with the largest gains

on Weather, Electricity, and Traffic. Its phase-normalized variant, ClockUp-V, further improves aggregate scoring performance on several datasets, revealing a trade-off with aggregate calibration.

Table 2: Performance comparison on seven real-world datasets in terms of MSE and MAE. The best and second-best results are highlighted in **bold** and underlined, respectively.

Dataset	ETTh1		ETTh2		ETTm1		ETTm2		Weather		Electricity		Traffic	
Method	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
DeepAR	0.797	0.624	1.800	0.992	0.870	0.697	0.535	0.508	0.259	0.333	0.304	0.377	0.968	0.524
TimeGrad	1.232	0.909	2.449	1.230	2.779	1.190	1.101	0.811	0.562	0.510	0.460	0.477	1.647	0.814
TimeDiff	0.493	0.473	0.544	0.527	0.539	0.480	0.286	0.352	0.266	0.316	0.836	0.758	1.422	0.788
D3VAE	0.906	0.707	3.392	1.520	0.815	0.637	2.106	1.140	0.445	0.449	0.520	0.528	1.097	0.577
DiffusionTS	0.917	0.714	2.941	1.318	1.036	0.755	1.500	0.918	0.452	0.490	0.594	0.585	0.989	0.560
TMDM	0.742	0.642	0.535	0.474	0.514	0.460	0.411	0.409	0.275	0.314	0.212	0.327	0.711	0.415
NsDiff	0.687	0.563	0.448	0.448	0.461	0.447	0.278	0.344	0.251	0.290	0.201	0.303	0.640	0.374
PPM	0.447	0.432	0.376	0.394	0.381	0.390	0.247	0.306	0.242	0.268	0.182	0.267	0.524	0.323
ClockUp	**0.439**	**0.426**	**0.363**	**0.384**	**0.358**	**0.375**	**0.231**	**0.290**	**0.211**	**0.242**	**0.154**	**0.244**	**0.464**	**0.280**
ClockUp-V	**0.442**	**0.425**	**0.363**	**0.384**	**0.359**	**0.373**	**0.230**	**0.290**	**0.207**	**0.239**	**0.154**	**0.244**	**0.471**	**0.281**

5.3 Ablation Studies

To investigate the contributions of the core components in ClockUp, we conduct ablation experiments on five datasets. Concretely, we consider the following variants: 1) **w/o f_s** removes the history-conditioned state prior; 2) **w/o τ** disables all clock-related conditioning; 3) **[IN(X), τ]** replaces the structured prior with simple concatenation of state and clock features; 4) **w/o σ_c** removes the phase-dependent clock variance; 5) **w/o $W^{\text{phase}}_{q(t)}$** removes the phase-conditioned output head; and 6) **r/w NLL+MM** replaces empirical CRPS with the objective used by PPM Li et al. (2026). As shown in Table 3, variant 1) reveals the complementarity between the clock and state priors, variant 6) highlights the advantage of the empirical CRPS loss in calibration, and variants 2)–5) collectively demonstrate that explicitly modeling phase-dependent uncertainty, rather than merely exposing the model to phase features, is the most critical driver of the observed gains.

Table 3: Ablation studies on five real-world datasets.

Dataset	**ClockUp**			w/o f_s			w/o τ			[IN($\mathbf{X}$),τ]			w/o σ_c			w/o $W^{\text{phase}}_{q(t)}$			r/w NLL+MM		
	CRPS	QICE	MSE	CRPS	QICE	MSE	CRPS	QICE	MSE	CRPS	QICE	MSE	CRPS	QICE	MSE	CRPS	QICE	MSE	CRPS	QICE	MSE
ETTh1	**0.326**	**1.231**	**0.439**	0.329	1.331	0.445	0.331	1.426	0.446	0.330	1.318	0.443	0.328	1.339	0.442	0.327	1.235	0.442	0.331	1.346	0.440
ETTm2	**0.223**	**1.360**	**0.231**	0.228	1.391	0.233	0.233	1.904	0.245	0.226	1.394	0.235	0.230	1.822	0.239	0.225	1.369	0.235	0.229	1.497	0.235
Weather	0.191	**1.971**	0.211	**0.189**	2.056	**0.206**	0.205	2.818	0.234	0.199	2.669	0.226	0.186	2.520	0.220	0.195	2.123	0.216	0.205	2.273	0.217
Electricity	**0.183**	**0.793**	**0.154**	0.186	0.828	0.156	0.200	1.109	0.185	0.186	0.913	0.158	0.194	0.869	0.168	0.188	0.805	0.164	0.195	2.084	0.154
Traffic	**0.217**	0.760	**0.464**	0.221	**0.710**	0.476	0.245	1.405	0.530	0.221	1.118	0.475	0.220	0.846	0.496	0.218	0.821	0.472	0.233	1.987	0.479

5.4 Visualization Analysis

To better illustrate the advantages of our method, we provide three visualizations covering phase-wise calibration, the learned structure of the latent priors, and representative forecasting trajectories. Figure 3 shows that while CRPS profiles are broadly similar across methods, the phase-wise QICE profiles reveal clearer differences. In particular, the concatenation-based variant attains a lower

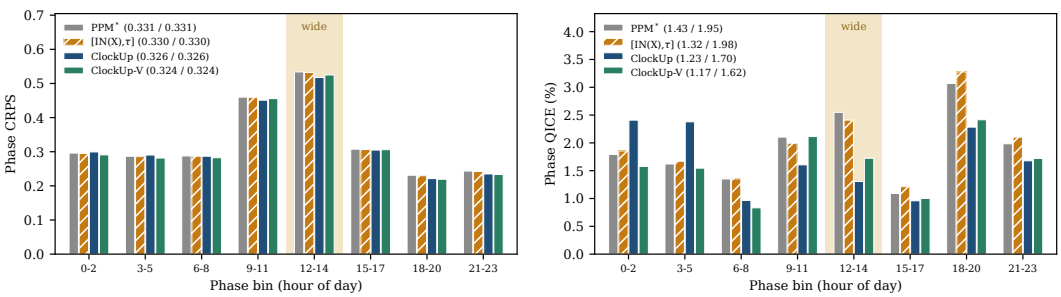


Figure 3: Phase-wise CRPS and QICE on ETTh1 using eight three-hour target-time bins. PPM* is reproduced with empirical CRPS loss. Legend values report (aggregate/mean phase-wise) CRPS and QICE. The shaded region marks the phase interval identified as having higher dispersion.

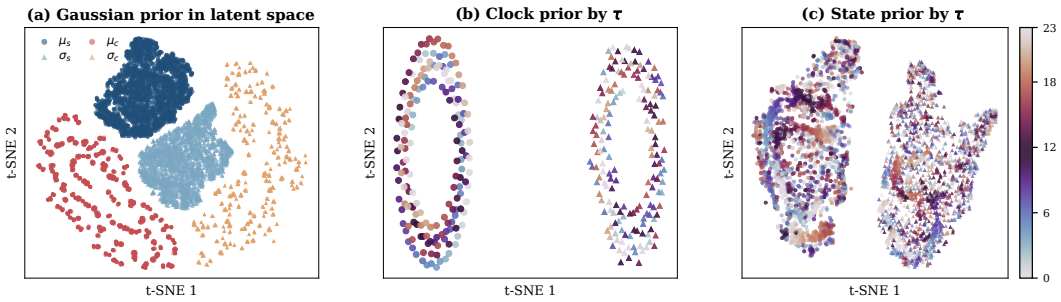


Figure 4: Visual semantic difference of state and clock priors in latent space.

global QICE than PPM* yet a slightly higher mean phase-wise QICE, whereas ClockUp improves both. This indicates that better aggregate calibration need not imply better phase-wise calibration, and supports reporting both aggregate and phase-wise diagnostics when evaluating phase-aware uncertainty models. Figure 4 visualizes the learned state and clock prior parameters using t-SNE Maaten & Hinton (2008). The clock parameters exhibit a more visibly phase-organized structure, whereas the state parameters show greater mixing across phase labels. The patterns are consistent with the intended conditioning roles of the two pathways: the state prior adapts to the observed history, while the clock prior models recurring variation through explicit conditioning on known cycle positions. Figure 5 compares ClockUp forecasts against PPM on two datasets. ClockUp tracks the trajectory more closely and yields better-calibrated intervals on both ETTm1 (CRPS $0.623 \rightarrow 0.433$; coverage $72\% \rightarrow 93\%$) and Traffic (CRPS $0.484 \rightarrow 0.308$; coverage $75\% \rightarrow 86\%$), reflecting improvements in both location and dispersion.

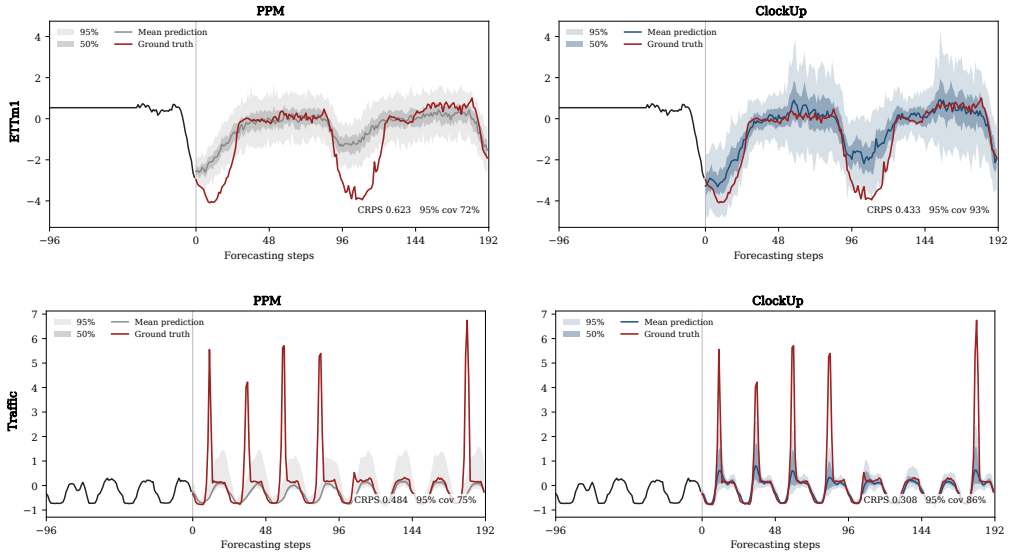


Figure 5: Probabilistic prediction comparisons on ETTm1 and Traffic datasets with PPM.

6 Conclusion

We propose ClockUp, a probabilistic forecasting framework that incorporates known cycle position into both an additive state–clock latent prior and a phase-conditioned decoder. This design complements history-based modeling with an explicit pathway for recurring phase-dependent variation, while retaining non-iterative probabilistic forecasting. Across seven benchmarks, ClockUp achieves consistent gains in CRPS and point forecasting accuracy over the evaluated baselines, with improved calibration on most datasets. Ablation studies and phase-wise diagnostics support the benefit of structured clock conditioning and further reveal dataset- and phase-dependent trade-offs, indicating that aggregate and phase-wise calibration capture complementary aspects of predictive quality. Phase-normalized training offers an additional mechanism for adjusting learning emphasis, though it does not uniformly improve all calibration metrics. Future work will explore adaptive and multiple periodicities, joint dependence modeling, and uncertainty estimation for rare extremes.

AI USE STATEMENT

AI-assisted tools were used only for language editing and manuscript polishing. The authors have reviewed and verified all resulting text and take full responsibility for the content of this paper.

ETHICS STATEMENT

This work focuses on probabilistic time series forecasting and does not involve human-subject experiments. The datasets used in our experiments were obtained and processed in accordance with their respective licenses and privacy requirements.

REPRODUCIBILITY STATEMENT

The experimental settings, data preprocessing procedures, model configurations, and evaluation metrics are described in the paper and its appendix. The code link is also provided in Appendix B.3.

REFERENCES

Pengsheng Cao, Piao Chen, and Der-Horng Lee. Embracing uncertainty: A review of bayesian methods for travel time reliability. *Reliability Engineering & System Safety*, pp. 112693, 2026.

Alireza Ermagun and David Levinson. Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews*, 38(6):786–814, 2018.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014.

Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with gaussian process priors for probabilistic time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 95435–95457, 2025.

Jinglin Li, Jun Tan, QI Fang, and Ning Gui. Parametric prior mapping framework for non-stationary probabilistic time series forecasting. *arXiv preprint arXiv:2605.23402*, 2026.

Qi Li, Zhenyu Zhang, Lei Yao, Zhaoxia Li, Tianyi Zhong, and Yong Zhang. Diffusion-based decoupled deterministic and uncertain framework for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 67475–67503, 2025.

Yan Li, Xinjiang Lu, Yaqing Wang, and Dejing Dou. Generative time series forecasting with diffusion, denoise, and disentanglement. *Advances in Neural Information Processing Systems*, 35: 23009–23022, 2022.

Yuxin Li, Wenchao Chen, Xinyue Hu, Bo Chen, Mingyuan Zhou, et al. Transformer-modulated diffusion models for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, volume 2024, pp. 18604–18622, 2024.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37:106315–106345, 2024.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, 2022.

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*, volume 2024, pp. 11116–11140, 2024.

Zhiding Liu, Mingyue Cheng, Zhi Li, Zhenya Huang, Qi Liu, Yanhu Xie, and Enhong Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. In *Advances in Neural Information Processing Systems*, 2023.

Jiaming Ma, Binwu Wang, Qihe Huang, Guanjun Wang, Pengkun Wang, Zhengyang Zhou, and Yang Wang. Mofo: Empowering long-term time series forecasting with periodic pattern modeling. *Advances in Neural Information Processing Systems*, 38:84871–84902, 2026.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

James E Matheson and Robert L Winkler. Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096, 1976.

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

Jinseong Park, Seungyun Lee, Woojin Jeong, Yujin Choi, and Jaewook Lee. Timebridge: Better diffusion prior design with bridge models for time series generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1135–1146, 2026.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, and et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. *Advances in Neural Information Processing Systems*, 31, 2018.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pp. 8857–8868, 2021.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction. In *International Conference on Machine Learning*, pp. 31016–31029, 2023.

Slawek Smyl, Boris N Oreshkin, Paweł Pełka, and Grzegorz Dudek. Any-quantile probabilistic forecasting of short-term electricity demand: Fusing uncertainties from diverse sources. *Information Fusion*, pp. 103637, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015.

Xian Teng, Sen Pei, and Yu-Ru Lin. Stocast: Stochastic disease forecasting with progression uncertainty. *IEEE Journal of Biomedical and Health Informatics*, 25(3):850–861, 2020.

Guillermo Terrén-Serrano, Ranjit Deshmukh, and Manel Martínez-Ramón. Probabilistic day-ahead forecasting of system-level renewable energy and electricity demand. *Nature Communications*, 17(1):3307, 2026.

Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*, 2017.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

Weiwei Ye, Songgaojun Deng, Qiaosha Zou, and Ning Gui. Frequency adaptive normalization for non-stationary time series forecasting. *Advances in Neural Information Processing Systems*, 37: 31350–31379, 2024.

Weiwei Ye, Zhuopeng Xu, and Ning Gui. Non-stationary diffusion for probabilistic time series forecasting. *arXiv preprint arXiv:2505.04278*, 2025.

Changtian Ying, Weiqing Wang, Jiong Yu, Qi Li, Donghua Yu, and Jianhua Liu. Deep learning for renewable energy forecasting: A taxonomy, and systematic literature review. *Journal of Cleaner Production*, 384:135414, 2023.

Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. In *International Conference on Learning Representations*, volume 2024, pp. 41582–41610, 2024.

Vincent Zhihao Zheng and Lijun Sun. Multivariate probabilistic time series forecasting with correlated errors. *Advances in Neural Information Processing Systems*, 37:54288–54329, 2024.

Zhihao Zheng, Seongjin Choi, and Lijun Sun. Better batch for deep probabilistic time series forecasting. In *International Conference on Artificial Intelligence and Statistics*, pp. 91–99, 2024.

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.

Yingbo Zhou, Yutong Ye, Shuhao Li, Rui Qian, Qiang Huang, Lemao Liu, Li Sun, and Dejing Dou. Pamod: Modeling cyclical shifts via phase-amplitude modulation for non-stationary time series forecasting. *arXiv preprint arXiv:2605.00466*, 2026.

A THEORETICAL PROPERTIES OF CLOCKUP

This section analyzes the statistical motivation, latent construction, decoder, and training objectives of ClockUp. We establish when phase information can improve the optimal marginal predictive distribution, characterize the additive Gaussian prior and the phase-conditioned decoder, and distinguish the propriety of population CRPS from the finite-sample behavior of empirical CRPS. We also formalize why aggregate calibration can conceal phase-specific errors. These results describe properties of the model construction and of the scoring objectives; they do not assume that neural-network optimization recovers a population optimum.

A.1 NOTATION AND ASSUMPTIONS

Let $\mathbf{X}$ denote the historical window and let $\overline{\mathbf{X}} = \mathrm{IN}(\mathbf{X})$ denote its instance-normalized representation. We write $\boldsymbol{\tau}$ for the known future phase sequence and b for a discrete target-time phase bin. The Fourier representation of $\boldsymbol{\tau}$ is deterministic and serves as the input to the clock encoder.

For the marginal scoring results, we consider a scalar target Y, corresponding to a fixed channel and forecast lead time; the results extend to an average over channels and lead times. Let C denote the forecasting context, let G_C be the true conditional CDF of Y given C, and let F_C be a candidate predictive CDF.

Unless stated otherwise, all distributions have finite first moments, all displayed expected scores are finite, and generated forecast samples are conditionally i.i.d. given C. Generated samples are also conditionally independent of the verification target given C. Decoder moment calculations additionally assume finite second moments. Parameters are treated as fixed when analyzing forecast distributions and Monte Carlo sampling.

A.2 THE PREDICTIVE VALUE OF PHASE INFORMATION

The motivation for explicit phase conditioning can be expressed as a comparison between two information sets. Let R be the history representation available to a phase-agnostic forecaster, such as the normalized lookback. Define

$$G_R(u) = \mathbb{P}(Y \le u \mid R), \qquad G_{R,\boldsymbol{\tau}}(u) = \mathbb{P}(Y \le u \mid R, \boldsymbol{\tau}). \tag{21}$$

Importantly, this comparison concerns the information retained in R; it does not assume that instance normalization always removes phase information.

Proposition A.1 (Bayes risk reduction from phase conditioning)**.** *Define the optimal population CRPS risks*

$$\mathcal{R}^*_R = \inf_{F_R} \mathbb{E}[\mathrm{CRPS}(F_R, Y)]\,, \tag{22}$$

$$\mathcal{R}^*_{R,\boldsymbol{\tau}} = \inf_{F_{R,\boldsymbol{\tau}}} \mathbb{E}[\mathrm{CRPS}(F_{R,\boldsymbol{\tau}}, Y)]\,, \tag{23}$$

where the infima range over all predictive CDFs measurable with respect to the indicated information. Then

$$\mathcal{R}^*_R - \mathcal{R}^*_{R,\boldsymbol{\tau}} = \mathbb{E}\int_{\mathbb{R}} \left(G_{R,\boldsymbol{\tau}}(u) - G_R(u)\right)^2 \, du \ge 0. \tag{24}$$

The reduction is strictly positive whenever the conditional CDFs differ on a set of positive measure under the product of the context distribution and Lebesgue measure on u.

Proof. For any context C and candidate CDF F_C, expansion of the CRPS integrand gives

$$\mathbb{E}\Big[\left(F_C(u) - \mathbf{1}_{Y \le u}\right)^2 \mid C\Big] = \left(F_C(u) - G_C(u)\right)^2 + G_C(u)\left(1 - G_C(u)\right). \tag{25}$$

Integrating over u yields

$$\mathbb{E}[\mathrm{CRPS}(F_C, Y) - \mathrm{CRPS}(G_C, Y) \mid C] = \int_{\mathbb{R}} \left(F_C(u) - G_C(u)\right)^2 \, du. \tag{26}$$

Thus the optimal forecasts are G_R and $G_{R,\boldsymbol{\tau}}$, respectively. Applying Eq. equation 26 with $C = (R, \boldsymbol{\tau})$ and candidate forecast $F_C = G_R$, and then averaging over the context, gives Eq. equation 24. □

Equation equation 24 characterizes phase blindness at the population level: ignoring phase incurs a nonzero cost when phase changes the conditional target distribution beyond the information contained in the retained history representation. If phase is already recoverable from R, or contributes no additional predictive information, this cost is zero.

This result motivates explicit phase conditioning but does not establish that a dedicated clock pathway is universally superior to feature concatenation. Both architectures may access the same information; their difference lies in parameterization, inductive bias, and finite-sample learning behavior, which we assess empirically.

A.3 Properties of the Additive State–Clock Prior

For a fixed context $C = (\mathbf{X}, \boldsymbol{\tau})$, flatten the latent coordinates into a vector of dimension d. ClockUp constructs

$$\mathbf{z}_s \mid C \sim \mathcal{N}\big(\boldsymbol{\mu}_s, \mathrm{diag}(\boldsymbol{\sigma}_s^2)\big), \quad \mathbf{z}_c \mid C \sim \mathcal{N}\big(\boldsymbol{\mu}_c, \mathrm{diag}(\boldsymbol{\sigma}_c^2)\big), \tag{27}$$

with $\mathbf{z}_s$ and $\mathbf{z}_c$ conditionally independent by construction. The state parameters depend on the normalized history, whereas the clock parameters depend on the Fourier phase representation.

Proposition A.2 (Exact Gaussian composition)**.** *The additive latent variable $\mathbf{z} = \mathbf{z}_s + \mathbf{z}_c$ satisfies*

$$\mathbf{z} \mid C \sim \mathcal{N}\big(\boldsymbol{\mu}_s + \boldsymbol{\mu}_c, \mathrm{diag}\left(\boldsymbol{\sigma}_s^2 + \boldsymbol{\sigma}_c^2\right)\big). \tag{28}$$

Consequently, sampling from the composed prior is distributionally equivalent to

$$\mathbf{z} = \boldsymbol{\mu}_s + \boldsymbol{\mu}_c + \sqrt{\boldsymbol{\sigma}_s^2 + \boldsymbol{\sigma}_c^2} \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I_d), \tag{29}$$

where the square root is elementwise.

Proof. Conditional independence implies that the characteristic function of the sum is the product of the component characteristic functions. Writing $\Sigma_s = \mathrm{diag}(\boldsymbol{\sigma}_s^2)$ and $\Sigma_c = \mathrm{diag}(\boldsymbol{\sigma}_c^2)$, for any $\mathbf{t} \in \mathbb{R}^d$,

$$\begin{aligned}\mathbb{E}\Big[e^{i\mathbf{t}^\top \mathbf{z}} \mid C\Big] &= \exp\big(i\mathbf{t}^\top \boldsymbol{\mu}_s - \tfrac{1}{2}\mathbf{t}^\top \Sigma_s \mathbf{t}\big) \exp\big(i\mathbf{t}^\top \boldsymbol{\mu}_c - \tfrac{1}{2}\mathbf{t}^\top \Sigma_c \mathbf{t}\big) \\ &= \exp\big(i\mathbf{t}^\top (\boldsymbol{\mu}_s + \boldsymbol{\mu}_c) - \tfrac{1}{2}\mathbf{t}^\top (\Sigma_s + \Sigma_c)\mathbf{t}\big). \end{aligned} \tag{30}$$

This is the characteristic function of the Gaussian distribution in Eq. equation 28. Eq. equation 29 has the same mean and covariance and is Gaussian, proving the sampling equivalence. □

The two components have different conditioning semantics by design. For a fixed phase sequence, the clock parameters are shared across forecasting windows regardless of their observed values. The state parameters, by contrast, adapt to normalized historical observations. Moreover, because the Fourier features repeat after one complete cycle, the clock parameters repeat when the forecasting origin is shifted by an integer multiple of the specified cycle length, with the horizon held fixed.

These properties do not require the representations to contain disjoint information. In particular, history may itself reveal phase. Conditional independence of latent draws also does not imply independence after averaging over contexts. When the relevant second moments exist, the law of total covariance gives

$$\mathrm{Cov}(\mathbf{z}_s, \mathbf{z}_c) = \mathrm{Cov}(\boldsymbol{\mu}_s(\mathbf{X}), \boldsymbol{\mu}_c(\boldsymbol{\tau})), \tag{31}$$

which need not vanish.

The composed prior depends on the sums of the means and variances. For example, for any constant vector $\mathbf{a}$, the alternative means

$$\boldsymbol{\mu}_s' = \boldsymbol{\mu}_s + \mathbf{a}, \qquad \boldsymbol{\mu}_c' = \boldsymbol{\mu}_c - \mathbf{a} \tag{32}$$

produce exactly the same composed prior. Similar reallocations of variance are possible whenever they preserve nonnegative component variances. Thus the construction specifies complementary conditioning pathways rather than a uniquely identifiable allocation of predictive uncertainty.

A.4 How the Phase-Conditioned Decoder Modulates Uncertainty

For a fixed channel v, let $\mathbf{h}_v$ be the stochastic hidden representation produced by the decoder trunk. Define

$$\mathbf{m}_{h,v} = \mathbb{E}[\mathbf{h}_v \mid C], \qquad \Sigma_{h,v} = \mathrm{Cov}(\mathbf{h}_v \mid C). \tag{33}$$

Writing the output-head parameters as column vectors, let

$$\mathbf{a}_t = \mathbf{w}_t^{\mathrm{lead}} + \mathbf{w}_{q(t)}^{\mathrm{phase}}. \tag{34}$$

The normalized forecast is $\overline{Y}_{v,t} = \mathbf{a}_t^\top \mathbf{h}_v$.

Proposition A.3 (Conditional decoder moments)**.** *The conditional mean and variance of the normalized forecast are*

$$\mathbb{E}[\overline{Y}_{v,t} \mid C] = \mathbf{a}_t^\top \mathbf{m}_{h,v}, \quad \mathrm{Var}(\overline{Y}_{v,t} \mid C) = \mathbf{a}_t^\top \Sigma_{h,v} \mathbf{a}_t. \tag{35}$$

For two lead times t and r,

$$\mathrm{Cov}(\overline{Y}_{v,t}, \overline{Y}_{v,r} \mid C) = \mathbf{a}_t^\top \Sigma_{h,v} \mathbf{a}_r. \tag{36}$$

Proof. The head weights are deterministic given C. The statements therefore follow from the expectation and covariance identities for linear transformations of a random vector. □

Expanding Eq. equation 35 gives

$$\mathrm{Var}(\overline{Y}_{v,t} \mid C) = (\mathbf{w}_t^{\mathrm{lead}})^\top \Sigma_{h,v} \mathbf{w}_t^{\mathrm{lead}} + 2(\mathbf{w}_t^{\mathrm{lead}})^\top \Sigma_{h,v} \mathbf{w}_{q(t)}^{\mathrm{phase}} + (\mathbf{w}_{q(t)}^{\mathrm{phase}})^\top \Sigma_{h,v} \mathbf{w}_{q(t)}^{\mathrm{phase}}. \tag{37}$$

Thus the phase-conditioned head can change predictive variance as well as predictive mean. Because of the cross term, its inclusion need not increase variance; it can increase or decrease the spread relative to the lead-time head acting on the same hidden distribution.

The same shared stochastic representation can induce dependence across forecast lead times through Eq. equation 36, despite the diagonal latent covariance. This statement does not by itself imply cross-channel dependence when channels are generated independently.

Finally, restoring the lookback level and scale gives

$$\mathbb{E}[\widehat{Y}_{v,t} \mid C] = m_v + s_v \mathbf{a}_t^\top \mathbf{m}_{h,v}, \quad \mathrm{Var}(\widehat{Y}_{v,t} \mid C) = s_v^2 \mathbf{a}_t^\top \Sigma_{h,v} \mathbf{a}_t. \tag{38}$$

The moment identities remain valid for a nonlinear trunk; no Gaussian assumption on $\mathbf{h}_v$ or the decoded forecast is required.

A.5 Scale Sensitivity and Phase-Weighted Population CRPS

For independent $Z, Z' \sim F$, the CRPS admits the representation

$$\mathrm{CRPS}(F, y) = \mathbb{E}|Z - y| - \frac{1}{2}\mathbb{E}|Z - Z'|. \tag{39}$$

Let $F^{a,b}$ denote the distribution of $a + bZ$, where $b > 0$. Eq. equation 39 immediately implies

$$\mathrm{CRPS}(F^{a,b}, a + by) = b\ \mathrm{CRPS}(F, y). \tag{40}$$

Therefore, when forecast–target pairs differ by an affine scale transformation, their CRPS values scale proportionally. This motivates scale adjustment, but does not imply that the target standard deviation alone determines forecast error for arbitrary phases.

For ClockUp-V, define the context-dependent weight

$$w(C) = \frac{1}{\widetilde{s}(b, v)}, \tag{41}$$

where $\widetilde{s}(b, v)$ is the clipped relative phase scale for phase bin b and channel v. The scales are estimated from training data and subsequently held fixed. We assume that relative scales are well-defined; a channel whose phase scales are all zero requires a positive scale floor or a specified fallback such as unit weights.

Proposition A.4 (Conditional propriety under fixed positive weights)**.** *Let $w(C) > 0$ be a fixed, measurable function of the conditioning context C. The same weight function is used for all candidate predictive distributions and does not depend on the distribution being evaluated. Then*

$$\mathbb{E}[w(C)\,\{\mathrm{CRPS}(F_C, Y) - \mathrm{CRPS}(G_C, Y)\}] = \mathbb{E}\left[w(C)\int_{\mathbb{R}} (F_C(u) - G_C(u))^2\; du\right]. \tag{42}$$

Consequently, the weighted population score is strictly proper for the conditional scalar distribution, up to contexts of probability zero.

Proof. Multiply Eq. equation 26 by $w(C)$ and apply the tower property of conditional expectation. The right-hand side is nonnegative. Since $w(C) > 0$ almost surely, equality requires equality of the conditional CDFs almost everywhere in u for almost every context; right-continuity then identifies the corresponding distributions. □

For clipping threshold $\kappa \geq 1$,

$$\frac{1}{\kappa} \leq \widetilde{s}(b, v) \leq \kappa, \qquad \frac{1}{\kappa} \leq w(C) \leq \kappa. \tag{43}$$

Let $\Delta(F)$ denote the unweighted population CRPS regret and let $\Delta_w(F)$ denote the weighted regret in Eq. equation 42. Then

$$\frac{1}{\kappa}\Delta(F) \leq \Delta_w(F) \leq \kappa\Delta(F). \tag{44}$$

Thus clipping prevents the weighted objective from assigning arbitrarily large or vanishing importance to a context.

Over all conditional distributions, weighting preserves the optimal forecast G_C. Over a restricted neural-network family, however, different contexts share parameters and cannot generally be fitted independently. Reweighting can therefore change the best attainable compromise across phases. Neither Proposition A.4 nor Eq. equation 44 guarantees an improvement in unweighted CRPS or QICE.

When the weights are estimated from training targets, Proposition A.4 describes the population scoring functional conditional on the fitted weights, evaluated on a new verification target. It is not a claim that reuse of the training data yields an unbiased generalization estimate.

A.6 Finite-Sample Properties of Empirical CRPS

The population scoring result must be distinguished from the objective computed using a finite forecast ensemble. For conditionally i.i.d. draws $\widehat{Y}_1, \ldots, \widehat{Y}_K \sim F_C$, define

$$F_{C,K} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\widehat{Y}_k}. \tag{45}$$

The empirical objective used by ClockUp is

$$\widehat{S}_K = \frac{1}{K}\sum_{k=1}^{K} |\widehat{Y}_k - Y| - \frac{1}{2K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} |\widehat{Y}_i - \widehat{Y}_j|. \tag{46}$$

This is exactly $\mathrm{CRPS}(F_{C,K}, Y)$.

Proposition A.5 (Finite-ensemble bias)**.** *Let*

$$D(F_C) = \mathbb{E}[|Z - Z'| \mid C], \qquad Z, Z' \overset{\text{i.i.d.}}{\sim} F_C. \tag{47}$$

Then, conditional on C and Y,

$$\mathbb{E}_{\mathrm{MC}}[\widehat{S}_K \mid C, Y] = \mathrm{CRPS}(F_C, Y) + \frac{D(F_C)}{2K}. \tag{48}$$

Proof. The expected first term of Eq. equation 46 is $\mathbb{E}[|Z - Y| \mid C, Y]$. The K diagonal pairwise terms vanish, while each of the $K(K-1)$ off-diagonal terms has conditional expectation $D(F_C)$. Hence

$$\begin{aligned}\mathbb{E}_{\mathrm{MC}}[\widehat{S}_K \mid C, Y] &= \mathbb{E}[|Z - Y| \mid C, Y] - \frac{K-1}{2K} D(F_C) \\ &= \mathrm{CRPS}(F_C, Y) + \frac{1}{2} D(F_C) - \frac{K-1}{2K} D(F_C) \\ &= \mathrm{CRPS}(F_C, Y) + \frac{D(F_C)}{2K},\end{aligned} \tag{49}$$

where the second equality substitutes Eq. equation 39. □

For a fixed underlying distribution, the bias is nonnegative and decreases as K^{-1}. Because $D(F_C)$ depends on the forecast distribution, it is not a constant offset during training. In expectation, the empirical objective adds a distribution-dependent dispersion penalty to the population CRPS risk. Consequently, strict propriety of population CRPS does not imply that the expected finite-ensemble objective is strictly proper as a functional of the underlying sampling distribution.

For example, when $K = 1$, the objective reduces to absolute error. If the true conditional distribution is uniform on $[0, 1]$, independently sampling from that distribution gives expected absolute error $1/3$, whereas the point forecast at $1/2$ gives expected absolute error $1/4$. Thus truthful sampling is not optimal for that finite-ensemble objective. This distinction does not invalidate Eq. equation 46: it remains the exact CRPS of the empirical forecast distribution actually submitted.

For $K \geq 2$, the off-diagonal estimator

$$\widehat{S}_K^{\mathrm{fair}} = \frac{1}{K} \sum_{k=1}^{K} |\widehat{Y}_k - Y| - \frac{1}{2K(K-1)} \sum_{i \neq j} |\widehat{Y}_i - \widehat{Y}_j| \tag{50}$$

satisfies

$$\mathbb{E}_{\mathrm{MC}}[\widehat{S}_K^{\mathrm{fair}} \mid C, Y] = \mathrm{CRPS}(F_C, Y). \tag{51}$$

The proof follows by counting the off-diagonal terms as above. This estimator is an unbiased alternative for scoring the underlying sampling distribution, rather than the exact CRPS of the realized empirical CDF. The reported ClockUp experiments use Eq. equation 46, not this alternative.

Multiplication by fixed phase weights gives

$$\mathbb{E}\Big[w(C)\widehat{S}_K\Big] = \mathbb{E}[w(C)\,\mathrm{CRPS}(F_C, Y)] + \frac{1}{2K}\mathbb{E}[w(C)D(F_C)]\,. \tag{52}$$

Thus phase weighting preserves conditional propriety at the population-CRPS level, but does not remove the finite-ensemble bias. For a fixed model with integrable $D(F_C)$, the additional term vanishes as K increases. This pointwise statement does not by itself establish convergence of trained parameters as the ensemble size changes.

A.7 Aggregate Calibration Can Hide Phase-Specific Errors

Let $U = F_C(Y)$ be the PIT for a continuous predictive CDF, and let B denote a phase bin with probability $\pi_b = \mathbb{P}(B = b)$. Partition $[0, 1]$ into S equal-length intervals $I_1, \ldots, I_S$, and define

$$r_{b,s} = \mathbb{P}(U \in I_s \mid B = b), \qquad r_s = \sum_b \pi_b r_{b,s}. \tag{53}$$

The population aggregate and phase-wise QICE values are

$$Q_{\mathrm{pool}} = \frac{1}{S} \sum_{s=1}^{S} \left| r_s - \frac{1}{S} \right|, \quad Q_b = \frac{1}{S} \sum_{s=1}^{S} \left| r_{b,s} - \frac{1}{S} \right|. \tag{54}$$

Proposition A.6 (Cancellation under pooling)**.** *The aggregate QICE satisfies*

$$Q_{\mathrm{pool}} \leq \sum_b \pi_b Q_b \leq \max_b Q_b. \tag{55}$$

The first inequality can be strict, including cases in which $Q_{\mathrm{pool}} = 0$ but every phase has nonzero QICE.

Proof. Since $\sum_b \pi_b = 1$, the triangle inequality gives

$$Q_{\text{pool}} = \frac{1}{S}\sum_{s=1}^{S}\left|\sum_b \pi_b\left(r_{b,s} - \frac{1}{S}\right)\right| \le \frac{1}{S}\sum_{s=1}^{S}\sum_b \pi_b\left|r_{b,s} - \frac{1}{S}\right| = \sum_b \pi_b Q_b. \tag{56}$$

The second inequality follows because a weighted average cannot exceed its largest component. □

As an example, consider two equally likely phases with conditional PIT densities

$$f_{U|B=0}(u) = 1 + \alpha\cos(2\pi u), \quad f_{U|B=1}(u) = 1 - \alpha\cos(2\pi u), \quad 0 < \alpha \le 1, \quad u \in [0,1]. \tag{57}$$

Both are valid densities, but neither is uniform, whereas their equally weighted mixture is exactly uniform. For $S = 4$, the first phase has interval probabilities

$$\left(\frac{1}{4} + \frac{\alpha}{2\pi}, \frac{1}{4} - \frac{\alpha}{2\pi}, \frac{1}{4} - \frac{\alpha}{2\pi}, \frac{1}{4} + \frac{\alpha}{2\pi}\right), \tag{58}$$

and the second phase has the opposite deviations. Therefore,

$$Q_{\text{pool}} = 0, \qquad Q_0 = Q_1 = \frac{\alpha}{2\pi} > 0. \tag{59}$$

The first phase places more PIT mass near the boundaries, while the second places more mass near the center. Their opposing discrepancies disappear after pooling.

The same pooling inequality holds for empirical occupancies when phase weights equal the observed phase frequencies and the same interval construction is used throughout. An unweighted mean over phase bins coincides with the frequency-weighted mean only when the bins have equal frequencies. Moreover, zero QICE at a fixed bin resolution does not establish full PIT uniformity, and phase-wise PIT uniformity does not establish calibration conditional on every aspect of the forecasting context.

A.8 Scope of the Marginal Scoring Guarantee

ClockUp produces forecast trajectories, but the reported CRPS objective is an average of scalar marginal scores:

$$S_{\text{marg}}(F, \mathbf{Y} \mid C) = \frac{1}{VH}\sum_{v=1}^{V}\sum_{t=1}^{H} w_{v,t}(C)\,\text{CRPS}(F_{v,t|C}, Y_{v,t}), \tag{60}$$

with $w_{v,t}(C) = 1$ for unweighted training and positive phase-dependent weights for ClockUp-V.

Applying Eq. equation 42 coordinatewise shows that the corresponding population regret is

$$\frac{1}{VH}\sum_{v=1}^{V}\sum_{t=1}^{H}\mathbb{E}\left[w_{v,t}(C)\int_{\mathbb{R}}\left(F_{v,t|C}(u) - G_{v,t|C}(u)\right)^2 du\right]. \tag{61}$$

Thus the score identifies the collection of conditional marginals, but not their joint dependence structure.

For example, the bivariate distributions of (Z, Z) and $(Z, -Z)$ with $Z \sim \mathcal{N}(0,1)$ have identical marginal distributions and therefore identical marginal CRPS scores for every observation vector, despite their opposite dependence structures. Accordingly, the theoretical scoring guarantees support marginal distributional accuracy. Correct temporal or cross-channel dependence requires additional evaluation beyond marginal CRPS and QICE.

Taken together, these results justify phase information as potentially useful predictive context, establish the exactness of the additive prior, and explain how the phase-conditioned decoder can modulate predictive location and dispersion. They also delimit the claims supported by the training objective: phase-weighted population CRPS retains conditional propriety, finite ensembles introduce a quantifiable sampling bias, and aggregate marginal diagnostics cannot fully characterize phase-wise calibration or joint predictive quality.

B Implementation Details

B.1 Dataset Details

We use seven well-known real-world time series datasets to evaluate the performance of our proposed ClockUp. The corresponding information is described below:

- **ETT** [1] consists of power load and oil temperature measurements from electricity transformers with seven features. The dataset includes four subsets, i.e., ETTh1, ETTh2, ETTm1, and ETTm2, which differ in sampling frequency.
- **Weather** [2] comprises 21 meteorological indicators recorded every 10 minutes in 2020 at the weather station of the Max Planck Institute for Biogeochemistry.
- **Electricity** [3] records hourly electricity consumption for 321 clients between 2012 and 2014.
- **Traffic** [4] contains hourly occupancy rates of different lanes, measured by various sensors on San Francisco highways and published by the California Department of Transportation.

Following , we set the sliding stride of the sampling window equal to the prediction length for high-dimensional multivariate datasets such as Weather, Electricity, and Traffic. This reduces the number of evaluated windows and the memory footprint while still covering the entire test horizon, so that all test data are used for inference and evaluation. All baselines use this setting during evaluation. Table 4 provides basic statistical information for these datasets. Based on the sampling frequency of each dataset, we can infer its maximum cycle length, which is consistent with Lin et al. (2024; 2025). We also report the average variance over the Fourier spectrum to examine the evolving dynamics Ye et al. (2024); Li et al. (2026), where a larger variance indicates more complex data dynamics.

Table 4: The basic statistical information on seven datasets.

Datasets	Channels	Frequency	Size	Cyclic Patterns	Dtaset Split	Variance
ETTh1	7	1 Hour	14307	Daily	(6:2:2)	3.690
ETTh2	7	1 Hour	14307	Daily	(6:2:2)	1.013
ETTm1	7	15 Minutes	69680	Daily	(6:2:2)	3.330
ETTm2	7	15 Minutes	69680	Daily	(6:2:2)	1.648
Electricity	321	1 Hour	26304	Daily & Weekly	(7:1:2)	2.645
Traffic	862	1 Hour	17544	Daily & Weekly	(7:1:2)	14.225
Weather	21	10 Minutes	52696	Daily	(7:1:2)	0.387

B.2 Baseline Information

We compare against several strong probabilistic forecasting baselines, including DeepAR Salinas et al. (2020), TimeGrad Rasul et al. (2021), D3VAE Li et al. (2022), TimeDiff Shen & Kwok (2023), DiffusionTS Yuan & Qiao (2024), TMDM Li et al. (2024), NsDiff Ye et al. (2025), and PPM Li et al. (2026). We briefly describe each baseline below:

- **DeepAR** adopts an autoregressive recurrent network to output the parameters of a parametric distribution at each step, yielding step-wise probabilistic forecasts.
- **TimeGrad** couples an autoregressive recurrence with a diffusion probabilistic model to characterize high-dimensional predictive distributions over time steps.
- **D3VAE** builds on a bidirectional variational autoencoder with component separation, and uses the resulting representation for data augmentation prior to diffusion generation.
- **TimeDiff** introduces future mixup together with autoregressive initialization, allowing a non-autoregressive diffusion model to produce forecasts.

[1] https://github.com/zhouhaoyi/ETDataset
[2] https://www.bgc-jena.mpg.de/wetter/
[3] https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014
[4] http://pems.dot.ca.gov

- **DiffusionTs** employs an encoder–decoder Transformer with disentangled temporal representations, reconstructing samples directly at each diffusion step and adding a Fourier-based loss to improve generation quality.
- **TMDM** injects conditional information into the forward diffusion process and uses a Transformer backbone to better estimate uncertainty distributions over a sequence.
- **NsDiff** introduces an explicit variance estimation module that parameterizes prior uncertainty, supervised by variances computed over sliding-window segments.
- **PPM** applies a parametric estimator to obtain a dynamic, adaptive prior, which guides learning of a complex predictive distribution through a learnable mapping.

B.3 Hyperparameter Details

We provide the hyperparameters of ClockUp in Table 5. The sensitivity analysis for the cycle length and the number of harmonics can be found in Appendix C.6. Code is available at `https://anonymous.4open.science/r/ClockUp-60B9/`

Table 5: Hyperparameters of ClockUp.

Datasets	Batch Size	Learning Rate	Hidden Dim	Cycle Length	Harmonics
ETTh1	7	1e-4	256	24	8
ETTh2	7	1e-4	256	24	8
ETTm1	7	1e-4	256	96	16
ETTm2	7	1e-4	256	96	16
Electricity	321	5e-4	256	168	24
Traffic	862	1e-3	256	168	24
Weather	21	1e-4	512	144	16

C Extra Results

C.1 Clock Parameterization

To evaluate the effectiveness of Fourier-based clock parameterization, we compare it with two alternatives: Query, adapted from CycleNet Lin et al. (2024), and Embedding, adapted from PAMod Zhou et al. (2026). As shown in Table 6, Fourier-based clock parameterization achieves the best results across all six datasets and all three metrics, yielding modest but consistent improvements. A plausible explanation is its continuous, periodic inductive bias: unlike discrete phase-indexed representations, Fourier features explicitly connect neighboring phases and preserve continuity across cycle boundaries. This structure may facilitate learning phase-dependent variation without requiring these relationships to be learned entirely from data.

Table 6: Performance comparison of different clock parameterizations in ClockUp.

Clock	ETTh1			ETTh2			ETTm1			ETTm2			Electricity			Traffic		
	CRPS	QICE	MSE	CRPS	QICE	MSE	CRPS	QICE	MSE	CRPS	QICE	MSE	CRPS	QICE	MSE	CRPS	QICE	MSE
Query	0.328	1.458	0.442	0.295	1.223	0.366	0.295	1.355	0.365	0.227	1.521	0.237	0.188	0.815	0.159	0.221	0.837	0.478
Embedding	0.330	1.301	0.444	0.298	1.278	0.369	0.301	1.438	0.378	0.228	1.714	0.239	0.195	0.950	0.174	0.231	1.716	0.488
Fourier	**0.326**	**1.231**	**0.439**	**0.291**	**1.164**	**0.363**	**0.292**	**1.332**	**0.358**	**0.223**	**1.360**	**0.231**	**0.183**	**0.793**	**0.154**	**0.217**	**0.760**	**0.464**

C.2 Sensitivity Analysis

Cycle Length and Harmonic Order. Figure 6 examines how clock configuration affects predictive accuracy and calibration. On ETTh1 and ETTm2, cycle lengths matching the daily period and some of their multiples generally yield favorable results, whereas mismatched lengths can degrade performance, particularly on ETTm2. On ECL, the weekly cycle achieves lower MSE than shorter cycles, though the best QICE occurs under a different configuration. These results suggest that cycle length affects both prediction aspects, and that their optimal settings need not coincide. In contrast, varying the number of harmonics causes relatively small MSE changes across datasets, while calibration

remains more sensitive, especially on ECL. Overall, the results support using a meaningful temporal cycle to configure ClockUp and show that competitive accuracy does not require fine-tuning the harmonic count.

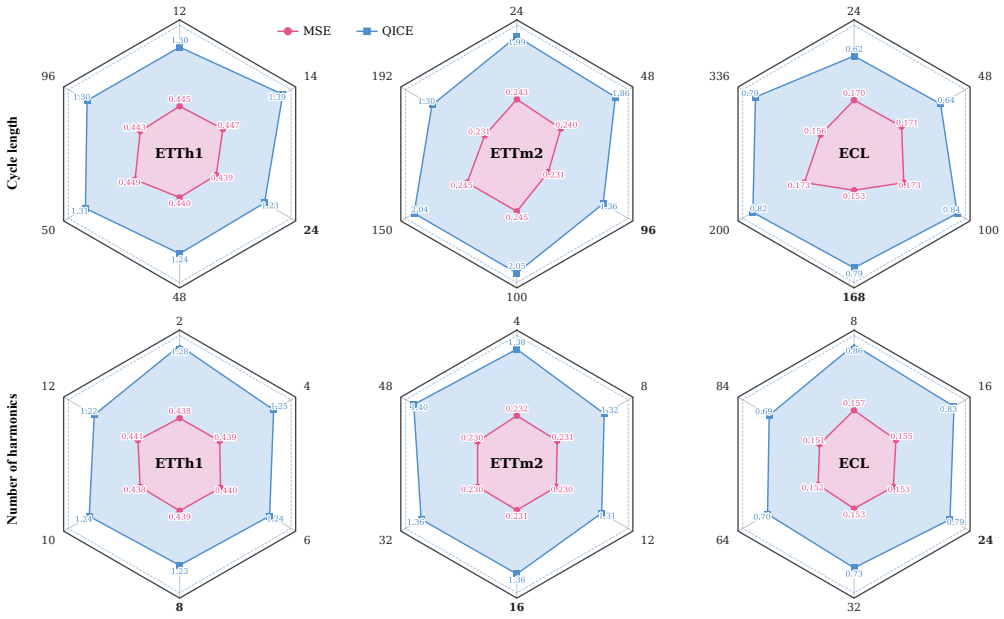


Figure 6: Analysis of cycle length and harmonic order.

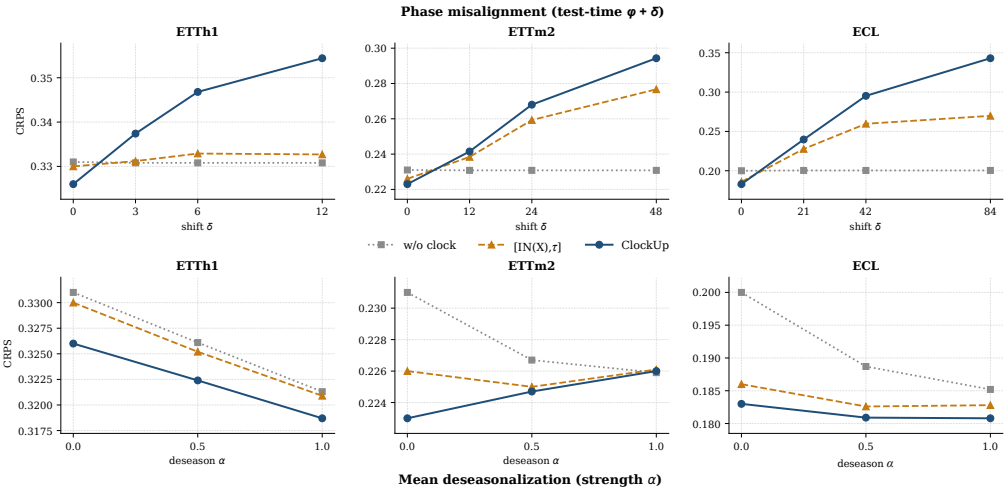


Figure 7: Analysis of phase misalignment and weakened mean seasonality.

Phase Misalignment and Weakened Seasonality. Figure 7 examines whether ClockUp benefits from accurate phase information and whether such benefits persist under weakened mean seasonality. With correctly aligned phases ϕ, ClockUp achieves the lowest CRPS among the three variants on all datasets. Increasing the test-time phase offset δ consistently degrades its performance, whereas the variant without the clock pathway remains unaffected. Direct phase concatenation shows smaller degradation but also yields less favorable results under correct alignment, supporting the usefulness of structured phase conditioning while revealing its reliance on accurate phase information. To weaken mean seasonality, we subtract the training-set-estimated phase-wise mean ($\mathbf{i}\mathbf{X} \leftarrow \mathbf{X} - \alpha\boldsymbol{\mu}(\boldsymbol{\phi})$) at deseasonalization strength $\alpha \in \{0, 0.5, 1\}$ and retrain each variant. ClockUp retains its advantage on ETTh1 and ECL, while its advantage on ETTm2 diminishes until the meth-

ods perform comparably. These benefits are therefore not solely attributable to the removed phase-dependent mean component, though their magnitude is dataset-dependent.

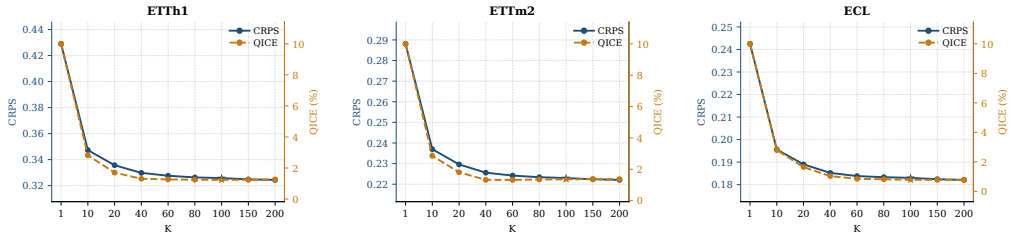


Figure 8: Analysis of sampling numbers K.

Sampling Number K. Figure 8 evaluates the effect of the number of predictive samples K. Across all three datasets, CRPS and QICE decrease rapidly when K increases from very small values, followed by diminishing improvements as the sample count grows. The curves become comparatively stable around $K = 100$, with limited additional gains up to $K = 200$. This suggests that ClockUp can obtain a sufficiently resolved empirical predictive distribution with a moderate sampling budget, supporting $K = 100$ as a practical accuracy–cost trade-off to maintain reliable performance.

C.3 Extra Visualizations

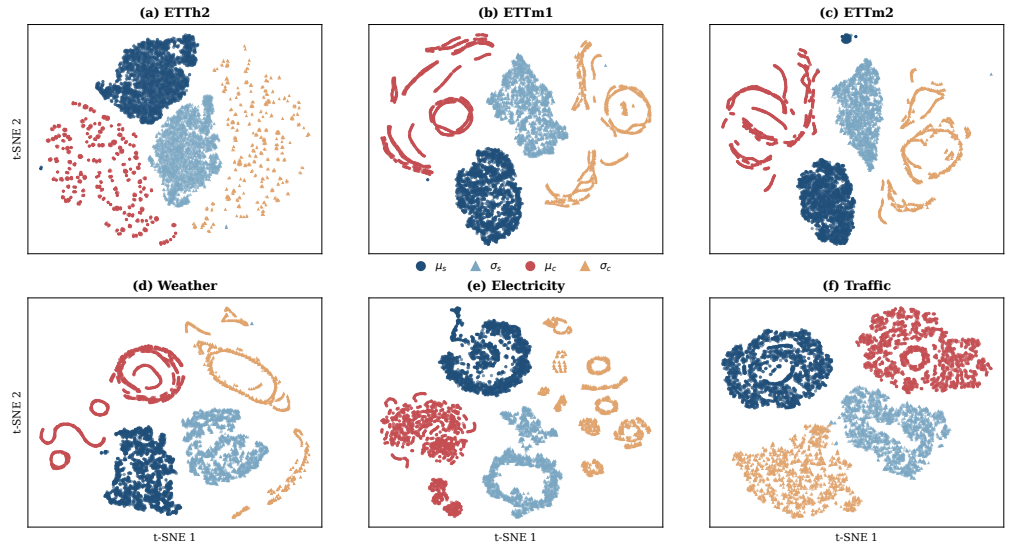


Figure 9: t-SNE visualizations of the learned semantics on six datasets.

State-Clock representations. Figure 9 visualizes the representations associated with the state-conditioned and clock-conditioned location and scale components. Across the six datasets, the four groups generally occupy distinct regions in the two-dimensional projections, suggesting that the corresponding pathways learn differentiated representations rather than collapsing to a common representation. On several datasets, the clock-conditioned components exhibit curved or loop-like structures, which are qualitatively consistent with their periodic parameterization. The variation in these patterns across datasets also suggests that the learned representations are not restricted to a single common geometry. These visualizations provide qualitative support for ClockUp's separation of state- and phase-dependent information.

Efficiency Analysis. As shown in Figure 10, ClockUp occupies the lower-left region of the plot, achieving the lowest CRPS and QICE among the compared methods while remaining within the low-latency group. Relative to PPM, it improves both metrics while retaining a similar order of inference cost; compared with the evaluated diffusion-based methods, it combines better predictive scores with substantially lower inference time. These results demonstrate a favorable qual-

ity–efficiency trade-off in the evaluated setting, showing that ClockUp's structured conditioning can improve probabilistic prediction without requiring the high inference cost observed for several competing methods.

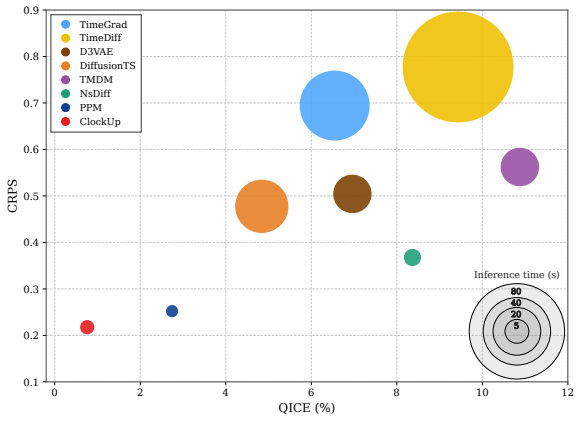


Figure 10: Inference time comparison across models on Traffic, with history window $T = 96$ and forecasting horizon $H = 192$. The number of samples is set to 100.

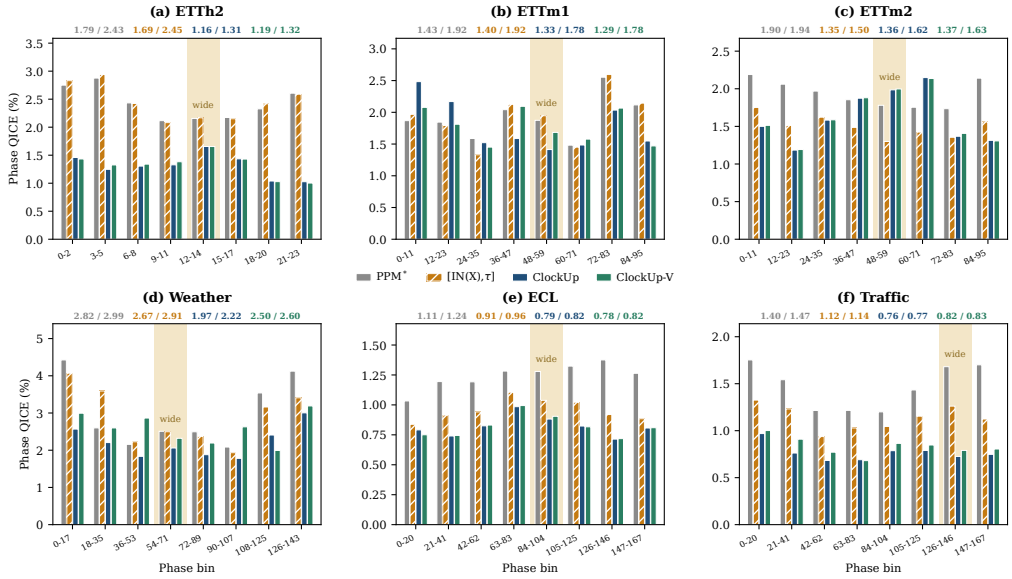
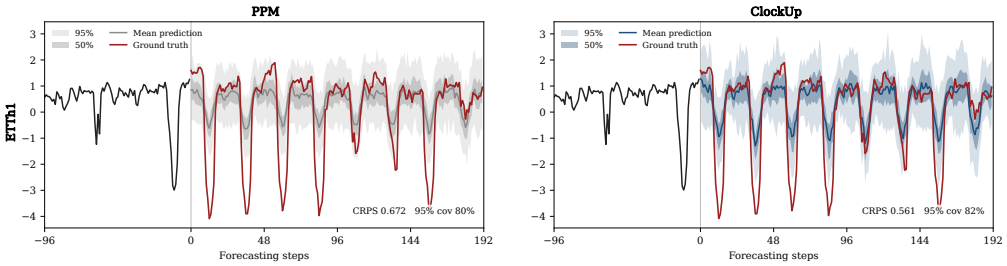


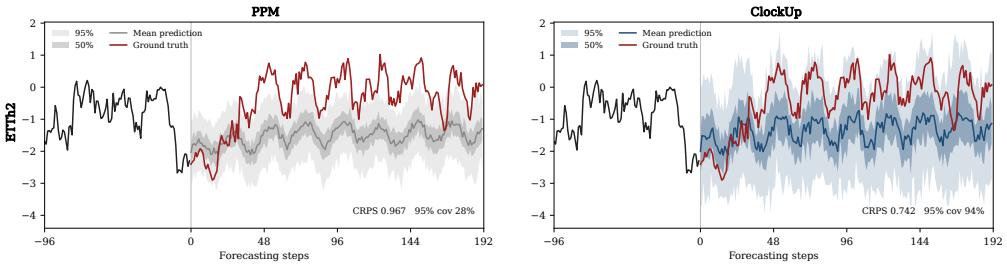
Figure 11: Phase-wise QICE on six datasets using eight target-time bins. PPM* is reproduced with empirical CRPS loss. Legend values report (aggregate/mean phase-wise) QICE, and the shaded region marks the phase interval identified as having higher dispersion.

Phase-wise Calibration. Figure 11 examines phase-wise calibration. ClockUp achieves lower mean phase-wise QICE than PPM* on all six datasets and than direct phase concatenation on five, with ETTm2 as the exception (e.g., from 2.45 to 1.31 on ETTh2 and 1.14 to 0.77 on Traffic), supporting the value of structured phase conditioning beyond merely exposing phase features. Moreover, improvements are nonetheless non-uniform: some higher-dispersion bins remain challenging, and ClockUp-V behaves dataset-dependently, matching ClockUp on several datasets while ClockUp attains lower calibration errors on Weather and Traffic. Overall, this observation indicates improved phase-resolved calibration in most settings, though not at every phase.

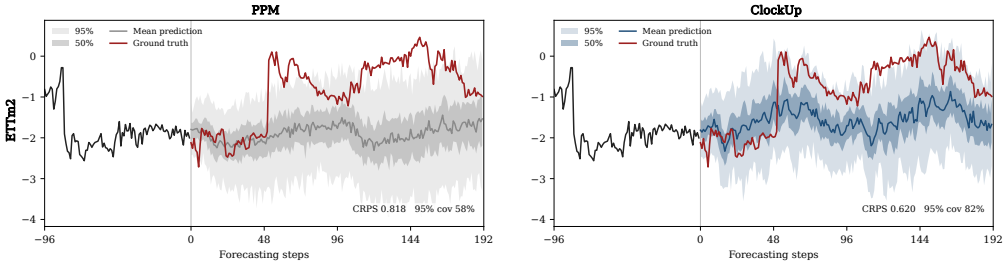
Qualitative Comparison of Predictions. To show the superiority of our ClockUp, we visualize more prediction results across five datasets compared with PPM in Figure 12.



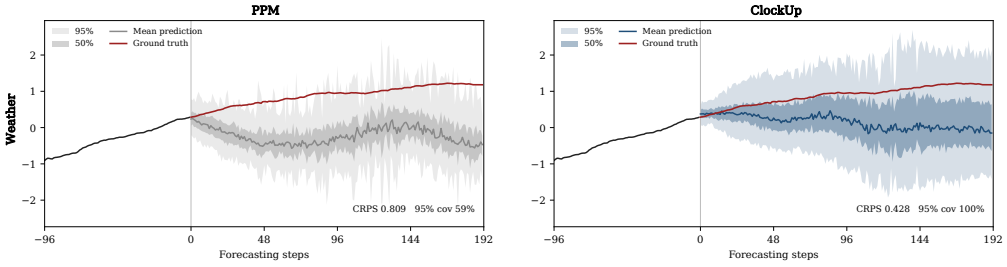
(a) ETTh1



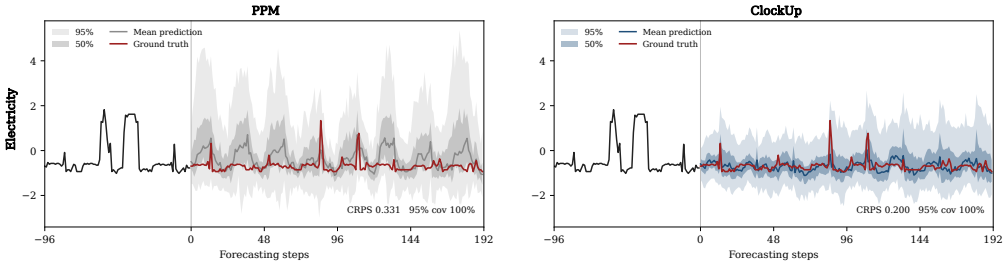
(b) ETTh2



(c) ETTm2



(d) Weather



(e) Electricity

Figure 12: Probabilistic prediction comparisons across five datasets with PPM.

D DISCUSSION

Limitations and Future Works. ClockUp relies on a predefined cycle and reusable phase-dependent structure. While the deseasonalization experiments support its applicability under weakened mean seasonality, its clock pathway may offer limited benefit for non-periodic series or when phase-dependent patterns are not stable over time. Furthermore, the additive latent prior imposes a structural assumption rather than an identifiable decomposition of uncertainty, and the current marginal forecasting objective does not inherently ensure accurate joint dependence across variables or forecast horizons. Future work could explore adaptive clock conditioning when phase information is weak or unreliable, more flexible state–phase interactions, and extensions to joint predictive distribution modeling.