# PREFIXFLOW: PROJECTIVE CONSISTENCY FOR PROBABILISTIC TIME-SERIES FORECASTING

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

## ABSTRACT

Probabilistic forecasts at different horizons share a common future prefix, yet direct generative forecasters need not yield consistent predictive distributions over it. We formalize this requirement as *horizon-wise projective consistency*. Using Energy Distance, we show that three recent generative forecasters exhibit systematic cross-horizon discrepancies beyond finite-sample Monte Carlo variation. To address this, we propose **PrefixFlow**, a variable-horizon Flow Matching model trained jointly across horizons and supporting any patch-aligned horizon up to a prescribed maximum. Unlike autoregressive forecasters that require sequential rollout, PrefixFlow generates the entire requested future in parallel. Its prefix-invariant conditioning and prefix-causal velocity field preserve the generative dynamics of an existing prefix under horizon extension; together with a projectively consistent Gaussian source, this yields projectively consistent predictive distributions. Across six benchmarks, PrefixFlow achieves the lowest average CRPS on five, with no detectable cross-horizon discrepancy beyond finite Monte Carlo variation.

## 1 INTRODUCTION

Probabilistic time-series forecasting models the conditional distribution of future sequences given an observed history. Recent work has increasingly adopted diffusion- and flow-based generative models for multi-step probabilistic forecasting (Rasul et al., 2021; Shen & Kwok, 2023; Li et al., 2025; Ye et al., 2025; Kollovieh et al., 2025; Solano Carrillo et al., 2026). In many forecasting settings, the same observed history may be queried at multiple forecast horizons. A variable-horizon probabilistic forecaster should induce a compatible family of predictive distributions across horizons. Since these forecasts share future prefixes, their distributions should agree on those prefixes under marginalization (Kallenberg, 2021).

For an observed history $\mathbf{x}$, let $Q_H(\cdot \mid \mathbf{x})$ denote the predictive distribution over an $H$-step future. For $h < H$, let $\pi_h(\mathbf{y}_{1:H}) = \mathbf{y}_{1:h}$ denote the projection onto the first $h$ future steps. We say that the predictive distributions across horizons satisfy *horizon-wise projective consistency* if, for every $h < H$,

$$Q_h(\cdot \mid \mathbf{x}) = (\pi_h)_{\#} Q_H(\cdot \mid \mathbf{x}). \tag{1}$$

It requires the predictive law on a shared future prefix to remain unchanged when the forecast is extended beyond that prefix.

Autoregressive models with a shared horizon-independent factorization are projectively consistent by construction, since all horizons share the same chain of one-step conditionals. Their generation, however, requires sequential rollout. Direct multi-step generative models avoid this sequential dependency, but predictive distributions obtained at different horizons need not be projectively consistent. Figure 1 shows an example for D3U (Li et al., 2025) on Weather: given the same history and shared 96-step future prefix, Direct-h and Crop-from-H assign visibly different predictive distributions, with the probability of the same directional event changing from $0.88$ to $0.13$.

We find that this behavior is systematic across three direct generative forecasters, three datasets, and three horizon pairs. In all 27 settings, the cross-horizon Energy Distance exceeds the within-distribution Monte Carlo baseline on every evaluated window. More importantly, this inconsistency can alter probability-based decisions. Across the evaluated settings, the Directional Event Probability Gap reaches $0.545$, and the Direction Flip Rate reaches $57.9\%$. Thus, the same observed history can lead to different decisions simply because the requested forecast horizon changes.



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.

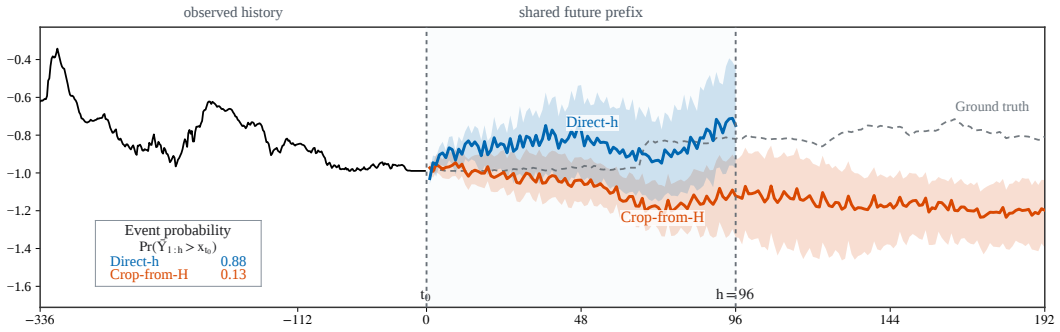


Figure 1: An example of cross-horizon inconsistency for D3U on Weather. Given the same 336-step history, Direct-$h$ ($h = 96$) and Crop-from-$H$ ($H = 192$, cropped to the first 96 steps) assign markedly different predictive distributions to the shared future prefix. The probability of the directional event $\Pr(\bar{Y}_{1:h} > x_{t_0})$ changes from 0.88 to 0.13. Colored lines denote predictive medians, and shaded regions denote the 10–90% predictive quantile intervals.

An alternative is to always generate the maximum horizon and crop it to the requested length, which guarantees projective consistency but incurs redundant computation for shorter requests. We instead seek a non-autoregressive model that directly generates only the requested horizon while preserving projective consistency. To this end, we introduce **PrefixFlow**, a variable-horizon Flow Matching model built on prefix-invariant conditioning and a prefix-causal velocity field. Together with a projectively consistent Gaussian source, these properties yield a projectively consistent predictive family. Our contributions are threefold:

- We formalize *horizon-wise projective consistency* as a requirement for probabilistic forecasting. A systematic diagnosis of direct generative models with fixed forecast horizons quantifies their cross-horizon inconsistencies and reveals the effects on probability-based decisions.
- We introduce **PrefixFlow**, a unified variable-horizon Flow Matching model that shares parameters across horizons and directly supports any patch-aligned horizon up to a prescribed maximum, generating the requested future in parallel.
- We prove that **PrefixFlow** yields projectively consistent predictive distributions. Across six forecasting benchmarks, it achieves strong point and probabilistic forecasting performance with no detectable cross-horizon discrepancy beyond finite Monte Carlo variation.

## 2 Related Work

**Autoregressive and direct probabilistic forecasting.** Autoregressive forecasters such as DeepAR and TimeGrad define the future through a shared sequence of one-step conditional distributions (Salinas et al., 2020; Rasul et al., 2021). Extending the horizon simply continues this factorization, naturally preserving the predictive law on an existing prefix at the cost of sequential generation. TEDM likewise adopts an autoregressive diffusion formulation that aligns diffusion time with physical time (Solano Carrillo et al., 2026). In contrast, TimeDiff, D3U, and NsDiff develop direct diffusion-based forecasting, while TSFlow uses Flow Matching (Shen & Kwok, 2023; Li et al., 2025; Ye et al., 2025; Kollovieh et al., 2025). For direct generative forecasters trained independently at each horizon, no structural inductive bias enforces agreement on overlapping prefixes. The resulting horizon-specific predictive distributions can therefore disagree on the same future prefix.

**Forecasting across varying horizons.** Supporting varying forecast horizons is an important capability of general-purpose time-series forecasting models. Time-series foundation models such as TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024), and Moirai (Woo et al., 2024) provide such flexibility within a shared pretrained model through different generation mechanisms. ElasTST considers non-autoregressive varied-horizon forecasting for deterministic prediction, with existing forecasts preserved as the inference horizon is extended (Zhang et al., 2024). For probabilistic forecasting, however, supporting multiple horizons is not enough: the resulting predictive distributions should also agree on their shared future prefixes.

**Projective consistency.** At the distributional level, the horizon-invariance question is naturally expressed through projective consistency, closely related to the Kolmogorov consistency conditions for finite-dimensional distributions of a stochastic process (Kallenberg, 2021). Related consistency requirements have also been studied in neural stochastic-process models, such as Markov Neural Processes (Xu et al., 2023). More directly related to probabilistic forecasting, MOSES studies marginalization-consistent forecasting for irregular time series over arbitrary target subsets (Yalavarthi et al., 2026). Its query-wise separable flows provide general subset consistency but constrain dependency modeling in the transformations. We instead leverage the nested-prefix structure of forecast horizons, using prefix-closed generative dynamics to preserve cross-horizon consistency while allowing causal interactions within the generated prefix.

# 3 Problem Formulation and Empirical Diagnosis

## 3.1 Horizon-wise Projective Consistency

Let $\mathbf{x} \in \mathbb{R}^{L \times D}$ denote an observed history, and let $\mathbf{y}_{1:H} \in \mathbb{R}^{H \times D}$ denote an $H$-step future sequence. For each forecast horizon $H \in \mathcal{H}$, let $Q_H(\cdot \mid \mathbf{x})$ be the predictive distribution over $\mathbf{y}_{1:H}$. Projective consistency requires Eq. (1) to hold for every $h, H \in \mathcal{H}$ with $h < H$. This means that extending the forecast horizon cannot change the predictive law induced on any existing future prefix. Our goal is to achieve this property while generating only the requested horizon, rather than always generating a fixed $H_{\max}$ trajectory and cropping it.

## 3.2 Empirical Diagnosis of Cross-Horizon Inconsistency

We evaluate cross-horizon consistency for D3U, NsDiff, and TSFlow on ETTh1, Weather, and Exchange using the horizon pairs $(96, 192)$, $(96, 720)$, and $(336, 720)$. Since these models require a predefined forecast horizon, we train a separate model for each horizon following their standard setups. For each pair $h < H$, *Direct* denotes the predictive distribution from the $h$-horizon model, whereas *Crop* is obtained by retaining the first $h$ steps of trajectories generated by the $H$-horizon model. Within each comparison, Direct and Crop use the same 336-step history and refer to the same future prefix. Each predictive distribution is represented by 256 generated trajectories.

We measure distributional discrepancy using Energy Distance (ED), a nonparametric statistical distance between multivariate distributions that vanishes when the distributions coincide (Székely & Rizzo, 2013). To distinguish cross-horizon discrepancy from finite-sample Monte Carlo variation, we split each 256-trajectory ensemble into two equal halves. Let $D_{\text{self}}$ denote the average ED between halves drawn from the same predictive distribution, and let $D_{\text{cross}}$ denote the average ED between matched Direct and Crop halves. Our primary measure is the Excess Energy Distance,

$$\Delta_{\text{ED}} = D_{\text{cross}} - D_{\text{self}}, \tag{2}$$

where a positive value indicates that the Direct–Crop discrepancy exceeds the empirical finite-sampling baseline. We additionally report the Cross $>$ Self Rate, i.e., the fraction of test windows with $D_{\text{cross}} > D_{\text{self}}$, the Quantile Gap at the $0.1/0.5/0.9$ predictive quantiles, the Directional Event Probability Gap for whether the temporal average over the shared future prefix exceeds the last observed value, and the corresponding Direction Flip Rate under a $0.5$ threshold. Exact estimators and the moving-block bootstrap procedure are provided in Appendix A.

Table 1 summarizes the cross-horizon discrepancy across models, datasets, and horizon pairs. For all 27 configurations, the 95% confidence intervals for Excess ED lie strictly above zero, and the Cross $>$ Self Rate is $1.000$ in every setting. Thus, the Direct–Crop discrepancy consistently exceeds the empirical within-distribution finite-sampling baseline.

Importantly, the inconsistency is not limited to widely separated horizons. Even for the closest pair, 96–192, the Direction Flip Rate reaches $0.520$ for D3U on Exchange and $0.294$ on ETTh1. Across all settings, the Directional Event Probability Gap reaches $0.545$ and the Direction Flip Rate $0.579$, showing that cross-horizon inconsistency can alter not only predictive distributions but also event probabilities and the resulting thresholded decisions. These effects appear for both diffusion- and flow-based forecasters across all three datasets. Cross-horizon inconsistency is therefore not specific to a particular generative formulation or dataset. Horizon-specific probabilistic forecasts may be individually accurate yet still fail to form a coherent predictive family across horizons. Standard

Table 1: Cross-horizon inconsistency diagnosis across three generative forecasters and three datasets. Within each metric cell, values correspond to 96–192 / 96–720 / 336–720. Full confidence intervals and additional statistics are reported in Appendix B.

| Model | Dataset | Excess ED | Quantile Gap | Event Prob. Gap | Flip Rate |
|---|---|---|---|---|---|
| D3U | ETTh1 | 3.038 / 33.524 / 64.153 | 0.298 / 1.280 / 1.149 | 0.166 / 0.507 / 0.363 | 0.294 / 0.550 / 0.317 |
| D3U | Exchange | 6.150 / 10.554 / 53.362 | 0.272 / 0.340 / 0.659 | 0.267 / 0.400 / 0.545 | 0.520 / 0.456 / 0.579 |
| D3U | Weather | 2.206 / 45.475 / 5.169 | 0.150 / 0.597 / 0.185 | 0.137 / 0.343 / 0.145 | 0.165 / 0.321 / 0.193 |
| NsDiff | ETTh1 | 4.311 / 4.781 / 8.649 | 0.139 / 0.154 / 0.131 | 0.183 / 0.129 / 0.155 | 0.184 / 0.140 / 0.177 |
| NsDiff | Exchange | 3.897 / 7.295 / 22.132 | 0.207 / 0.297 / 0.362 | 0.206 / 0.222 / 0.194 | 0.195 / 0.214 / 0.199 |
| NsDiff | Weather | 4.132 / 7.812 / 15.420 | 0.186 / 0.254 / 0.276 | 0.143 / 0.163 / 0.154 | 0.167 / 0.163 / 0.145 |
| TSFlow | ETTh1 | 1.492 / 4.427 / 11.587 | 0.122 / 0.238 / 0.471 | 0.092 / 0.183 / 0.180 | 0.080 / 0.157 / 0.166 |
| TSFlow | Exchange | 0.586 / 24.288 / 45.932 | 0.096 / 0.538 / 0.686 | 0.062 / 0.370 / 0.305 | 0.057 / 0.389 / 0.274 |
| TSFlow | Weather | 1.964 / 4.952 / 11.985 | 0.100 / 0.161 / 0.205 | 0.054 / 0.128 / 0.163 | 0.064 / 0.156 / 0.194 |

forecasting objectives do not explicitly enforce compatibility on overlapping future prefixes. This motivates treating cross-horizon consistency as a structural property of the forecasting model itself.

# 4 METHOD

## 4.1 OVERVIEW

PrefixFlow is a channel-independent variable-horizon Flow Matching model. For channel $j \in \{1, \dots, D\}$, let $\mathbf{x}_j \in \mathbb{R}^L$ denote the observed history and $\mathbf{y}_j \in \mathbb{R}^H$ the corresponding $H$-step future. Given patch length $P$, we define $N_x = L/P$ and $N_H = H/P$ for any supported patch-aligned horizon $H \le H_{\max}$.

During training at horizon $H$, following Flow Matching (Lipman et al., 2023), we sample $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_H)$ and $t \sim \mathcal{U}(0,1)$, and use the linear interpolation

$$\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{y}_j, \qquad \mathbf{u} = \mathbf{y}_j - \mathbf{z}_0. \tag{3}$$

Here, $\mathbf{u}$ is the target velocity along the interpolation path. PrefixFlow learns a velocity field $\mathbf{v}_\theta^{(H)}(\mathbf{z}_t, t, \mathbf{x}_j, j)$, with parameters $\theta$ shared across channels and all supported horizons.

To maintain consistency under horizon extension, PrefixFlow ensures that existing condition representations remain unchanged and that prefix velocities do not depend on newly appended suffix states. As illustrated in Fig. 2, these properties are enforced by prefix-invariant conditioning and a prefix-causal velocity field, respectively. Together with a projectively consistent Gaussian source, they induce the projectively consistent predictive family established in Sec. 4.5.

## 4.2 PREFIX-INVARIANT CONDITIONING

The conditioning path is designed so that extending the requested horizon only appends new future queries, without altering the conditioning of existing future positions. For a requested horizon $H$, we patch the observed history into $N_x$ history tokens (Nie et al., 2023) and instantiate $N_H = H/P$ future queries, one for each requested future patch:

$$\begin{aligned} \mathbf{h}_i &= \mathbf{W}_x \mathbf{X}_i + \mathbf{p}_i, \qquad i = 1, \dots, N_x, \\ \mathbf{q}_p &= \mathbf{q}_{\text{learn}} + \mathbf{p}_{N_x+p}, \quad p = 1, \dots, N_H. \end{aligned} \tag{4}$$

Here, $\mathbf{W}_x \in \mathbb{R}^{d \times P}$ is the history-patch projection, $\mathbf{p}_k \in \mathbb{R}^d$ is a learnable positional embedding, and $\mathbf{q}_{\text{learn}} \in \mathbb{R}^d$ is a shared future-query embedding. Only the $N_H$ queries corresponding to the requested horizon are instantiated.

The sequence $\mathbf{E}^{(H)} = [\mathbf{h}_1, \dots, \mathbf{h}_{N_x}, \mathbf{q}_1, \dots, \mathbf{q}_{N_H}]$ is processed by a prefix-masked Transformer (Vaswani et al., 2017). History tokens attend only to the observed history, while each future query $\mathbf{q}_p$ attends to the full history and the query prefix $\mathbf{q}_{1:p}$. Consequently, extending the horizon appends new queries without changing the attention context of existing ones.

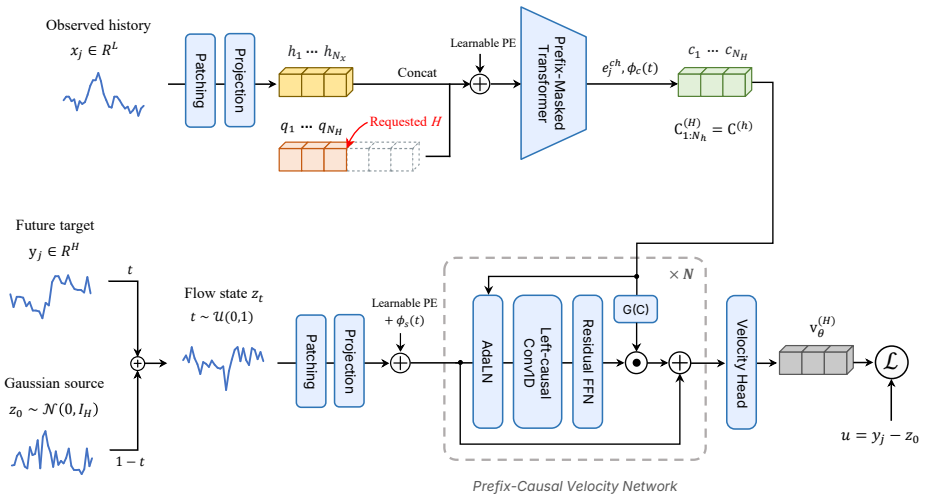


Figure 2: Overall architecture of **PrefixFlow** during training. For a requested horizon $H$, the observed history and $N_H$ future queries are encoded into condition tokens $\mathbf{C}^{(H)}$, while the interpolated Flow Matching state $\mathbf{z}_t$ is projected into aligned state tokens $\mathbf{S}^{(0)}$. The prefix-causal velocity network uses $\mathbf{C}^{(H)}$ to condition the state sequence and predicts the velocity $\mathbf{v}_\theta^{(H)}$.

Let $\mathbf{r}_p \in \mathbb{R}^d$ denote the Transformer representation associated with $\mathbf{q}_p$. Flow-time and channel information are then incorporated to form the condition token:

$$\mathbf{c}_p = \mathrm{LN}\left(\mathbf{W}_c\left[\mathbf{r}_p + \boldsymbol{\phi}_c(t); \mathbf{e}_j^{\mathrm{ch}}\right] + \mathbf{b}_c\right), \qquad p = 1, \ldots, N_H. \tag{5}$$

The condition sequence is $\mathbf{C}^{(H)} = [\mathbf{c}_1, \ldots, \mathbf{c}_{N_H}] \in \mathbb{R}^{N_H \times d}$.

Since horizon extension only appends new future queries, which do not affect the existing ones, the corresponding condition tokens remain unchanged. Thus, for any supported patch-aligned $h < H$, with $N_h = h/P$,

$$\boxed{\mathbf{C}_{1:N_h}^{(H)} = \mathbf{C}^{(h)}}. \tag{6}$$

We refer to this property as *condition-prefix invariance*.

### 4.3 Prefix-Causal Velocity Field

The velocity field is designed so that the dynamics of an existing state prefix are unaffected by appended suffix states. The Flow Matching state is partitioned into $N_H$ patches $\mathbf{Z}_{t,p} \in \mathbb{R}^P$ aligned with the condition tokens. Each patch is embedded as

$$\mathbf{s}_p^{(0)} = \mathbf{W}_z \mathbf{Z}_{t,p} + \mathbf{b}_z + \mathbf{p}_p^z + \boldsymbol{\phi}_s(t), \qquad p = 1, \ldots, N_H, \tag{7}$$

where $\mathbf{W}_z \in \mathbb{R}^{d \times P}$ and $\mathbf{b}_z \in \mathbb{R}^d$ denote the state-patch projection and bias, $\mathbf{p}_p^z \in \mathbb{R}^d$ is a learnable positional embedding, and $\boldsymbol{\phi}_s(t) \in \mathbb{R}^d$ is the flow-time embedding.

The resulting state tokens form $\mathbf{S}^{(0)} = [\mathbf{s}_1^{(0)}, \ldots, \mathbf{s}_{N_H}^{(0)}] \in \mathbb{R}^{N_H \times d}$.

As shown in the lower-right part of Fig. 2, the velocity network stacks $N$ prefix-causal blocks followed by a shared velocity head. Each block uses AdaLN (Peebles & Xie, 2023) to modulate the state with its aligned conditions, while a left-causal Conv1D restricts state mixing to current and earlier positions. A position-wise FFN is applied within a condition-dependent gated residual update.

At block $l$, this computation can be summarized as

$$\begin{aligned}\widetilde{\mathbf{S}}^{(l)} &= \mathrm{AdaLN}^{(l)}\big(\mathbf{S}^{(l-1)}; \mathbf{C}^{(H)}\big),\\ \mathbf{M}^{(l)} &= \mathrm{CausalConv1D}^{(l)}\big(\widetilde{\mathbf{S}}^{(l)}\big),\\ \mathbf{S}^{(l)} &= \mathbf{S}^{(l-1)} + \mathbf{G}^{(l)}(\mathbf{C}^{(H)}) \odot \Big[\mathbf{M}^{(l)} + \mathrm{FFN}^{(l)}(\mathbf{M}^{(l)})\Big].\end{aligned} \tag{8}$$

All operations within each block are position-wise except the left-causal convolution. Consequently, information can propagate only from earlier to later positions, preventing an appended suffix from affecting an existing state prefix. Importantly, prefix causality constrains information flow within the velocity network but does not impose autoregressive generation: the velocities of all requested future patches are predicted jointly in a single forward pass.

After $N$ blocks, a shared output head produces the velocity for each future patch:

$$\begin{aligned}\widehat{\mathbf{u}}_p &= \mathbf{W}_o\,\mathrm{LN}(\mathbf{s}_p^{(N)}) + \mathbf{b}_o,\\ \mathbf{v}_\theta^{(H)}(\mathbf{z}_t, t, \mathbf{x}_j, j) &= [\widehat{\mathbf{u}}_1; \ldots; \widehat{\mathbf{u}}_{N_H}] \in \mathbb{R}^H.\end{aligned} \tag{9}$$

Together with the condition-prefix invariance in Eq. (6), this ensures that extending the horizon does not alter the velocity on an existing prefix. Therefore, for every supported patch-aligned $h < H$,

$$\boxed{\pi_h\, \mathbf{v}_\theta^{(H)}(\mathbf{z}_t, t, \mathbf{x}_j, j) = \mathbf{v}_\theta^{(h)}(\pi_h \mathbf{z}_t, t, \mathbf{x}_j, j)}. \tag{10}$$

We refer to this property as *velocity prefix closure*.

### 4.4 Multi-Horizon Training and Inference

**Multi-horizon training.** Let $\mathcal{H}_{\mathrm{train}}$ denote the set of training horizons. For each $H \in \mathcal{H}_{\mathrm{train}}$, we construct the corresponding Flow Matching training pair $(\mathbf{z}_t, \mathbf{u})$ using Eq. (3).

Because earlier future positions are included in more training horizons, we balance their contributions according to their horizon coverage. Let $c_i = \sum_{H' \in \mathcal{H}_{\mathrm{train}}} \mathbb{I}(i \le H')$ denote the number of training horizons containing position $i$. We define the normalized weight as

$$\bar{w}_i^{(H)} = \frac{1/\sqrt{c_i}}{\sum_{k=1}^{H} 1/\sqrt{c_k}}.$$

The joint training objective is

$$\mathcal{L} = \frac{1}{|\mathcal{H}_{\mathrm{train}}|} \sum_{H \in \mathcal{H}_{\mathrm{train}}} \mathbb{E}\left[\sum_{i=1}^{H} \bar{w}_i^{(H)} \left(v_{\theta,i}^{(H)} - u_i\right)^2\right]. \tag{11}$$

The losses across horizons are optimized jointly through the shared parameters $\theta$.

**Inference.** For a requested patch-aligned horizon $H$, PrefixFlow instantiates only $N_H$ future queries and state tokens and samples $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_H)$. A forecast trajectory is obtained by solving

$$\frac{d\mathbf{z}_t}{dt} = \mathbf{v}_\theta^{(H)}(\mathbf{z}_t, t, \mathbf{x}_j, j), \qquad t \in [0, 1], \tag{12}$$

from $\mathbf{z}_0$ to $\widehat{\mathbf{y}}_j = \mathbf{z}_1$. Thus, inference cost depends on the requested horizon rather than on $H_{\max}$. We use a four-step non-uniform Euler solver in all experiments.

### 4.5 Projective Consistency

The preceding sections establish prefix-invariant conditioning and velocity prefix closure under horizon extension; together with the projectively consistent Gaussian source family, these properties yield projectively consistent predictive distributions.

For a fixed channel $j$ and observed history $\mathbf{x}_j$, let $Q_H^{(j)}(\cdot \mid \mathbf{x}_j)$ denote the predictive distribution at horizon $H$. Since PrefixFlow generates channels independently, the full predictive law factorizes across these channel-wise distributions. PrefixFlow uses the standard Gaussian source $\mu_H = \mathcal{N}(\mathbf{0}, \mathbf{I}_H)$, which satisfies $(\pi_h)_{\#}\mu_H = \mu_h$ for every $h < H$.

**Proposition 1** (Projective consistency)**.** *For any supported patch-aligned horizons $h < H$, suppose the corresponding Flow Matching ODEs admit unique solutions. Then PrefixFlow satisfies*

$$Q_h(\cdot \mid \mathbf{x}) = (\pi_h)_{\#} Q_H(\cdot \mid \mathbf{x}). \tag{13}$$

*Proof sketch.* Fix a channel $j$ and let $\Phi_H^t$ denote the ODE flow at horizon $H$. Let $\mathbf{z}_0^{(H)} \sim \mu_H$ and set $\mathbf{z}_0^{(h)} = \pi_h \mathbf{z}_0^{(H)} \sim \mu_h$. By velocity prefix closure in Eq. (10), the projected $H$-horizon trajectory and the $h$-horizon trajectory satisfy the same initial-value problem. Uniqueness therefore gives

$$\pi_h \Phi_H^t(\mathbf{z}_0) = \Phi_h^t(\pi_h \mathbf{z}_0), \qquad t \in [0, 1].$$

Hence the flow maps commute with prefix projection: $\pi_h \circ \Phi_H^t = \Phi_h^t \circ \pi_h$.

Since $Q_H^{(j)} = (\Phi_H^1)_{\#} \mu_H$, we obtain

$$\begin{aligned} (\pi_h)_{\#} Q_H^{(j)} &= (\pi_h \circ \Phi_H^1)_{\#} \mu_H \\ &= (\Phi_h^1 \circ \pi_h)_{\#} \mu_H \\ &= (\Phi_h^1)_{\#} (\pi_h)_{\#} \mu_H \\ &= (\Phi_h^1)_{\#} \mu_h = Q_h^{(j)}. \end{aligned} \tag{14}$$

Channel-wise factorization then yields Eq. (13) for the full predictive law. □

The same argument extends to Euler discretization when the same time grid is used across horizons. A complete continuous- and discrete-time proof is provided in Appendix C.

# 5 Experiments

## 5.1 Experimental Setup

**Datasets and Settings.** We evaluate PrefixFlow on six widely used time-series forecasting datasets: ETTh1 and ETTm1 from the ETT benchmark (Zhou et al., 2021), and Weather, Exchange, ECL, and ILI (Wu et al., 2021). For the first five datasets, we use a history length of 336 and evaluate forecast horizons $H \in \{96, 192, 336, 720\}$. For ILI, we use a history length of 36 and horizons $H \in \{24, 36, 48, 60\}$.

**Baselines.** We compare PrefixFlow with five recent generative probabilistic forecasters covering different generative and decoding paradigms: the non-autoregressive diffusion models TimeDiff (Shen & Kwok, 2023), D3U (Li et al., 2025), and NsDiff (Ye et al., 2025); the autoregressive diffusion model TEDM (Solano Carrillo et al., 2026); and the Flow-Matching-based TSFlow (Kollovieh et al., 2025). Unlike the baselines, which are trained separately for each forecast horizon, PrefixFlow uses a single model jointly trained across horizons. All methods use the same data splits, history lengths, and forecasting targets.

**Evaluation Protocol.** For each test instance, we generate 100 predictive samples. Their sample mean defines the point forecast used for MSE (Zhou et al., 2021), and CRPS (Gneiting & Raftery, 2007) is estimated directly from the predictive samples. Both metrics are averaged over all test instances, forecast steps, and variables. We report MSE and CRPS in the main text; additional evaluation and implementation details are provided in Appendix D.3, while the full results, including MAE and QICE, are reported in Appendix E.

## 5.2 Forecasting Performance

Table 2 summarizes the forecasting performance across the six benchmarks. PrefixFlow achieves the lowest average CRPS on five datasets. Relative to the best competing baseline, PrefixFlow reduces average CRPS by 14.1% on Weather, 14.0% on Exchange, and 30.2% on ILI. It also performs strongly in point forecasting, achieving the lowest average MSE on five datasets. Importantly, PrefixFlow achieves these results with a single model jointly trained across horizons, whereas the competing methods are trained separately for each horizon. This indicates that projective consistency can be enforced without sacrificing forecasting accuracy.

Table 2: Forecasting performance averaged over four prediction horizons. Lower values are better. Best and second-best results are shown in **bold** and underlined, respectively.

| | ETTh1 | | ETTm1 | | Weather | | Exchange | | ECL | | ILI | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | MSE | CRPS | MSE | CRPS | MSE | CRPS | MSE | CRPS | MSE | CRPS | MSE | CRPS |
| TimeDiff | 0.4289 | 0.4126 | 0.3733 | 0.3677 | 0.2701 | 0.2921 | 0.5025 | 0.5001 | 0.2056 | 0.3013 | 3.5109 | 1.2817 |
| D3U | 0.4590 | 0.3479 | 0.3640 | 0.2859 | 0.2535 | 0.2256 | 0.6151 | 0.4608 | 0.1642 | **0.1935** | 2.5094 | 0.8890 |
| NsDiff | 0.5940 | 0.3990 | 0.4590 | 0.3401 | 0.2798 | 0.2544 | 0.7708 | 0.4996 | 0.2198 | 0.2881 | 2.8542 | 0.9232 |
| TSFlow | 0.5343 | 0.4021 | 0.7023 | 0.4799 | 0.2924 | 0.2504 | 0.4224 | 0.4165 | 0.3649 | 0.3208 | 4.6242 | 1.3468 |
| TEDM | 0.7614 | 0.4297 | 0.8146 | 0.4379 | 0.3135 | 0.2710 | **0.3841** | 0.3689 | 0.8504 | 0.6408 | 4.9721 | 1.3249 |
| **PrefixFlow (Ours)** | **0.4195** | **0.3182** | **0.3359** | **0.2751** | **0.2239** | **0.1938** | 0.4102 | **0.3172** | **0.1633** | 0.1977 | **2.0016** | **0.6201** |

### 5.3 Cross-Horizon Consistency

Figure 3(a) tests whether PrefixFlow preserves the predictive distribution over a shared future prefix across requested horizons. On ETTh1, Weather, and Exchange, all nine horizon-pair comparisons yield Excess Energy Distances close to zero, and all 95% bootstrap confidence intervals include zero. Thus, we find no statistically detectable cross-horizon discrepancy beyond finite-sample variation. Exchange shows the largest Excess ED for the $(336, 720)$ pair, likely because fewer shared forecast origins make the estimate more sensitive to finite-sample variation. Its confidence interval nevertheless still includes zero. The qualitative example in Figure 3(b) further shows that direct $H = 96$ and $H = 192$ forecasts produce closely aligned predictive medians and 10–90% predictive quantile intervals over their shared 96-step prefix. These empirical results are consistent with the projective-consistency guarantee of PrefixFlow. Complete consistency diagnostics are reported in Appendix F.1.

### 5.4 Model Analysis

Table 3: Ablation study on ETTh1. Consistency is evaluated for the $(96, 720)$ horizon pair. MSE and CRPS are averaged over $H \in \{96, 192, 336, 720\}$.

| Variant | Excess ED | Cross > Self | MSE | CRPS |
|---|---|---|---|---|
| Full | -0.0009 | 45.8% | 0.4195 | 0.3182 |
| w/o Prefix Mask | 0.1122 | 99.6% | 0.4179 | 0.3168 |
| w/o Causal Conv | 0.0702 | 98.5% | 0.4255 | 0.3210 |
| w/o Both | 0.1688 | 100.0% | 0.4191 | 0.3151 |
| Separate Training | 1.5939 | 100.0% | – | – |

**Ablation Study.** Table 3 isolates the mechanisms that enforce cross-horizon consistency. The full model yields an Excess ED close to zero, with a Cross $>$ Self rate of 45.8%, close to the 50% reference expected under cross-horizon consistency. Removing the prefix mask increases Excess ED to 0.1122 and raises this rate to 99.6%, showing that later requested positions must not alter the conditioning representation of an existing prefix. Replacing the causal convolution with a symmetric one also introduces cross-horizon discrepancies, highlighting the importance of prefix closure in the velocity field. Removing both constraints further increases Excess ED to 0.1688. In contrast, MSE and CRPS remain nearly unchanged across these variants, indicating that cross-horizon consistency is not captured by standard forecasting metrics and that PrefixFlow achieves it without sacrificing forecasting performance.

Separate Training provides a complementary reference. This variant uses the same architecture as *w/o Both*, but trains the $H = 96$ and $H = 720$ models independently, removing parameter sharing and multi-horizon supervision. Its Excess ED rises from 0.1688 to 1.5939, with Cross $>$ Self on every test window. This result shows that fixed-horizon models trained independently need not form a compatible predictive family. Prefix-causal structure provides the consistency guarantee, while shared multi-horizon training further encourages empirical alignment across horizons.

**Generalization to Unseen Horizons.** Although trained only on $H \in \{96, 192, 336, 720\}$, PrefixFlow directly forecasts unseen patch-aligned horizons $H \in \{120, 240, 480, 600\}$ without retraining. As shown in Figure 3(c), MSE and CRPS at these horizons follow the same overall trend as the seen horizons on both ETTh1 and Weather, with no systematic degradation. We further evaluate cross-horizon consistency among unseen horizons on ETTh1. Across the three evaluated unseen–unseen horizon pairs, the 95% confidence intervals include zero for Excess ED and 50% for Cross $>$ Self (Appendix F.3). Thus, PrefixFlow generalizes to unseen horizons within the supported range and preserves both forecasting quality and cross-horizon consistency.

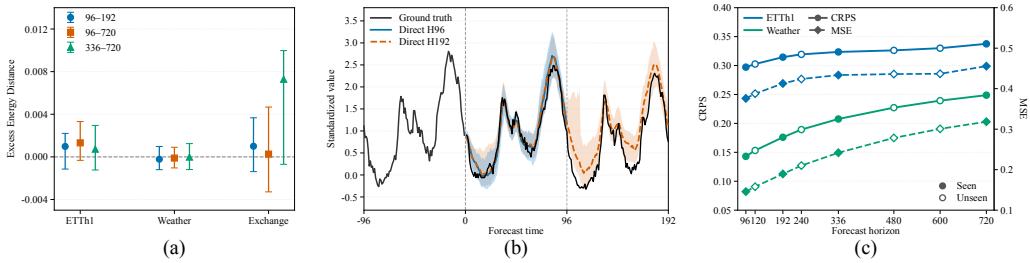


Figure 3: (a) Cross-horizon consistency measured by Excess Energy Distance on ETTh1, Weather, and Exchange for three representative horizon pairs. Error bars denote 95% moving-block bootstrap confidence intervals, and the dashed line marks zero excess discrepancy. (b) Qualitative comparison of direct $H = 96$ and $H = 192$ forecasts on ETTm1. Both forecasts condition on 336 historical steps, of which only the last 96 are shown for clarity. Solid lines denote predictive medians, and shaded regions denote the 10–90% predictive quantile intervals. (c) Forecasting performance at seen and unseen patch-aligned horizons on ETTh1 and Weather, evaluated by CRPS and MSE.

Table 4: Inference efficiency on ETTh1 with batch size 128, comparing direct requested-horizon generation against always generating $H = 720$ and cropping.

| | Latency | | | Peak Memory | | |
|---|---|---|---|---|---|---|
| $H$ | Direct (ms) | $H_{720}$-Crop (ms) | Speedup | Direct (MiB) | $H_{720}$-Crop (MiB) | Reduction |
| 96 | 228.03±12.92 | 878.08±36.36 | **3.85**× | 381.5 | 1216.2 | **68.6%** |
| 192 | 312.94±15.74 | 879.67±42.31 | 2.81× | 472.3 | 1216.2 | 61.2% |
| 336 | 439.09±22.86 | 899.80±40.13 | 2.05× | 647.3 | 1216.2 | 46.8% |
| 720 | 884.92±32.19 | 904.95±42.34 | 1.02× | 1216.2 | 1216.2 | 0.0% |

**Inference Efficiency.** Generating only the requested horizon avoids the redundant computation of forecasting to $H_{\max}$ and then cropping. As shown in Table 4, PrefixFlow achieves $3.85\times$, $2.81\times$, and $2.05\times$ speedups at $H = 96$, 192, and 336, respectively, while reducing peak GPU memory by 68.6%, 61.2%, and 46.8%. At $H = 720$, both strategies have nearly identical computational cost because they generate the same forecast length.

## 6 Conclusion

We study horizon-wise projective consistency in variable-horizon probabilistic forecasting. Probabilistic forecasts at different horizons are not separate predictions, but marginals of the same future distribution. Our diagnosis shows that direct generative forecasters trained separately for different horizons can produce systematic discrepancies over shared future prefixes beyond finite-sample Monte Carlo variation, which can further affect event probabilities and thresholded decisions. To address this issue, we introduce PrefixFlow, a variable-horizon Flow Matching model that directly generates the requested horizon with projective consistency guaranteed by construction. Across six benchmarks, PrefixFlow achieves strong point and probabilistic forecasting performance while maintaining projective consistency across horizons.

### AI use statement

The research ideas, methodology, experiments, analysis, and initial manuscript were developed by the authors. Generative AI tools were used for language polishing, refining the conceptual and mathematical presentation, assisting with the presentation of mathematical proofs, and improving the clarity and organization of the manuscript. All AI-assisted content and mathematical claims were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

## Reproducibility Statement

The appendix provides details on data preprocessing, model configuration, training and inference settings, baseline protocols, and evaluation procedures. An anonymized implementation of PrefixFlow is provided in the supplementary material.

## References


Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Expert Certification.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167, 21–27 Jul 2024.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

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

Olav Kallenberg. *Foundations of Modern Probability*. Springer Cham, 3rd edition, 2021. doi: 10.1007/978-3-030-61871-1.

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*, pp. 95435–95457, 2025.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. doi: 10.1214/aos/1176347265.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 95–104, 2018.

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*, pp. 67475–67503, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8857–8868, 18–24 Jul 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. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2019.07.001.

Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31016–31029, 23–29 Jul 2023.

Edgardo Solano Carrillo, Sreerag Vadakkemeppully Naveenachandran, and Julia Niebling. TEDM: Time series forecasting with elucidated diffusion models. In *International Conference on Learning Representations*, pp. 6696–6726, 2026.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. ISSN 0378-3758. doi: 10.1016/j.jspi.2013.03.018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164, 21–27 Jul 2024.

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

Jin Xu, Emilien Dupont, Kaspar Märtens, Thomas Rainforth, and Yee Whye Teh. Deep stochastic processes via functional Markov transition operators. In *Advances in Neural Information Processing Systems*, volume 36, pp. 37975–37994, 2023.

Vijaya Krishna Yalavarthi, Randolf Scholz, Christian Klötergens, Kiran Madhusudhanan, Stefan Born, and Lars Schmidt-Thieme. Reliable probabilistic forecasting of irregular time series through marginalization-consistent flows. In *International Conference on Learning Representations*, pp. 85403–85427, 2026.

Weiwei Ye, Zhuopeng Xu, and Ning Gui. Non-stationary diffusion for probabilistic time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72112–72130, 13–19 Jul 2025.

Jiawen Zhang, Shun Zheng, Xumeng Wen, Xiaofang Zhou, Jiang Bian, and Jia Li. ElasTST: Towards robust varied-horizon forecasting with elastic time-series transformer. In *Advances in Neural Information Processing Systems*, volume 37, pp. 119174–119197, 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

# APPENDICES

## A CONSISTENCY EVALUATION PROTOCOL

This section details the protocol used to evaluate whether predictive distributions remain compatible when the requested forecast horizon changes. For a horizon pair $h < H$, we compare the distribution obtained by directly forecasting $h$ steps with the marginal induced by an $H$-step forecast over the same $h$-step future prefix.

### A.1 COMPARISON SETUP

For an observed history $\mathbf{x}_i$, let

$$\mathbf{Y}_i^D \sim Q_h(\cdot \mid \mathbf{x}_i) \tag{15}$$

denote the *Direct* forecast at horizon $h$, and let

$$\mathbf{Y}_i^C \sim (\pi_h)_{\#} Q_H(\cdot \mid \mathbf{x}_i) \tag{16}$$

denote the *Crop* forecast obtained by retaining the first $h$ steps of an $H$-step prediction. The main diagnosis considers the horizon pairs $(96, 192)$, $(96, 720)$, and $(336, 720)$.

For every pair, Direct and Crop share the same forecast origin, 336-step history, and future target prefix. Each predictive distribution is represented by 256 generated trajectories. Only forecast origins available for both horizons are retained, and all samples are canonicalized to the shape $(\text{samples}, \text{horizon}, \text{variables})$ before comparison.

Because adjacent rolling windows overlap substantially, we evaluate every eighth archived window. This subsampling is deterministic and independent of the metric values. For ETTh1, the diagnosis uses the target variable `OT`. For Weather and Exchange, Energy Distance is evaluated on the full multivariate trajectory, whereas Quantile Gap, Directional Event Probability Gap, and Direction Flip Rate are computed separately for each variable and then macro-averaged across variables.

### A.2 ENERGY-BASED CONSISTENCY MEASURES

For random vectors $X \sim P$ and $Y \sim Q$, the Energy Distance is

$$ED(P, Q) = 2\mathbb{E}\|X - Y\|_2 - \mathbb{E}\|X - X'\|_2 - \mathbb{E}\|Y - Y'\|_2, \tag{17}$$

where $X'$ and $Y'$ are independent copies (Székely & Rizzo, 2013). Given samples $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$, we use the V-statistic estimator

$$\begin{aligned}\widehat{ED}(X, Y) = \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \|x_i - y_j\|_2 \\ - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - x_j\|_2 - \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \|y_i - y_j\|_2.\end{aligned} \tag{18}$$

Each trajectory is flattened across the shared future steps and variables before computing the Euclidean distances, so the metric compares the joint trajectory distribution over the shared prefix.

To separate cross-horizon discrepancy from finite-sample Monte Carlo variation, for each evaluated test window $i$, the 256 Direct samples are split into two equal subsets $D_{i,1}, D_{i,2}$, and the 256 Crop samples into $C_{i,1}, C_{i,2}$. We define

$$D_{\text{self},i} = \frac{ED(D_{i,1}, D_{i,2}) + ED(C_{i,1}, C_{i,2})}{2}, \tag{19}$$

and

$$D_{\text{cross},i} = \frac{ED(D_{i,1}, C_{i,1}) + ED(D_{i,2}, C_{i,2})}{2}. \tag{20}$$

Both quantities therefore compare sample sets of equal size.

We use the Excess Energy Distance as the primary measure of cross-horizon discrepancy. For window $i$,

$$\Delta_{\mathrm{ED},i} = D_{\mathrm{cross},i} - D_{\mathrm{self},i}, \tag{21}$$

and the reported Excess Energy Distance is

$$\Delta_{\mathrm{ED}} = \frac{1}{N}\sum_{i=1}^{N} \Delta_{\mathrm{ED},i}, \tag{22}$$

where $N$ is the number of evaluated test windows. This statistic quantifies the average cross-horizon discrepancy beyond the empirical finite-sampling baseline.

We additionally report the Cross $>$ Self Rate,

$$R_{\mathrm{CS}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[D_{\mathrm{cross},i} > D_{\mathrm{self},i}\right]. \tag{23}$$

When Direct and Crop differ only by finite-sample variation, $D_{\mathrm{cross}}$ has no systematic tendency to exceed $D_{\mathrm{self}}$, giving a reference level near $0.5$. Values close to $1.0$ indicate that the cross-horizon discrepancy exceeds the self baseline for nearly all evaluated windows.

As an auxiliary statistic, we also report the aggregate Cross/Self Energy Ratio

$$R_{\mathrm{ED}} = \frac{\frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{cross},i}}{\frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{self},i}}. \tag{24}$$

A value near $1.0$ indicates comparable cross- and within-distribution discrepancies. Because the ratio is sensitive to small values of $D_{\mathrm{self}}$, we use it only as a descriptive statistic and do not assign it a separate bootstrap interval. Because Excess ED retains the scale of Energy Distance, its magnitude is not directly compared across settings with different trajectory dimensions.

### A.3 Prefix-Level Distributional Diagnostics

Energy Distance measures discrepancy in the joint trajectory distribution. We further quantify how this discrepancy appears in predictive quantiles and event probabilities over the same shared future prefix.

For quantile levels $q \in \{0.1, 0.5, 0.9\}$, let $Q^{D}_{i,c,t,q}$ and $Q^{C}_{i,c,t,q}$ denote the Direct and Crop predictive quantiles for test window $i$, variable $c$, and future step $t$. For each window–variable pair, we define

$$G^{Q}_{i,c} = \frac{1}{3h}\sum_{q\in\{0.1,0.5,0.9\}}\sum_{t=1}^{h} \left|Q^{D}_{i,c,t,q} - Q^{C}_{i,c,t,q}\right|. \tag{25}$$

For multivariate evaluation, we first average across variables within each window:

$$G^{Q}_{i} = \frac{1}{D}\sum_{c=1}^{D} G^{Q}_{i,c}. \tag{26}$$

The reported Quantile Gap is then

$$G^{Q} = \frac{1}{N}\sum_{i=1}^{N} G^{Q}_{i}. \tag{27}$$

All quantile gaps are computed in the train-standardized space.

For window $i$ and variable $c$, we define the directional event

$$E_{i,c} = \left\{\frac{1}{h}\sum_{t=1}^{h} Y_{i,t,c} > x_{i,\mathrm{last},c}\right\}, \tag{28}$$

where $x_{i,\mathrm{last},c}$ is the last historical observation. Using all 256 predictive trajectories, let $p^{D}_{i,c}$ and $p^{C}_{i,c}$ denote the empirical probabilities of this event under Direct and Crop, respectively. The window–variable probability gap is

$$G^{P}_{i,c} = \left|p^{D}_{i,c} - p^{C}_{i,c}\right|. \tag{29}$$

For multivariate evaluation,

$$G_i^P = \frac{1}{D}\sum_{c=1}^{D} G_{i,c}^P, \tag{30}$$

and the reported Directional Event Probability Gap is

$$G^P = \frac{1}{N}\sum_{i=1}^{N} G_i^P. \tag{31}$$

Finally, thresholding the same event probabilities at $0.5$ gives

$$a_{i,c}^D = \mathbf{1}\left[p_{i,c}^D > 0.5\right], \qquad a_{i,c}^C = \mathbf{1}\left[p_{i,c}^C > 0.5\right]. \tag{32}$$

We define the corresponding flip indicator as

$$F_{i,c} = \mathbf{1}\left[a_{i,c}^D \neq a_{i,c}^C\right]. \tag{33}$$

For multivariate evaluation, the window-level flip rate is

$$F_i = \frac{1}{D}\sum_{c=1}^{D} F_{i,c}, \tag{34}$$

and the reported Direction Flip Rate is

$$R_{\text{flip}} = \frac{1}{N}\sum_{i=1}^{N} F_i. \tag{35}$$

For ETTh1, $D = 1$ because the diagnosis is performed on the target variable `OT`. For Weather and Exchange, the scalar diagnostics above are macro-averaged across variables within each test window before averaging across windows.

### A.4 Bootstrap Confidence Intervals

Adjacent rolling forecast windows share substantial temporal context and cannot be treated as independent observations. We therefore estimate uncertainty using a moving-block bootstrap (Künsch, 1989). With a window stride of $8$, an original block length of $336$ time steps corresponds to

$$B_{\text{eff}} = \left\lceil \frac{336}{8} \right\rceil = 42 \tag{36}$$

evaluated windows.

Each bootstrap replicate samples contiguous blocks of 42 evaluated windows with replacement until the original evaluation length is reached. We use 1000 bootstrap replicates and report percentile-based 95% confidence intervals for Excess Energy Distance, Cross $>$ Self Rate, Quantile Gap, Directional Event Probability Gap, and Direction Flip Rate. The aggregate Energy Ratio is reported without a bootstrap interval.

## B Full Cross-Horizon Diagnosis Results

Table 5 reports the complete cross-horizon diagnosis for D3U, NsDiff, and TSFlow across ETTh1, Weather, and Exchange. For each model–dataset combination, we evaluate the horizon pairs $(96, 192)$, $(96, 720)$, and $(336, 720)$ using the protocol described in Appendix A.

Across all 27 evaluated settings, the 95% confidence intervals of Excess Energy Distance lie strictly above zero, and the Cross $>$ Self Rate is $1.000$ in every case. Cross-horizon inconsistency is therefore not an isolated behavior of a particular model or dataset, but is consistently observed across all three forecasters and all three datasets. Notably, this pattern appears in both the diffusion-based models, D3U and NsDiff, and the Flow-Matching-based TSFlow, indicating that it is not specific to a single generative paradigm.

Table 5: Complete cross-horizon diagnosis across three generative forecasters, three datasets, and three horizon pairs. Excess ED denotes Excess Energy Distance; Ratio denotes the aggregate Cross/Self Energy Ratio; Prob. Gap denotes the Directional Event Probability Gap. Bracketed values are 95% moving-block bootstrap confidence intervals.

| Model | Dataset | Pair | Excess ED | Cross > Self | Ratio | Quantile Gap | Prob. Gap | Flip Rate |
|---|---|---|---|---|---|---|---|---|
| D3U | ETTh1 | 96–192 | 3.038 [2.782, 3.365] | 1.000 [1.000, 1.000] | 11.534 | 0.298 [0.221, 0.396] | 0.166 [0.150, 0.191] | 0.294 [0.228, 0.371] |
| D3U | ETTh1 | 96–720 | 33.524 [33.151, 35.908] | 1.000 [1.000, 1.000] | 183.403 | 1.280 [1.144, 1.515] | 0.507 [0.498, 0.560] | 0.550 [0.498, 0.675] |
| D3U | ETTh1 | 336–720 | 64.153 [62.610, 68.778] | 1.000 [1.000, 1.000] | 218.206 | 1.149 [1.006, 1.382] | 0.363 [0.325, 0.431] | 0.317 [0.255, 0.413] |
| D3U | Exchange | 96–192 | 6.150 [5.624, 6.902] | 1.000 [1.000, 1.000] | 33.231 | 0.272 [0.263, 0.296] | 0.267 [0.237, 0.311] | 0.520 [0.415, 0.665] |
| D3U | Exchange | 96–720 | 10.554 [8.589, 11.456] | 1.000 [1.000, 1.000] | 79.602 | 0.340 [0.318, 0.364] | 0.400 [0.360, 0.412] | 0.456 [0.399, 0.494] |
| D3U | Exchange | 336–720 | 53.362 [42.271, 59.804] | 1.000 [1.000, 1.000] | 188.513 | 0.659 [0.583, 0.709] | 0.545 [0.446, 0.582] | 0.579 [0.476, 0.613] |
| D3U | Weather | 96–192 | 2.206 [2.114, 2.316] | 1.000 [1.000, 1.000] | 6.445 | 0.150 [0.147, 0.155] | 0.137 [0.130, 0.145] | 0.165 [0.149, 0.182] |
| D3U | Weather | 96–720 | 45.475 [41.340, 50.520] | 1.000 [1.000, 1.000] | 245.552 | 0.597 [0.549, 0.648] | 0.343 [0.330, 0.358] | 0.321 [0.302, 0.339] |
| D3U | Weather | 336–720 | 5.169 [4.923, 5.433] | 1.000 [1.000, 1.000] | 6.672 | 0.185 [0.179, 0.190] | 0.145 [0.136, 0.153] | 0.193 [0.172, 0.212] |
| NsDiff | ETTh1 | 96–192 | 4.311 [3.830, 4.952] | 1.000 [1.000, 1.000] | 10.410 | 0.139 [0.116, 0.164] | 0.183 [0.110, 0.233] | 0.184 [0.098, 0.240] |
| NsDiff | ETTh1 | 96–720 | 4.781 [4.020, 5.821] | 1.000 [1.000, 1.000] | 11.137 | 0.154 [0.144, 0.162] | 0.129 [0.075, 0.172] | 0.140 [0.077, 0.196] |
| NsDiff | ETTh1 | 336–720 | 8.649 [7.034, 10.065] | 1.000 [1.000, 1.000] | 10.727 | 0.131 [0.121, 0.143] | 0.155 [0.098, 0.203] | 0.177 [0.107, 0.225] |
| NsDiff | Exchange | 96–192 | 3.897 [2.652, 4.649] | 1.000 [1.000, 1.000] | 24.215 | 0.207 [0.170, 0.229] | 0.206 [0.156, 0.240] | 0.195 [0.144, 0.232] |
| NsDiff | Exchange | 96–720 | 7.295 [5.292, 8.743] | 1.000 [1.000, 1.000] | 42.667 | 0.297 [0.259, 0.343] | 0.222 [0.198, 0.311] | 0.214 [0.188, 0.299] |
| NsDiff | Exchange | 336–720 | 22.132 [16.338, 28.902] | 1.000 [1.000, 1.000] | 72.597 | 0.362 [0.326, 0.423] | 0.194 [0.180, 0.293] | 0.199 [0.184, 0.295] |
| NsDiff | Weather | 96–192 | 4.132 [3.497, 4.880] | 1.000 [1.000, 1.000] | 10.432 | 0.186 [0.176, 0.199] | 0.143 [0.131, 0.155] | 0.167 [0.154, 0.182] |
| NsDiff | Weather | 96–720 | 7.812 [6.638, 9.071] | 1.000 [1.000, 1.000] | 21.335 | 0.254 [0.236, 0.271] | 0.163 [0.151, 0.173] | 0.163 [0.150, 0.176] |
| NsDiff | Weather | 336–720 | 15.420 [13.515, 17.371] | 1.000 [1.000, 1.000] | 21.675 | 0.276 [0.259, 0.292] | 0.154 [0.139, 0.167] | 0.145 [0.129, 0.160] |
| TSFlow | ETTh1 | 96–192 | 1.492 [1.336, 1.726] | 1.000 [1.000, 1.000] | 9.452 | 0.122 [0.119, 0.124] | 0.092 [0.068, 0.107] | 0.080 [0.047, 0.113] |
| TSFlow | ETTh1 | 96–720 | 4.427 [3.725, 5.427] | 1.000 [1.000, 1.000] | 20.862 | 0.238 [0.211, 0.253] | 0.183 [0.090, 0.212] | 0.157 [0.045, 0.175] |
| TSFlow | ETTh1 | 336–720 | 11.587 [10.077, 13.255] | 1.000 [1.000, 1.000] | 26.208 | 0.471 [0.380, 0.548] | 0.180 [0.075, 0.197] | 0.166 [0.049, 0.188] |
| TSFlow | Exchange | 96–192 | 0.586 [0.424, 0.690] | 1.000 [1.000, 1.000] | 4.169 | 0.096 [0.088, 0.102] | 0.062 [0.046, 0.073] | 0.057 [0.041, 0.072] |
| TSFlow | Exchange | 96–720 | 24.288 [23.847, 24.636] | 1.000 [1.000, 1.000] | 248.430 | 0.538 [0.531, 0.542] | 0.370 [0.307, 0.372] | 0.389 [0.310, 0.397] |
| TSFlow | Exchange | 336–720 | 45.932 [45.270, 46.445] | 1.000 [1.000, 1.000] | 151.547 | 0.686 [0.682, 0.687] | 0.305 [0.253, 0.304] | 0.274 [0.214, 0.274] |
| TSFlow | Weather | 96–192 | 1.964 [1.220, 2.695] | 1.000 [1.000, 1.000] | 9.143 | 0.100 [0.096, 0.104] | 0.054 [0.051, 0.058] | 0.064 [0.059, 0.072] |
| TSFlow | Weather | 96–720 | 4.952 [4.397, 5.432] | 1.000 [1.000, 1.000] | 21.461 | 0.161 [0.153, 0.168] | 0.128 [0.114, 0.145] | 0.156 [0.135, 0.182] |
| TSFlow | Weather | 336–720 | 11.985 [10.671, 13.177] | 1.000 [1.000, 1.000] | 24.473 | 0.205 [0.193, 0.215] | 0.163 [0.141, 0.192] | 0.194 [0.165, 0.234] |

A stronger pattern emerges when the shared future prefix is fixed at 96 steps. Comparing $(96, 192)$ with $(96, 720)$ changes only the requested forecast horizon while evaluating the same 96-step prefix. Excess ED increases in all nine model–dataset combinations, and Quantile Gap follows the same pattern. Thus, as the requested forecast extends further into the future, the marginal distribution induced over the same 96-step prefix can drift further away from the distribution obtained by forecasting those 96 steps directly.

This discrepancy is also reflected in more specific predictive summaries, although the different diagnostics do not move in lockstep. Across all settings, Quantile Gap ranges from $0.096$ to $1.280$, Directional Event Probability Gap from $0.054$ to $0.545$, and Direction Flip Rate from $0.057$ to $0.579$. For example, on ETTh1 with NsDiff, Excess ED increases from $4.311$ for $(96, 192)$ to $4.781$ for $(96, 720)$, while the Directional Event Probability Gap decreases from $0.183$ to $0.129$ and the Direction Flip Rate from $0.184$ to $0.140$. Hence, a larger discrepancy in the joint trajectory distribution does not necessarily translate into a larger change in every downstream summary. The different diagnostics therefore characterize complementary manifestations of cross-horizon inconsistency, from changes in the full predictive distribution to changes in quantiles, event probabilities, and thresholded decisions.

## C Proof of Projective Consistency

We first establish projective consistency for a fixed channel and then extend the result to the full multivariate predictive law. Fix a channel $j$ and its observed history $\mathbf{x}_j$. Let $P$ denote the patch length and $H_{\max}$ the maximum supported prediction horizon, with

$$\mathcal{H} = \{mP : m \in \mathbb{N}_+,\ mP \le H_{\max}\}. \tag{37}$$

Thus, the guarantee applies to all patch-aligned horizons in $\mathcal{H}$.

For $h, H \in \mathcal{H}$ with $h < H$, let

$$\pi_h : \mathbb{R}^H \to \mathbb{R}^h$$

denote the projection onto the first $h$ forecast coordinates. At horizon $H$, PrefixFlow defines the conditional ODE

$$\frac{d\mathbf{z}_t^{(H)}}{dt} = \mathbf{v}_\theta^{(H)}\left(\mathbf{z}_t^{(H)}, t, \mathbf{x}_j, j\right), \qquad \mathbf{z}_0^{(H)} \sim \mu_H, \tag{38}$$

where

$$\mu_H = \mathcal{N}(\mathbf{0}, \mathbf{I}_H) \tag{39}$$

is the source distribution. Under the uniqueness assumption of Proposition 1, let $\Phi_H^t : \mathbb{R}^H \to \mathbb{R}^H$ denote the corresponding flow map.

**Source consistency.** The standard Gaussian source family is projectively consistent:

$$(\pi_h)_\# \mu_H = \mu_h, \tag{40}$$

since the first $h$ coordinates of an $H$-dimensional standard Gaussian are themselves standard Gaussian.

**Prefix closure and flow consistency.** For any supported $h < H$, the prefix-masked condition encoder leaves the first $h/P$ condition tokens unchanged when the requested horizon is extended from $h$ to $H$. Moreover, all state mixing in the velocity network is left-causal, so suffix state and condition tokens cannot affect the first $h/P$ velocity patches. Hence, for every $\mathbf{z} \in \mathbb{R}^H$ and $t \in [0, 1]$,

$$\pi_h \mathbf{v}_\theta^{(H)}(\mathbf{z}, t, \mathbf{x}_j, j) = \mathbf{v}_\theta^{(h)}(\pi_h \mathbf{z}, t, \mathbf{x}_j, j). \tag{41}$$

For any initial state $\mathbf{z}_0 \in \mathbb{R}^H$, define

$$\boldsymbol{\eta}_t = \pi_h \Phi_H^t(\mathbf{z}_0).$$

By Eq. (41),

$$\frac{d\boldsymbol{\eta}_t}{dt} = \mathbf{v}_\theta^{(h)}(\boldsymbol{\eta}_t, t, \mathbf{x}_j, j), \qquad \boldsymbol{\eta}_0 = \pi_h \mathbf{z}_0. \tag{42}$$

Thus, $\boldsymbol{\eta}_t$ satisfies the same initial-value problem as $\Phi_h^t(\pi_h \mathbf{z}_0)$. By uniqueness,

$$\pi_h \Phi_H^t(\mathbf{z}_0) = \Phi_h^t(\pi_h \mathbf{z}_0), \qquad t \in [0, 1], \tag{43}$$

and therefore the flow maps commute with prefix projection:

$$\pi_h \circ \Phi_H^t = \Phi_h^t \circ \pi_h. \tag{44}$$

**Channel-wise projective consistency.** The predictive distribution at horizon $H$ is

$$Q_H^{(j)}(\cdot \mid \mathbf{x}_j) = (\Phi_H^1)_\# \mu_H. \tag{45}$$

Using Eq. (44) and source consistency,

$$\begin{aligned}(\pi_h)_\# Q_H^{(j)}(\cdot \mid \mathbf{x}_j) &= (\pi_h \circ \Phi_H^1)_\# \mu_H \\ &= (\Phi_h^1 \circ \pi_h)_\# \mu_H \\ &= (\Phi_h^1)_\# (\pi_h)_\# \mu_H \\ &= (\Phi_h^1)_\# \mu_h = Q_h^{(j)}(\cdot \mid \mathbf{x}_j).\end{aligned} \tag{46}$$

Hence, for each channel $j$, PrefixFlow defines a projectively consistent family over all supported horizons.

**Multivariate extension.** Overloading $\pi_h$ to denote the temporal-prefix projection applied componentwise across channels, Eq. (46) implies

$$\begin{aligned}(\pi_h)_\# Q_H(\cdot \mid \mathbf{x}) &= \bigotimes_{j=1}^{D} (\pi_h)_\# Q_H^{(j)}(\cdot \mid \mathbf{x}_j) \\ &= \bigotimes_{j=1}^{D} Q_h^{(j)}(\cdot \mid \mathbf{x}_j) = Q_h(\cdot \mid \mathbf{x}).\end{aligned} \tag{47}$$

Therefore, the full multivariate product distribution induced by PrefixFlow is projectively consistent across all supported horizons.

**Consistency under discrete Euler inference.** The implementation uses a finite-step Euler solver rather than the exact continuous-time flow. Let $0 = t_0 < t_1 < \cdots < t_K = 1$ be a time grid shared across horizons, with $\Delta t_k = t_{k+1} - t_k$. The horizon-$H$ update is

$$\mathbf{z}_{k+1}^{(H)} = \mathbf{z}_k^{(H)} + \Delta t_k \mathbf{v}_\theta^{(H)} \left( \mathbf{z}_k^{(H)}, t_k, \mathbf{x}_j, j \right). \tag{48}$$

Couple the initial states by

$$\mathbf{z}_0^{(h)} = \pi_h \mathbf{z}_0^{(H)}.$$

If $\mathbf{z}_k^{(h)} = \pi_h \mathbf{z}_k^{(H)}$, then Eq. (41) gives

$$\begin{aligned} \pi_h \mathbf{z}_{k+1}^{(H)} &= \pi_h \mathbf{z}_k^{(H)} + \Delta t_k \pi_h \mathbf{v}_\theta^{(H)} \left( \mathbf{z}_k^{(H)}, t_k, \mathbf{x}_j, j \right) \\ &= \mathbf{z}_k^{(h)} + \Delta t_k \mathbf{v}_\theta^{(h)} \left( \mathbf{z}_k^{(h)}, t_k, \mathbf{x}_j, j \right) = \mathbf{z}_{k+1}^{(h)}. \end{aligned} \tag{49}$$

Therefore, by induction,

$$\mathbf{z}_k^{(h)} = \pi_h \mathbf{z}_k^{(H)}, \qquad k = 0, \ldots, K. \tag{50}$$

Thus, Euler inference preserves the same projective-consistency property whenever the same solver grid is used across horizons. The multivariate extension follows from the same channel-wise factorization above.

We use the same four-step non-uniform Euler grid for all horizons: $[0, 0.0625, 0.25, 0.5625, 1]$. The noise coupling above is used only to establish the pathwise result: we sample $z_0^{(H)} \sim \mathcal{N}(0, I_H)$ and define $z_0^{(h)} = \pi_h z_0^{(H)}$, which preserves the correct Gaussian marginal for the shorter horizon. In the empirical consistency analysis, Direct and Crop are instead represented by independently generated finite sample sets. Consequently, their empirical discrepancy need not be exactly zero even when the underlying predictive distributions are projectively consistent.

# D Experimental Details

## D.1 Datasets and Splits

We evaluate PrefixFlow on six widely used real-world multivariate time-series forecasting benchmarks: ETTh1, ETTm1, Weather, Exchange, ECL, and ILI.

**ETTh1 and ETTm1** are subsets of the Electricity Transformer Temperature (ETT) benchmark (Zhou et al., 2021). The ETT data contain electricity-transformer measurements collected from July 2016 to July 2018. Each subset contains seven variables: six load-related measurements and the oil temperature (OT). ETTh1 is recorded hourly, whereas ETTm1 is recorded every 15 minutes.

**Weather** contains 21 meteorological variables recorded every 10 minutes throughout 2020 at the Weather Station of the Max Planck Institute for Biogeochemistry.

**Exchange** contains daily exchange rates from eight countries between 1990 and 2016 (Lai et al., 2018).

**ECL** contains hourly electricity-consumption measurements from 321 clients between 2012 and 2014.

**ILI** contains weekly influenza-like illness surveillance records from the U.S. Centers for Disease Control and Prevention between 2002 and 2021. The benchmark contains seven variables describing ILI prevalence, age-group counts, the total number of ILI cases, the number of reporting providers, and the total number of patients.

We follow the standard chronological data splits used in prior long-term forecasting benchmarks (Wu et al., 2021). The ETT datasets use a 6:2:2 train/validation/test split, corresponding to the standard 12/4/4-month protocol, while Weather, Exchange, ECL, and ILI use a 7:1:2 split. For ETTh1, ETTm1, Weather, Exchange, and ECL, we use a history length of 336 and evaluate $H \in \{96, 192, 336, 720\}$. For ILI, we use a history length of 36 and $H \in \{24, 36, 48, 60\}$.

Table 6 summarizes the dataset statistics and forecasting configurations.

Table 6: Dataset statistics and forecasting configurations.

| Dataset | Length | Variables | Frequency | History | Horizons | Split |
|---|---|---|---|---|---|---|
| ETTh1 | 17,420 | 7 | 1 hour | 336 | {96, 192, 336, 720} | 6:2:2 |
| ETTm1 | 69,680 | 7 | 15 min | 336 | {96, 192, 336, 720} | 6:2:2 |
| Weather | 52,696 | 21 | 10 min | 336 | {96, 192, 336, 720} | 7:1:2 |
| Exchange | 7,588 | 8 | 1 day | 336 | {96, 192, 336, 720} | 7:1:2 |
| ECL | 26,304 | 321 | 1 hour | 336 | {96, 192, 336, 720} | 7:1:2 |
| ILI | 966 | 7 | 1 week | 36 | {24, 36, 48, 60} | 7:1:2 |

### D.2 Evaluation Metrics

We evaluate point forecasting accuracy using mean squared error (MSE) and mean absolute error (MAE), and probabilistic forecasting quality using the Continuous Ranked Probability Score (CRPS) and Quantile Interval Coverage Error (QICE). For each test instance, we generate $S = 100$ predictive trajectories, and use their sample mean as the point forecast for computing MSE and MAE.

CRPS evaluates the predictive distribution rather than only its mean (Gneiting & Raftery, 2007). Given predictive samples $\{\widehat{y}^{(s)}\}_{s=1}^{S}$ and observation $y$, we use the empirical estimator

$$\widehat{\mathrm{CRPS}} = \frac{1}{S}\sum_{s=1}^{S}\left|\widehat{y}^{(s)} - y\right| - \frac{1}{2S^2}\sum_{s=1}^{S}\sum_{r=1}^{S}\left|\widehat{y}^{(s)} - \widehat{y}^{(r)}\right|. \tag{51}$$

The reported CRPS is averaged over all test windows, forecast steps, and variables.

QICE measures predictive calibration by comparing the empirical occupancy of equal-probability quantile intervals with their nominal proportions (Han et al., 2022). We pool all test windows, forecast steps, and variables, yielding $M = NHD$ scalar forecast points, and divide each predictive distribution into $K = 10$ quantile intervals. Let $\widehat{q}_{m,k}$ denote the empirical $k/K$ quantile of the $S = 100$ predictive samples for point $m$. Following our implementation, the interval index of observation $y_m$ is

$$\widetilde{b}_m = \min\left\{K,\ \max\left\{1,\ \sum_{k=0}^{K}\mathbf{1}[y_m > \widehat{q}_{m,k}]\right\}\right\}. \tag{52}$$

Let

$$\widehat{p}_j = \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\left[\widetilde{b}_m = j\right], \qquad j = 1, \ldots, K. \tag{53}$$

Then

$$\mathrm{QICE} = \frac{1}{K}\sum_{j=1}^{K}\left|\widehat{p}_j - \frac{1}{K}\right|. \tag{54}$$

Lower QICE indicates better calibration, with $\mathrm{QICE} = 0$ corresponding to exact occupancy of the nominal quantile intervals.

### D.3 Implementation and Evaluation Details

PrefixFlow uses a patch length of $P = 12$, hidden dimension $d = 128$, and four prefix-causal velocity blocks with a causal Conv1D kernel size of 3. The condition encoder contains two Transformer layers with four attention heads and a feed-forward dimension of 256. The time embedding dimension is 128, and the channel embedding dimension is 16. Each variable is standardized using statistics from the training split, followed by a window-level normalization based on the observed history.

We optimize PrefixFlow with AdamW (Loshchilov & Hutter, 2019) with cosine learning-rate annealing and gradient clipping at 1.0. At each optimization step, one mini-batch is sampled for each training horizon, with independent Gaussian source noise and flow time for each horizon. The Flow Matching loss uses the coverage-aware weighting introduced in Sec. 4.4.

Table 7 reports the dataset-specific training hyperparameters used in our experiments. For ETTh1, ETTm1, Weather, Exchange, and ECL, joint training uses the horizon set $\{96, 192, 336, 720\}$; ILI uses $\{24, 36, 48, 60\}$. Checkpoints are selected according to the mean validation CRPS across the training horizons.

Table 7: Dataset-specific training configurations for PrefixFlow.

| Dataset | Batch | Learning Rate | Weight Decay | Dropout | Max Epochs | Patience |
|---|---|---|---|---|---|---|
| ETTh1 | 128 | $1 \times 10^{-4}$ | $1 \times 10^{-4}$ | 0.05 | 20 | 3 |
| ETTm1 | 128 | $1 \times 10^{-4}$ | $1 \times 10^{-4}$ | 0.05 | 20 | 3 |
| Weather | 128 | $1 \times 10^{-4}$ | $1 \times 10^{-4}$ | 0.05 | 20 | 5 |
| Exchange | 128 | $6 \times 10^{-5}$ | $1 \times 10^{-3}$ | 0.25 | 20 | 4 |
| ECL | 512 | $1 \times 10^{-4}$ | $1 \times 10^{-4}$ | 0.05 | 20 | 4 |
| ILI | 64 | $1 \times 10^{-4}$ | $1 \times 10^{-4}$ | 0.05 | 40 | 6 |

At inference time, PrefixFlow generates forecasts directly at the requested horizon, without first generating the maximum horizon $H_{\max}$ and then cropping. We use the same four-step non-uniform Euler solver for validation and testing, with the time grid

$$t_k = \left(\frac{k}{4}\right)^2, \qquad [t_0, \ldots, t_4] = [0, 0.0625, 0.25, 0.5625, 1]. \tag{55}$$

Accordingly, all forecasts are generated with four neural function evaluations (NFE $= 4$).

The main forecasting evaluation uses 100 predictive samples per test instance, whereas the consistency analyses use 256 samples per predictive distribution. We use a fixed random seed of 2026 for PrefixFlow and all baseline methods throughout the experiments.

All experiments are conducted using PyTorch on NVIDIA GeForce RTX 4090 GPUs with 24 GB of memory. For the efficiency evaluation, we measure batched inference latency and peak GPU memory on a single RTX 4090 with a batch size of 128, using 100 predictive samples and NFE $= 4$. Latency is averaged over ten runs after three warm-up runs.

### D.4 Baseline Protocols

We reproduce the results of TimeDiff, D3U, NsDiff, TSFlow, and TEDM under a common experimental protocol. For each baseline, a separate model is trained for each forecast horizon. All methods use the same chronological train/validation/test splits, history lengths and random seed of 2026. Reported metrics are computed in the same train-standardized evaluation space across all methods.

For each baseline, we follow the corresponding model formulation, training objective, and method-specific sampling procedure, with hyperparameter settings chosen according to the corresponding paper or available implementation whenever applicable. The same test protocol is used for all methods to ensure that the reported forecasting results are directly comparable.

### D.5 Ablation Settings

To isolate the effects of prefix-invariant conditioning, the prefix-causal velocity field, and shared multi-horizon training, we compare the full model with three architectural ablations and a separate-training variant. Table 8 summarizes their exact configurations.

Table 8: Definitions of the ablation variants.

| Variant | Condition Path | State Path | Training |
|---|---|---|---|
| Full | Prefix Mask | Causal Conv (kernel 3) | Joint (96, 192, 336, 720) |
| w/o Prefix Mask | Bidirectional Attention | Causal Conv (kernel 3) | Joint (96, 192, 336, 720) |
| w/o Causal Conv | Prefix Mask | Symmetric Conv (kernel 3) | Joint (96, 192, 336, 720) |
| w/o Both | Bidirectional Attention | Symmetric Conv (kernel 3) | Joint (96, 192, 336, 720) |
| Separate Training | Bidirectional Attention | Symmetric Conv (kernel 3) | Separate (96 / 720) |

The *w/o Prefix Mask* variant replaces prefix-masked conditioning with bidirectional attention while retaining the causal state path. The *w/o Causal Conv* variant retains prefix-masked conditioning but replaces the causal convolution with a symmetric non-causal convolution of the same kernel size. The

*w/o Both* variant removes both prefix-causal constraints while retaining joint multi-horizon training. These three variants, together with the Full model, are trained jointly on $H \in \{96, 192, 336, 720\}$.

The *Separate Training* baseline uses the same non-causal condition and state paths as *w/o Both*, but trains the $H = 96$ and $H = 720$ models independently. Therefore, this comparison removes both shared parameterization and joint multi-horizon supervision and should not be interpreted as an ablation of joint training alone.

# E Full Forecasting Results

Table 9: Deterministic forecasting results on all six datasets. Lower values are better. Best and second-best results are shown in **bold** and underlined, respectively.

| Dataset | $H$ | MSE | | | | | | MAE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | TimeDiff | D3U | NsDiff | TSFlow | TEDM | **PrefixFlow** | TimeDiff | D3U | NsDiff | TSFlow | TEDM | **PrefixFlow** |
| ETTh1 | 96 | 0.3837 | 0.4036 | 0.5453 | 0.4262 | 0.6749 | **0.3760** | 0.4046 | 0.4219 | 0.5062 | 0.4464 | 0.5299 | **0.4023** |
| | 192 | 0.4220 | 0.4629 | 0.5288 | 0.5081 | 0.7006 | **0.4128** | 0.4322 | 0.4588 | 0.4995 | 0.4991 | 0.5434 | **0.4242** |
| | 336 | 0.4524 | 0.4748 | 0.5477 | 0.5668 | 0.7128 | **0.4336** | 0.4523 | 0.4700 | 0.5104 | 0.5440 | 0.5556 | **0.4359** |
| | 720 | 0.4572 | 0.4949 | 0.7542 | 0.6359 | 0.9574 | **0.4556** | 0.4671 | 0.5042 | 0.6255 | 0.6108 | 0.6561 | **0.4601** |
| | **Avg.** | 0.4289 | 0.4590 | 0.5940 | 0.5343 | 0.7614 | **0.4195** | 0.4390 | 0.4637 | 0.5354 | 0.5251 | 0.5713 | **0.4306** |
| ETTm1 | 96 | 0.3056 | 0.2953 | 0.3993 | 0.7088 | 0.7068 | **0.2789** | 0.3523 | 0.3551 | 0.4175 | 0.6165 | 0.5321 | **0.3312** |
| | 192 | 0.3445 | 0.3350 | 0.4017 | 0.6830 | 0.7377 | **0.3138** | 0.3727 | 0.3723 | 0.4190 | 0.6073 | 0.5482 | **0.3551** |
| | 336 | 0.3875 | 0.3924 | 0.4923 | 0.7190 | 0.7551 | **0.3454** | 0.4096 | 0.4140 | 0.4597 | 0.6081 | 0.5588 | **0.3759** |
| | 720 | 0.4554 | 0.4334 | 0.5427 | 0.6983 | 1.0586 | **0.4057** | 0.4519 | 0.4259 | 0.4988 | 0.5636 | 0.6586 | **0.4106** |
| | **Avg.** | 0.3733 | 0.3640 | 0.4590 | 0.7023 | 0.8146 | **0.3359** | 0.3966 | 0.3918 | 0.4487 | 0.5989 | 0.5744 | **0.3682** |
| Weather | 96 | 0.1858 | 0.1683 | 0.1863 | 0.1935 | 0.2340 | **0.1458** | 0.2404 | 0.2457 | 0.2411 | 0.2528 | 0.2576 | **0.2029** |
| | 192 | 0.2478 | 0.2112 | 0.2522 | 0.2539 | 0.2781 | **0.1895** | 0.2995 | 0.2753 | 0.2947 | 0.3030 | 0.2949 | **0.2478** |
| | 336 | 0.2959 | 0.3004 | 0.2944 | 0.2952 | 0.3360 | **0.2418** | 0.3279 | 0.3414 | 0.3268 | 0.3365 | 0.3368 | **0.2895** |
| | 720 | 0.3508 | 0.3341 | 0.3863 | 0.4269 | 0.4058 | **0.3184** | 0.3705 | 0.3468 | 0.3773 | 0.3846 | 0.3857 | **0.3438** |
| | **Avg.** | 0.2701 | 0.2535 | 0.2798 | 0.2924 | 0.3135 | **0.2239** | 0.3096 | 0.3023 | 0.3100 | 0.3192 | 0.3188 | **0.2710** |
| Exchange | 96 | 0.1837 | 0.1587 | 0.1598 | 0.2195 | **0.0839** | 0.0877 | 0.3126 | 0.2772 | 0.2914 | 0.3526 | **0.1984** | 0.2045 |
| | 192 | 0.3550 | 0.2234 | 0.2660 | 0.3159 | **0.1808** | 0.1844 | 0.4413 | 0.3336 | 0.3957 | 0.4305 | **0.3007** | 0.3030 |
| | 336 | 0.6150 | 1.0324 | 0.7328 | 0.4073 | **0.3421** | 0.3497 | 0.6009 | 0.8227 | 0.6590 | 0.4920 | **0.4254** | 0.4268 |
| | 720 | 0.8563 | 1.0460 | 1.9245 | **0.7469** | 0.9294 | 1.0191 | **0.7085** | 0.7770 | 0.9914 | 0.7179 | 0.7283 | 0.7635 |
| | **Avg.** | 0.5025 | 0.6151 | 0.7708 | 0.4224 | **0.3841** | 0.4102 | 0.5158 | 0.5526 | 0.5844 | 0.4982 | **0.4132** | 0.4245 |
| ECL | 96 | 0.1780 | **0.1332** | 0.2028 | 0.3795 | 0.8232 | 0.1333 | 0.2907 | 0.2321 | 0.3056 | 0.3833 | 0.7462 | **0.2249** |
| | 192 | 0.1860 | **0.1489** | 0.2023 | 0.5402 | 0.8407 | 0.1494 | 0.2969 | 0.2464 | 0.3063 | 0.4886 | 0.7556 | **0.2396** |
| | 336 | 0.2009 | 0.1723 | 0.2324 | 0.2304 | 0.8552 | **0.1661** | 0.3093 | 0.2666 | 0.3323 | 0.3269 | 0.7618 | **0.2566** |
| | 720 | 0.2575 | **0.2024** | 0.2416 | 0.3094 | 0.8827 | 0.2043 | 0.3617 | 0.2927 | 0.3330 | 0.3259 | 0.7716 | **0.2888** |
| | **Avg.** | 0.2056 | 0.1642 | 0.2198 | 0.3649 | 0.8504 | **0.1633** | 0.3146 | 0.2595 | 0.3193 | 0.3812 | 0.7588 | **0.2525** |
| ILI | 24 | 4.1036 | **1.8638** | 3.0314 | 4.7734 | 4.6114 | 2.2267 | 1.4344 | **0.8228** | 1.1601 | 1.4886 | 1.4175 | 0.8320 |
| | 36 | 3.3855 | 2.5896 | 3.4385 | 4.7966 | 5.2729 | **2.1269** | 1.2869 | 1.0248 | 1.2059 | 1.4966 | 1.5861 | **0.8351** |
| | 48 | 3.0984 | 2.6959 | 2.4854 | 4.2851 | 5.2764 | **1.8874** | 1.2054 | 1.0514 | 1.0216 | 1.4009 | 1.6117 | **0.8027** |
| | 60 | 3.4560 | 2.8882 | 2.4614 | 4.6417 | 4.7275 | **1.7653** | 1.2617 | 1.0901 | 1.0539 | 1.4252 | 1.5141 | **0.8052** |
| | **Avg.** | 3.5109 | 2.5094 | 2.8542 | 4.6242 | 4.9721 | **2.0016** | 1.2971 | 0.9973 | 1.1104 | 1.4528 | 1.5323 | **0.8187** |

Table 10: Probabilistic forecasting results on all six datasets. Lower values are better. Best and second-best results are shown in **bold** and underlined, respectively.

| Dataset | $H$ | CRPS | | | | | | QICE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | TimeDiff | D3U | NsDiff | TSFlow | TEDM | **PrefixFlow** | TimeDiff | D3U | NsDiff | TSFlow | TEDM | **PrefixFlow** |
| ETTh1 | 96 | 0.3852 | 0.3206 | 0.3801 | 0.3468 | 0.3748 | **0.2972** | 0.1438 | **0.0112** | 0.0130 | 0.0520 | 0.0130 | 0.0353 |
| | 192 | 0.4069 | 0.3491 | 0.3772 | 0.3899 | 0.3934 | **0.3145** | 0.1395 | 0.0325 | **0.0128** | 0.0588 | 0.0195 | 0.0344 |
| | 336 | 0.4230 | 0.3518 | 0.3848 | 0.4206 | 0.4067 | **0.3235** | 0.1378 | 0.0169 | **0.0122** | 0.0615 | 0.0221 | 0.0323 |
| | 720 | 0.4355 | 0.3702 | 0.4539 | 0.4510 | 0.5438 | **0.3375** | 0.1378 | 0.0183 | **0.0126** | 0.0552 | 0.0787 | 0.0273 |
| | **Avg.** | 0.4126 | 0.3479 | 0.3990 | 0.4021 | 0.4297 | **0.3182** | 0.1397 | 0.0197 | **0.0126** | 0.0569 | 0.0333 | 0.0323 |
| ETTm1 | 96 | 0.3249 | 0.2615 | 0.3134 | 0.5057 | 0.3771 | **0.2468** | 0.1316 | 0.0300 | **0.0125** | 0.0961 | 0.0216 | 0.0481 |
| | 192 | 0.3407 | 0.2726 | 0.3175 | 0.4846 | 0.3985 | **0.2651** | 0.1278 | **0.0170** | 0.0290 | 0.0805 | 0.0272 | 0.0489 |
| | 336 | 0.3805 | 0.2995 | 0.3494 | 0.4876 | 0.4132 | **0.2810** | 0.1351 | **0.0203** | 0.0290 | 0.0675 | 0.0305 | 0.0485 |
| | 720 | 0.4246 | 0.3099 | 0.3803 | 0.4419 | 0.5626 | **0.3073** | 0.1412 | **0.0054** | 0.0380 | 0.0521 | 0.0876 | 0.0478 |
| | **Avg.** | 0.3677 | 0.2859 | 0.3401 | 0.4799 | 0.4379 | **0.2751** | 0.1339 | **0.0182** | 0.0271 | 0.0741 | 0.0417 | 0.0483 |
| Weather | 96 | 0.2231 | 0.1770 | 0.2004 | 0.1956 | 0.2177 | **0.1430** | 0.1220 | 0.0404 | 0.0422 | **0.0267** | 0.1078 | 0.0400 |
| | 192 | 0.2859 | 0.1993 | 0.2409 | 0.2322 | 0.2490 | **0.1760** | 0.1296 | **0.0107** | 0.0356 | 0.0397 | 0.1086 | 0.0445 |
| | 336 | 0.3113 | 0.2615 | 0.2684 | 0.2598 | 0.2855 | **0.2076** | 0.1287 | 0.0619 | **0.0352** | 0.0412 | 0.1098 | 0.0479 |
| | 720 | 0.3480 | 0.2644 | 0.3081 | 0.3140 | 0.3316 | **0.2488** | 0.1349 | **0.0217** | 0.0480 | 0.0882 | 0.1138 | 0.0528 |
| | **Avg.** | 0.2921 | 0.2256 | 0.2544 | 0.2504 | 0.2710 | **0.1938** | 0.1288 | **0.0337** | 0.0403 | 0.0490 | 0.1100 | 0.0463 |
| Exchange | 96 | 0.2937 | 0.2124 | 0.2330 | 0.2744 | 0.1551 | **0.1506** | 0.1378 | 0.0296 | 0.0412 | 0.0575 | 0.0581 | **0.0233** |
| | 192 | 0.4273 | 0.2514 | 0.3132 | 0.3280 | 0.2517 | **0.2221** | 0.1473 | 0.0322 | 0.0685 | 0.0473 | 0.0875 | **0.0253** |
| | 336 | 0.5869 | 0.6750 | 0.5622 | 0.3713 | 0.3871 | **0.3140** | 0.1501 | 0.1288 | 0.0928 | 0.0388 | 0.1234 | **0.0367** |
| | 720 | 0.6926 | 0.7045 | 0.8899 | 0.6924 | 0.6815 | **0.5819** | 0.1491 | 0.1294 | 0.1059 | 0.1529 | 0.1365 | **0.0717** |
| | **Avg.** | 0.5001 | 0.4608 | 0.4996 | 0.4165 | 0.3689 | **0.3172** | 0.1461 | 0.0800 | 0.0771 | 0.0741 | 0.1014 | **0.0393** |
| ECL | 96 | 0.2772 | **0.1723** | 0.2908 | 0.3213 | 0.6073 | 0.1762 | 0.1425 | **0.0089** | 0.0757 | 0.1213 | 0.0995 | 0.0574 |
| | 192 | 0.2837 | **0.1825** | 0.2940 | 0.4215 | 0.6568 | 0.1879 | 0.1433 | **0.0073** | 0.0771 | 0.1274 | 0.1255 | 0.0560 |
| | 336 | 0.2961 | 0.2016 | 0.2810 | 0.2759 | 0.6523 | **0.2009** | 0.1438 | **0.0274** | 0.0567 | 0.0892 | 0.1206 | 0.0557 |
| | 720 | 0.3483 | **0.2174** | 0.2865 | 0.2646 | 0.6469 | 0.2256 | 0.1469 | **0.0128** | 0.0604 | 0.0793 | 0.1118 | 0.0552 |
| | **Avg.** | 0.3013 | **0.1935** | 0.2881 | 0.3208 | 0.6408 | 0.1977 | 0.1441 | **0.0141** | 0.0675 | 0.1043 | 0.1144 | 0.0561 |
| ILI | 24 | 1.4188 | 0.7228 | 0.9783 | 1.4049 | 1.2235 | **0.6126** | 0.1568 | 0.1039 | 0.0611 | 0.1332 | 0.0992 | **0.0332** |
| | 36 | 1.2717 | 0.8992 | 1.0068 | 1.4001 | 1.3772 | **0.6297** | 0.1566 | 0.1011 | 0.0619 | 0.1283 | 0.1077 | **0.0375** |
| | 48 | 1.1891 | 0.9533 | 0.8479 | 1.2754 | 1.3904 | **0.6158** | 0.1559 | 0.1195 | 0.0481 | 0.1191 | 0.1102 | **0.0405** |
| | 60 | 1.2472 | 0.9805 | 0.8599 | 1.3066 | 1.3085 | **0.6223** | 0.1569 | 0.1143 | 0.0540 | 0.1274 | 0.1079 | **0.0410** |
| | **Avg.** | 1.2817 | 0.8890 | 0.9232 | 1.3468 | 1.3249 | **0.6201** | 0.1565 | 0.1097 | 0.0562 | 0.1270 | 0.1062 | **0.0381** |

# F Additional Experimental Results

## F.1 Complete Cross-Horizon Consistency Diagnostics

Table 11 reports the complete cross-horizon consistency diagnostics for PrefixFlow. Across all nine comparisons, the 95% confidence intervals of Excess ED contain zero, and the Energy Ratio remains close to one, indicating no detectable joint-distribution discrepancy beyond finite-sample variation.

Exchange exhibits noticeably higher Direction Flip Rates than ETTh1 and Weather. This metric is sensitive to finite-sample variation because it thresholds an empirically estimated event probability at $0.5$ and, unlike Excess ED, does not include a self-discrepancy correction. In particular, approximately 49–55% of the evaluated Exchange window–variable cases have at least one of the two estimated event probabilities within $[0.45, 0.55]$, where small Monte Carlo fluctuations can readily change the resulting binary decision.

We therefore quantify the flip rate expected from finite sampling alone under a common-distribution null. For a window–variable case $(i, c)$, let $p_{i,c}$ denote the common underlying probability of the directional event under Direct and Crop. With $S = 256$ predictive trajectories, two independent empirical event counts under this null satisfy

$$K^D_{i,c}, K^C_{i,c} \overset{\text{i.i.d.}}{\sim} \text{Binomial}(256, p_{i,c}). \tag{56}$$

Because the direction decision is positive when the estimated event probability exceeds $0.5$, this is equivalent to $K_{i,c} \geq 129$. Define

$$q(p) = \Pr[\text{Binomial}(256, p) \geq 129]\,. \tag{57}$$

The probability that two independent 256-sample estimates produce opposite direction decisions under the same underlying event probability is then

$$r(p) = 2q(p)\big(1 - q(p)\big). \tag{58}$$

For each window–variable case, we estimate the common event probability by pooling the Direct and Crop event samples,

$$\hat{p}_{i,c} = \frac{K^D_{i,c} + K^C_{i,c}}{2S}, \qquad S = 256. \tag{59}$$

We then evaluate $r(\hat{p}_{i,c})$ for each Exchange window–variable case and aggregate the resulting values using the same macro-averaging scheme as the observed Direction Flip Rate.

The resulting null flip rates are 12.87%, 14.97%, and 8.04% for 96–192, 96–720, and 336–720, respectively, whereas the observed rates are 11.82%, 12.88%, and 7.38%. Thus, the observed Exchange flip rates do not exceed the level expected from two independent 256-sample estimates of the same underlying event probability. Together with the Excess ED confidence intervals containing zero and the Energy Ratios remaining close to one, this indicates that the larger raw Direction Flip Rates on Exchange are consistent with finite-sample threshold variability rather than evidence of systematic cross-horizon inconsistency.

Table 11: Complete cross-horizon consistency diagnostics for PrefixFlow. Bracketed values denote 95% moving-block bootstrap confidence intervals. Ratio denotes the aggregate Cross/Self Energy Ratio, and Prob. Gap denotes the Directional Event Probability Gap.

| Dataset | Pair | Excess ED | Cross > Self | Ratio | Quantile Gap | Prob. Gap | Direction Flip |
|---|---|---|---|---|---|---|---|
| ETTh1 | 96–192 | 0.00098 [-0.00114, 0.00220] | 0.534 [0.466, 0.573] | 1.004 | 0.01384 [0.01285, 0.01485] | 0.01943 [0.01787, 0.02149] | 0.0178 [0.0059, 0.0267] |
| | 96–720 | 0.00132 [-0.00033, 0.00332] | 0.513 [0.465, 0.550] | 1.005 | 0.01361 [0.01245, 0.01450] | 0.01826 [0.01498, 0.02063] | 0.0111 [0.0037, 0.0221] |
| | 336–720 | 0.00072 [-0.00123, 0.00295] | 0.502 [0.450, 0.542] | 1.001 | 0.01524 [0.01389, 0.01628] | 0.01345 [0.01042, 0.01490] | 0.0074 [0, 0.0148] |
| Weather | 96–192 | -0.00022 [-0.00120, 0.00098] | 0.501 [0.472, 0.533] | 0.999 | 0.02018 [0.01961, 0.02077] | 0.01700 [0.01618, 0.01776] | 0.0142 [0.0120, 0.0161] |
| | 96–720 | -0.00011 [-0.00104, 0.00090] | 0.506 [0.483, 0.536] | 1.000 | 0.02017 [0.01948, 0.02086] | 0.01681 [0.01608, 0.01752] | 0.0137 [0.0117, 0.0157] |
| | 336–720 | -0.00002 [-0.00118, 0.00125] | 0.502 [0.476, 0.531] | 1.000 | 0.02558 [0.02498, 0.02609] | 0.01425 [0.01350, 0.01505] | 0.0111 [0.0096, 0.0125] |
| Exchange | 96–192 | 0.00100 [-0.00138, 0.00367] | 0.494 [0.452, 0.548] | 1.006 | 0.02883 [0.02857, 0.02939] | 0.03504 [0.03360, 0.03647] | 0.1182 [0.1047, 0.1348] |
| | 96–720 | 0.00026 [-0.00328, 0.00468] | 0.490 [0.410, 0.580] | 1.002 | 0.02861 [0.02809, 0.02916] | 0.03347 [0.03225, 0.03491] | 0.1288 [0.0788, 0.1788] |
| | 336–720 | 0.00728 [-0.00070, 0.00997] | 0.630 [0.560, 0.660] | 1.015 | 0.04638 [0.04585, 0.04681] | 0.03478 [0.03426, 0.03548] | 0.0738 [0.0363, 0.1263] |

Table 12: Forecasting performance at unseen prediction horizons. The model is trained only on $H \in \{96, 192, 336, 720\}$. Lower values are better.

| | ETTh1 | | | Weather | | |
|---|---|---|---|---|---|---|
| $H$ | MSE | MAE | CRPS | MSE | MAE | CRPS |
| 120 | 0.3877 | 0.4091 | 0.3026 | 0.1579 | 0.2168 | 0.1532 |
| 240 | 0.4239 | 0.4303 | 0.3193 | 0.2101 | 0.2653 | 0.1893 |
| 480 | 0.4361 | 0.4416 | 0.3262 | 0.2784 | 0.3153 | 0.2272 |
| 600 | 0.4368 | 0.4494 | 0.3300 | 0.3008 | 0.3311 | 0.2393 |
| Avg. | 0.4211 | 0.4326 | 0.3195 | 0.2368 | 0.2822 | 0.2022 |

### F.2 Generalization to Unseen Horizons

We further evaluate PrefixFlow at patch-aligned prediction horizons that are not included during training. Table 12 reports the results on ETTh1 and Weather for $H \in \{120, 240, 480, 600\}$.

Despite being trained on only four horizons, PrefixFlow directly produces forecasts at unseen patch-aligned horizons within the supported range without retraining.

### F.3 Consistency at Unseen Horizons

We finally examine whether cross-horizon consistency is preserved between prediction horizons that are not used during training. Table 13 reports the results on ETTh1.

Table 13: Cross-horizon consistency at unseen prediction horizons on ETTh1. Bracketed values denote 95% moving-block bootstrap confidence intervals.

| Pair | Excess ED | Cross $>$ Self | Ratio | Quantile Gap | Prob. Gap | Direction Flip |
|---|---|---|---|---|---|---|
| 120–240 | -0.00019 [-0.00186, 0.00184] | 0.5045 [0.4591, 0.5680] | 0.99934 | 0.01428 [0.01321, 0.01519] | 0.01817 [0.01584, 0.02070] | 0.02719 [0.01511, 0.04532] |
| 240–600 | -0.00111 [-0.00268, 0.00158] | 0.4895 [0.4510, 0.5559] | 0.99736 | 0.01486 [0.01372, 0.01582] | 0.01337 [0.01222, 0.01554] | 0.00699 [0, 0.01399] |
| 480–600 | -0.00038 [-0.00269, 0.00242] | 0.4965 [0.4476, 0.5700] | 0.99939 | 0.01606 [0.01467, 0.01722] | 0.00994 [0.00911, 0.01149] | 0.00350 [0, 0.00350] |

The Excess ED confidence intervals again contain zero for all three unseen horizon pairs, with Cross $>$ Self close to $0.5$ and the Energy Ratio close to $1.0$. Thus, the cross-horizon consistency of PrefixFlow extends beyond the prediction horizons used during training.

## G VISUALIZATION OF CROSS-HORIZON CONSISTENCY

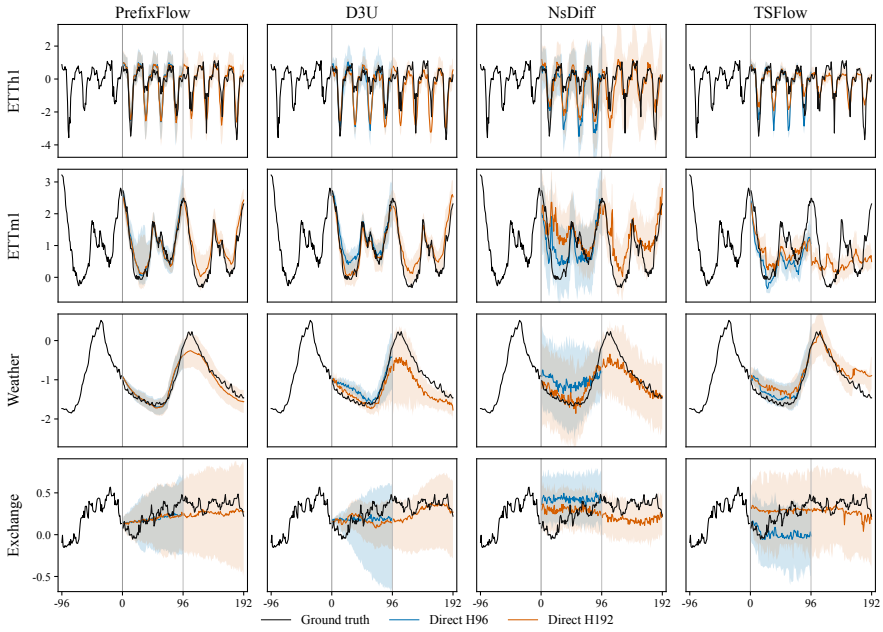


Figure 4: Cross-horizon forecast visualization for PrefixFlow, D3U, NsDiff, and TSFlow on ETTh1, ETTm1, Weather, and Exchange. All forecasts condition on a 336-step observed history; for visual clarity, only the last 96 historical steps are shown. Direct forecasts at $H = 96$ and $H = 192$ are generated from the same history and overlap on the first 96 future steps. Colored solid lines denote predictive medians, and shaded regions denote the 10–90% predictive quantile intervals.