CADRE: Distributional Refinement for Probabilistic Time Series Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Existing probabilistic time series forecasters often map temporal observations to predictive distributions through predefined parametric heads, such as Student-t and Gaussian. However, these distributional assumptions can limit their ability to capture asymmetric and irregular conditional distributions in real time series. To address this challenge, we introduce CAlibration via Distributional REfinement (CADRE), a framework that uses a base forecast to guide a more flexible predictive distribution. CADRE constructs input-adaptive bins from the forecaster's predicted quantiles and uses a lightweight refinement head to learn each bin's probability mass from shared temporal features. Experiments on six real world datasets demonstrate consistent improvements in probabilistic forecasting and gains in point forecast accuracy. Further experiments show that CADRE improves forecasting performance when transferring models across datasets. A channel-level analysis on Weather reveals an association between greater marginal skewness and larger probabilistic gains, suggesting that CADRE is particularly beneficial for forecasting asymmetric time series.

1 Introduction

Probabilistic time series forecasting requires predictive distributions that capture how uncertainty changes with the observed history and forecast horizon (Gneiting & Katzfuss, 2014). Their quality depends on both the temporal representation and the distributional head used to express uncertainty. Gaussian and Student-t heads restrict each forecast to a symmetric family (Salinas et al., 2020), which can limit their ability to capture asymmetric conditional distributions. Finite quantile heads allow more varied shapes, but constructing a complete distribution requires interpolation between predicted quantiles and extrapolation beyond them (Wen et al., 2017; Park et al., 2022). Increasing backbone capacity alone does not remove these restrictions. Flexible quantile functions offer richer representations (Gasthaus et al., 2019; Park et al., 2022; Tagasovska & Lopez-Paz, 2019; Bremnes, 2020). Distributional refinement takes a complementary approach: it uses an existing forecast's estimates of location and scale to guide a more flexible predictive distribution.

Existing methods refine or recalibrate predictive distributions after fitting (Avanzi et al., 2026; Dey et al., 2025; Kuleshov et al., 2018; Song et al., 2019; Kuleshov & Deshpande, 2022; Rumack et al., 2022) or during training (Dheur & Ben Taieb, 2024; 2023; Chung et al., 2021). For refinement based on bins, a central design choice is where to place their boundaries. The same deviation from a predicted value can be common during a highly variable period but rare during a stable one. A fixed partition shared across forecasts may therefore provide too little resolution near likely outcomes and devote too many bins to unlikely regions. Quantiles of the current predictive distribution provide a natural way to adapt this partition: equally spaced probability levels produce narrower bins where the forecast concentrates mass and wider bins elsewhere. Learning probability masses over these bins allows corrections to follow each forecast's conditional uncertainty.

We propose CAlibration via Distributional REfinement (CADRE), a framework for multivariate, multi-horizon probabilistic time series forecasting. A base distributional head produces a predictive distribution whose quantiles define bins for each input, channel, and forecast horizon. A lightweight refinement head uses the shared temporal features to learn probability masses across all bins, including the two unbounded tails. Finite bins use uniform densities, while each tail follows the base distribution conditioned on that interval. This construction yields a predictive distribution over the entire real line

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.

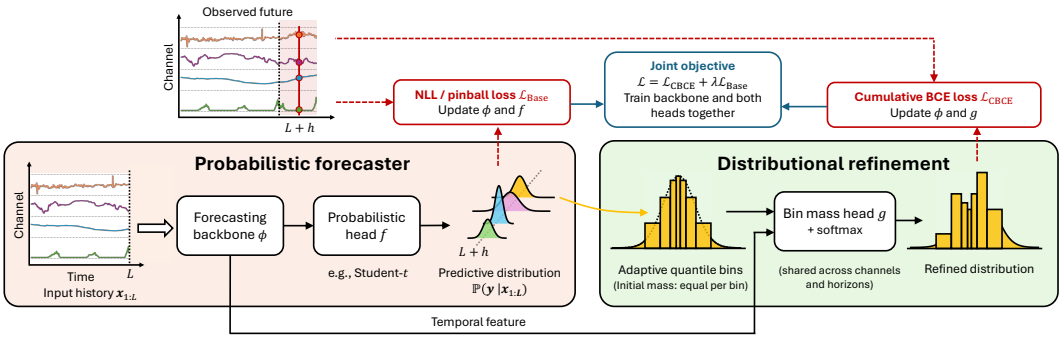

Figure 1: Overview of CADRE. The backbone ϕ extracts temporal features, and the base probabilistic head f produces a predictive distribution whose quantiles define adaptive bins for each channel and horizon. The refinement head g uses the shared features to learn bin probabilities, yielding a flexible predictive distribution. After pretraining the base forecaster, the backbone and both heads are jointly optimized: the base loss (negative log-likelihood (NLL) or pinball loss, depending on the head) updates ϕ and f, while cumulative binary cross-entropy (CBCE) updates ϕ and g. Further details are provided in Section 4.

and allows refinement to change both its interior shape and the relative probabilities of central and extreme outcomes.

After pretraining the base forecaster, CADRE jointly optimizes the backbone and both heads. The base head retains its negative log-likelihood or pinball objective, while cumulative binary cross-entropy (CBCE) supervises the refined cumulative probabilities (Li et al., 2021). This refinement objective accounts for the ordering of the bins by learning the probability that an observation falls at or below each boundary. Both objectives update the shared temporal representation, and the quantile boundaries evolve as the base forecast changes during training. Inference requires one shared backbone pass. Figure 1 summarizes the framework.

Our main contributions are as follows:

- We introduce input-adaptive quantile partitions derived from the base predictive distribution. CADRE learns the probability masses of all bins, including both unbounded tails, enabling corrections to tail probabilities while preserving the base distribution's tail decay.
- We develop a framework (Figure 1) that jointly trains the temporal backbone ϕ, base distributional head f, and lightweight refinement head g. Joint optimization with the base forecasting objective (e.g., NLL or pinball loss) and CBCE allows the temporal features, bin boundaries, and probability masses to adapt together.
- We evaluate CADRE on real-world datasets across multiple base distributional heads and forecasting backbones. Results show consistent improvements in probabilistic forecasting and gains in point forecast accuracy in most settings. Further experiments demonstrate benefits under cross-dataset domain shift and reveal an association between greater marginal skewness and larger probabilistic gains.

2 Related Work

Distributional refinement. Distributional refinement uses an existing predictive distribution as a starting point for learning a more flexible forecast. The Distributional Refinement Network (DRN) (Avanzi et al., 2026) learns residual adjustments to baseline bin probabilities on shared value-space outpoints. This has two restrictions: (1) the partition cannot follow each forecast's conditional distribution, and (2) the unchanged exterior density prevents correction of tail probabilities. CADRE addresses both restrictions. It constructs bins from each forecast's quantiles and learns their masses across the full support, including the two unbounded outer bins. The partition therefore follows changes in predictive location, scale, and shape, while refinement can redistribute mass between the center and tails. Spliced Binned-Pareto forecasting also combines a binned body with adaptive tail splice points, but fits separate generalized Pareto components (Ehrlich et al., 2021). CADRE instead

retains the current base distribution's conditional tail shapes and learns their masses; it adjusts tail frequency without introducing a separate family of tail distributions.

Discretized probabilistic forecasting. Discretization represents predictive uncertainty through probabilities assigned to ordered target intervals (Li et al., 2021; Bergsma et al., 2022; Schulz & Lerch, 2022). Chronos predicts probabilities over quantized time series values (Ansari et al., 2024), while QuantWeather jointly trains regression and probability heads using historical climatological quantile categories (Chen et al., 2026). However, these partitions lack an explicit correspondence to the current forecast's conditional quantiles. CADRE defines its bins through a shared probability grid mapped by each forecast's quantile function. Its thresholds adapt to the sample, channel, and horizon, so the refinement head learns corrections relative to the forecast being refined. For supervision, CADRE adopts the joint binary cross-entropy (JBCE) loss of Deep Distribution Regression (Li et al., 2021) to train cumulative bin probabilities; cumulative softmax masses ensure CDF monotonicity.

Recalibration and joint distribution learning. Recalibration aims to align predicted probabilities with observed targets, either after fitting or during model training (Gneiting et al., 2007; Kuleshov et al., 2018; Song et al., 2019; Kuleshov & Deshpande, 2022; Dheur & Ben Taieb, 2023; Rasp & Lerch, 2018). Cal-PIT learns feature-dependent probability maps to reshape a fitted predictive distribution (Dey et al., 2025). Such a post-hoc correction leaves the underlying representation unchanged. Quantile Recalibration Training addresses this limitation by incorporating recalibration into optimization (Dheur & Ben Taieb, 2024). Collaborating Networks also jointly learn a CDF and its inverse, using binary supervision at learned quantile locations and an inverse-consistency objective (Zhou et al., 2021b). CADRE uses a different coupling: the base head retains its likelihood or pinball objective, while the refinement head learns masses on the base head's evolving quantile partition. Both objectives update the temporal backbone. Gradients through reference boundaries and summaries are stopped, allowing JBCE to improve the shared features without differentiating through quantile construction. The result is a jointly trained forecaster in which the reference distribution, the mass corrections, and the temporal representation all adapt during refinement.

3 Problem Setup and Preliminaries

3.1 Probabilistic Time Series Forecasting

We consider multivariate, multi-horizon probabilistic time series forecasting. Let $\mathbf{x}_{1:L} = (\mathbf{x}_1, \ldots, \mathbf{x}_L) \in \mathbb{R}^{C\times L}$ denote an observed historical window, where C is the number of channels and L is the context length. For a forecasting horizon H, the goal is to learn a model that maps this historical window to conditional predictive distributions for the future observations $\mathbf{y}_{1:H} := \mathbf{x}_{L+1:L+H} \in \mathbb{R}^{C\times H}$. We express this model as the composition of a forecasting backbone ϕ and a base distributional head f:

$$(f \circ \phi)(\mathbf{x}_{1:L}) = \mathbb{P}(\cdot \mid \mathbf{x}_{1:L}) \in \mathcal{P}(\mathbb{R})^{C\times H}, \tag{1}$$

where $\mathcal{P}(\mathbb{R})$ denotes the set of probability distributions on $\mathbb{R}$, and $\mathbb{P}(\cdot \mid \mathbf{x}_{1:L})$ denotes a $C \times H$ collection of conditional marginal predictive distributions, one for each channel and future time step. The backbone ϕ extracts temporal representations from the historical window, and the head f maps these representations to the corresponding predictive distributions. This notation specifies marginal forecasts and does not, by itself, parameterize their joint dependence across channels and horizons.

Point forecast. A standard point forecasting model maps the historical window $\mathbf{x}_{1:L}$ to a single predicted trajectory $\hat{\mathbf{y}}_{1:H} \in \mathbb{R}^{C\times H}$, assigning one value to each channel and future time step. A probabilistic forecasting model can also produce point forecasts by taking the *mean* or *median* of each predictive distribution. The predictive mean minimizes expected squared error, while the predictive median minimizes expected absolute error. A probabilistic model can therefore be evaluated through both the accuracy of its point forecasts and the quality of its predictive distributions.

3.2 Distributional Evaluation

Given two probability distributions on $\mathbb{R}$ with cumulative distribution functions (CDFs) F and F', their misalignment can be quantified by the integrated squared CDF difference, $d^2_{\mathrm{CDF}}(F, F') =$

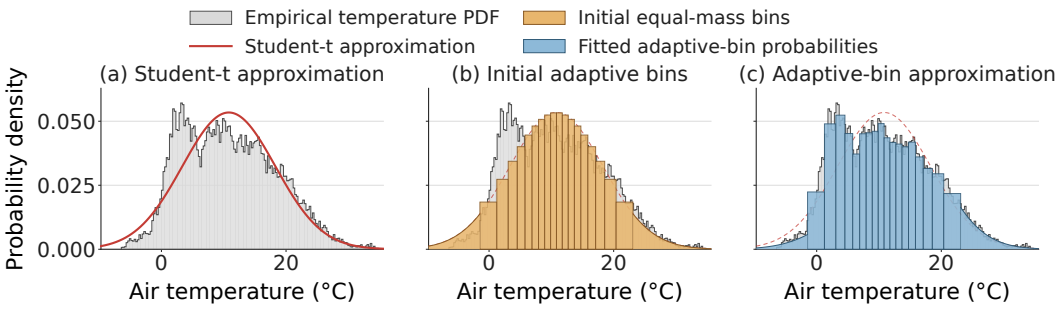

Figure 2: Illustration of CADRE. (a) Empirical density (gray) and Student-t approximation (red). (b) Adaptive bins initialized with equal probability mass. (c) Fitting the bin probabilities captures asymmetry and local structure in the empirical density.

$\int_{-\infty}^{\infty}(F(u)-F'(u))^2\,\mathrm{d}u$. This discrepancy is nonnegative and equals zero if and only if the distributions coincide. Evaluation in time series forecasting is more challenging because the oracle conditional distribution is generally unknown. A historical window $\mathbf{x}_{1:L}$ is typically paired with only one realized future trajectory, providing a single observation for each channel and future step. The corresponding oracle CDF for a fixed channel and future step is therefore unavailable for direct comparison with the predictive CDF. To address these challenges, following prior work (Gneiting & Katzfuss, 2014; Jordan et al., 2019), we evaluate predictive distributions using proper scoring rules and average the resulting scores over observed outcomes, which reflect distributional misalignment in expectation.

Continuous ranked probability score (CRPS). For a fixed historical window, channel, and future step, let $\widehat{F}$ denote the predictive CDF and y the observed outcome. CRPS measures the integrated squared difference between $\widehat{F}$ and the CDF of a point mass at y:

$$\mathrm{CRPS}(\widehat{F},y)=\int_{-\infty}^{\infty}(\widehat{F}(u)-\mathbf{1}\{y\le u\})^2\,\mathrm{d}u, \tag{2}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. This score requires only the predictive CDF and the observed outcome. Let $F^\star$ denote the corresponding oracle conditional CDF. If both distributions have finite first moments, then $\mathbb{E}_{Y\sim F^\star}[\mathrm{CRPS}(\widehat{F},Y)-\mathrm{CRPS}(F^\star,Y)]=d^2_{\mathrm{CDF}}(\widehat{F},F^\star)$ (Gneiting & Raftery, 2007). Thus, the excess expected CRPS relative to the oracle equals the integrated squared CDF discrepancy, and minimizing expected CRPS is equivalent to minimizing this discrepancy, with the unique minimizer $\widehat{F}=F^\star$.

Weighted quantile loss (wQL). Weighted quantile loss evaluates predicted quantiles using normalized pinball loss (Kan et al., 2022). Let y_i and $\widehat{F}_i$ denote the observed value and predictive conditional CDF, where i indexes all evaluated windows, channels, and future steps. For a quantile level $\alpha\in(0,1)$, the predicted quantile is $\hat{q}_i(\alpha)=\inf\{u\in\mathbb{R}:\widehat{F}_i(u)\ge\alpha\}$. Denote the pinball loss for a residual r by $\rho_\alpha(r)=\max\{\alpha r,(\alpha-1)r\}$. The corresponding wQL is

$$\mathrm{wQL}(\alpha)=\frac{2\sum_i\rho_\alpha(y_i-\hat{q}_i(\alpha))}{\sum_i|y_i|}. \tag{3}$$

Throughout the paper, we report mean wQL, computed by averaging this score over the evaluated quantile levels $\alpha\in\{0.1,0.2,\ldots,0.9\}$.

4 Calibration via Distributional Refinement

In this section, we introduce CAlibration via Distributional REfinement (CADRE), a framework for refining the predictive distributions of a probabilistic forecaster (as illustrated in Figure 1).

4.1 Base Probabilistic Forecaster

For each channel c and forecast horizon h, a shared forecasting backbone ϕ and a base distributional head f produce

$$\mathbf{z}_{c,h}=\phi(\mathbf{x}_{1:L})_{c,h},\qquad \widehat{\mathbb{P}}_{c,h}(\cdot\mid\mathbf{x}_{1:L})=f(\mathbf{z}_{c,h}), \tag{4}$$

where $\mathbf{z}_{c,h} \in \mathbb{R}^d$ denotes the temporal features for channel c and horizon h. The mapping $f : \mathbb{R}^d \to \mathcal{F}_{\mathcal{P}} \subseteq \mathcal{P}(\mathbb{R})$ encompasses the prediction head and its associated distributional construction, with $\mathcal{F}_{\mathcal{P}}$ denoting the family of predictive distributions represented by the head. For example, a Gaussian head uses $\mathcal{F}_{\mathcal{P}} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R},\ \sigma > 0\}$. Note that f is shared across all channels and horizons. The base predictive distribution $f(\mathbf{z}_{c,h})$ determines the adaptive bin boundaries, while the temporal features $\mathbf{z}_{c,h}$ are also used by the distributional refinement head to learn context-dependent probabilities for these bins.

4.2 Constructing Adaptive Bins

Given the predictive distribution $\widehat{\mathbb{P}}_{c,h}(\cdot \mid \mathbf{x}_{1:L})$ from (4), we construct adaptive bins using its quantiles. Figure 2(a) shows the empirical distribution of air temperature and its Student-t approximation, plotted in red. Let K denote the total number of bins. As illustrated in Fig. 2(b), we divide the entire probability range into K intervals with equal probability $p = 1/K$. The corresponding quantiles define the bin boundaries in target space, yielding narrower bins where the base distribution concentrates probability and wider bins in less likely regions.

Distributions like Gaussian and Student-t have support over the entire real line. Their quantile-based partitions therefore contain two unbounded tail bins, together with the finite interior bins. Within each finite bin, probability mass is distributed uniformly. A uniform density cannot be normalized over an unbounded interval, so each tail bin instead uses the base distribution conditioned on that interval to determine its shape. In Fig. 2(b), the two outer bins consequently follow the Student-t tail shapes indicated by the dashed red curve, preserving their characteristic decay. The subsequent refinement step learns the probability masses of all bins, including both tails, allowing the probabilities of extreme outcomes to change.

4.3 Distributional Refinement

CADRE learns a context-dependent allocation of probability across these bins in Section 4.2 to better represent the conditional distribution of future observations. As shown in Figure 1, the refinement head g maps the temporal features $\mathbf{z}_{c,h}$ (defined in (1)) to probabilities over the K adaptive bins, where g is a linear projection followed by softmax:

$$g : \mathbb{R}^d \to \Delta^{K-1}, \qquad g(\mathbf{z}_{c,h}) = \mathrm{softmax}(\mathbf{W}^\top \mathbf{z}_{c,h}), \tag{5}$$

Here, $\mathbf{W} \in \mathbb{R}^{d \times K}$ parameterizes the prediction head g, and $\Delta^{K-1} \subset \mathbb{R}^K$ denotes the simplex of probability distributions over the K bins. Setting $\mathbf{W} = 0$ assigns an initial probability of $1/K$ to each adaptive bin defined in Section 4.2.

For a given base forecast, refinement keeps the bin boundaries fixed and rescales each bin's density to integrate to its predicted probability. Finite bins remain uniform, while the unbounded bins retain their conditional base-distribution shapes up to this scaling. Figure 2(c) presents a toy example illustrating how the adaptive bin probabilities are updated to approximate the target distribution. Their corresponding distributional evaluations are reported in Table 1.

Table 1: Distributional evaluation in Figure 2. Student-t and CADRE correspond to (a) and (c), respectively. Lower CRPS (in °C)) and wQL are better; Δ denotes the percentage change relative to the Student-t approximation.

	CRPS	Δ (%)	wQL	Δ (%)
Student-t	4.278	—	0.408	—
CADRE	**4.257**	**−0.488**	**0.406**	**−0.467**

Cumulative binary cross-entropy. A natural approach is to minimize categorical cross-entropy using the bin containing the observed outcome as the class label. However, this objective does not account for the ordering of the bins: for a fixed probability assigned to the observed bin, the loss is unchanged whether the remaining probability is placed in neighboring or distant bins. To incorporate this ordering, we use cumulative binary cross-entropy (CBCE), following Li et al. (2021). CBCE supervises the probability that the outcome falls at or below each finite bin boundary, making the loss sensitive to how mass is allocated across the ordered bins. For any distribution $\mathbb{P}$, let $\mathbf{b}(\mathbb{P}) \in \mathbb{R}^{K-1}$ denote the $K-1$ finite boundaries constructed as in Section 4.2. For a single forecast with temporal features $\mathbf{z}$ and observed value y, write $b_k = [\mathbf{b}(f(\mathbf{z}))]_k$ and $\widehat{F}_k = \sum_{j=1}^{k} [g(\mathbf{z})]_j$ for the k-th boundary and the corresponding refined cumulative probability, where $[\cdot]_k$ denotes the k-th component of a vector.

The CBCE loss is defined as

$$\ell_{\mathrm{CBCE}}(y, \mathbf{z}) = -\frac{1}{K-1} \sum_{k=1}^{K-1} \Big[\mathbf{1}\{y \le b_k\} \log \widehat{F}_k + (1 - \mathbf{1}\{y \le b_k\}) \log(1 - \widehat{F}_k) \Big]. \tag{6}$$

The training objective $\mathcal{L}_{\mathrm{CBCE}}$ averages ℓ_{CBCE} over historical windows, channels, and forecast horizons. Cumulative sums of the softmax probabilities ensure that the predicted CDF values are nondecreasing.

Joint training objective. After pretraining the base forecaster, we jointly optimize the backbone ϕ, base head f, and refinement head g using the objective

$$\mathcal{L} = \mathcal{L}_{\mathrm{CBCE}} + \lambda \mathcal{L}_{\mathrm{Base}}, \tag{7}$$

where $\lambda > 0$ controls the contribution of the base forecasting loss. For Gaussian and Student-t heads, we use the negative log-likelihood $-\log \hat{p}(y)$, where $\hat{p}$ is the density of the base forecast $f(\mathbf{z})$. For quantile heads, we average the pinball loss $\rho_\alpha(y - \hat{q}(\alpha))$ over the predicted quantiles $\hat{q}(\alpha)$, using ρ_α defined in Section 3.2. In both cases, $\mathcal{L}_{\mathrm{Base}}$ averages the resulting per-forecast loss over historical windows, channels, and horizons. As illustrated in Figure 1, the base loss supervises the forecast that determines the adaptive bins, while CBCE trains the refined probability allocation. During joint training, CBCE updates g and the shared backbone ϕ through the temporal features $\mathbf{z}_{c,h}$, while the base loss updates f and ϕ.

5 Experiments

5.1 Experimental Setup

Datasets and forecasting backbones. We evaluate CADRE on six multivariate time-series datasets: Electricity, PEMS07, Solar, Traffic, Weather, and Wind. For each dataset, we predict $H = 96$ future steps from $L = 336$ historical observations, using training, validation, and test splits in a 7:2:1 ratio. Each base forecaster and its CADRE counterpart share the same data partitions and forecasting windows; distribution-shift experiments use the alternative partitions specified in their respective tables. Dataset descriptions are provided in Appendix A.1. We use DLinear for the main comparison and additionally evaluate PatchTST and iTransformer to assess CADRE across forecasting architectures. Results are reported in Tables 2 and 3, respectively.

Base predictive heads. We consider three base predictive heads: two parametric distribution heads (Gaussian and Student-t) and a quantile regression head. Each head is trained jointly with its forecasting backbone.

- ***Gaussian:*** For each input, the Gaussian head predicts a location $\mu_{c,h}$ and a positive scale $\sigma_{c,h}$ for each channel c and horizon h, defining the predictive distribution

$$\widehat{\mathbb{P}}_{c,h}(\cdot \mid \mathbf{x}_{1:L}) = \mathcal{N}\big(\mu_{c,h}, \sigma^2_{c,h}\big).$$

- ***Student-t:*** For each input, the Student-t head predicts a location $\mu_{c,h}$, a positive scale $\sigma_{c,h}$, and degrees of freedom $\nu_{c,h}$ for each channel and horizon, defining the predictive distribution

$$\widehat{\mathbb{P}}_{c,h}(\cdot \mid \mathbf{x}_{1:L}) = t_{\nu_{c,h}}(\mu_{c,h}, \sigma_{c,h}),$$

  where $t_\nu(\mu, \sigma)$ denotes the Student-t distribution with location μ, scale σ, and ν degrees of freedom.

- ***Quantile regression:*** For each input, the quantile regression head predicts $M = 21$ quantiles at quantile levels

$$\alpha = \{0.01, 0.05, 0.10, \ldots, 0.90, 0.95, 0.99\}.$$

  The head is trained by minimizing average pinball loss over these levels.

Table 2: Gaussian-, Student-t-, and quantile-based forecasting before and after CADRE, using DLinear as the forecasting backbone and $K = 64$ adaptive bins. Results are reported as the mean ± sample standard deviation over five random seeds. Blue (negative) Δ values indicate improvement; red (positive) values indicate deterioration. Lower values (bold) are better.

		Gaussian-based			Student-t-based			Quantile-based		
	Dataset	Base	CADRE	Δ (%)	Base	CADRE	Δ (%)	Base	CADRE	Δ (%)
RMSE	Electricity	$0.375_{\pm.001}$	$\mathbf{0.362}_{\pm.002}$	−3.50	$0.377_{\pm.000}$	$\mathbf{0.369}_{\pm.006}$	−2.25	$0.377_{\pm.001}$	$\mathbf{0.373}_{\pm.007}$	−1.06
	PEMS07	$0.432_{\pm.002}$	$\mathbf{0.317}_{\pm.002}$	−26.55	$0.432_{\pm.001}$	$\mathbf{0.327}_{\pm.004}$	−24.42	$0.454_{\pm.011}$	$\mathbf{0.318}_{\pm.002}$	−29.86
	Solar	$0.516_{\pm.001}$	$\mathbf{0.469}_{\pm.008}$	−9.07	$0.524_{\pm.001}$	$\mathbf{0.464}_{\pm.011}$	−11.52	$0.474_{\pm.006}$	$\mathbf{0.421}_{\pm.005}$	−11.12
	Traffic	$\mathbf{0.641}_{\pm.000}$	$0.660_{\pm.007}$	+3.00	$0.671_{\pm.003}$	$\mathbf{0.662}_{\pm.008}$	−1.25	$1.045_{\pm.091}$	$\mathbf{0.659}_{\pm.041}$	−36.93
	Weather	$0.469_{\pm.001}$	$\mathbf{0.448}_{\pm.002}$	−4.52	$0.483_{\pm.001}$	$\mathbf{0.452}_{\pm.002}$	−6.33	$0.476_{\pm.001}$	$\mathbf{0.453}_{\pm.006}$	−4.82
	Wind	$\mathbf{0.885}_{\pm.000}$	$0.893_{\pm.004}$	+0.94	$0.897_{\pm.001}$	$\mathbf{0.893}_{\pm.004}$	−0.46	$0.896_{\pm.004}$	$\mathbf{0.891}_{\pm.005}$	−0.53
CRPS	Electricity	$0.185_{\pm.000}$	$\mathbf{0.164}_{\pm.002}$	−11.10	$0.179_{\pm.000}$	$\mathbf{0.167}_{\pm.007}$	−6.66	$0.179_{\pm.000}$	$\mathbf{0.174}_{\pm.007}$	−2.82
	PEMS07	$0.213_{\pm.001}$	$\mathbf{0.141}_{\pm.001}$	−33.75	$0.206_{\pm.000}$	$\mathbf{0.141}_{\pm.001}$	−31.66	$0.198_{\pm.001}$	$\mathbf{0.142}_{\pm.001}$	−28.07
	Solar	$0.203_{\pm.003}$	$\mathbf{0.160}_{\pm.003}$	−21.29	$0.202_{\pm.001}$	$\mathbf{0.160}_{\pm.010}$	−20.79	$0.176_{\pm.003}$	$\mathbf{0.139}_{\pm.002}$	−21.26
	Traffic	$0.227_{\pm.001}$	$\mathbf{0.185}_{\pm.000}$	−18.39	$0.217_{\pm.000}$	$\mathbf{0.187}_{\pm.001}$	−13.96	$0.227_{\pm.003}$	$\mathbf{0.192}_{\pm.006}$	−15.31
	Weather	$0.230_{\pm.001}$	$\mathbf{0.183}_{\pm.002}$	−20.54	$0.209_{\pm.001}$	$\mathbf{0.184}_{\pm.001}$	−11.86	$0.210_{\pm.001}$	$\mathbf{0.186}_{\pm.007}$	−11.35
	Wind	$0.464_{\pm.000}$	$\mathbf{0.449}_{\pm.001}$	−3.26	$0.465_{\pm.005}$	$\mathbf{0.449}_{\pm.002}$	−3.53	$0.467_{\pm.003}$	$\mathbf{0.458}_{\pm.006}$	−2.02
Mean wQL	Electricity	$0.242_{\pm.001}$	$\mathbf{0.215}_{\pm.002}$	−11.12	$0.234_{\pm.000}$	$\mathbf{0.219}_{\pm.009}$	−6.59	$0.234_{\pm.001}$	$\mathbf{0.227}_{\pm.009}$	−2.80
	PEMS07	$0.270_{\pm.001}$	$\mathbf{0.179}_{\pm.001}$	−33.69	$0.260_{\pm.001}$	$\mathbf{0.179}_{\pm.001}$	−31.32	$0.252_{\pm.002}$	$\mathbf{0.181}_{\pm.001}$	−28.32
	Solar	$0.280_{\pm.004}$	$\mathbf{0.223}_{\pm.004}$	−20.35	$0.277_{\pm.001}$	$\mathbf{0.223}_{\pm.013}$	−19.53	$0.244_{\pm.004}$	$\mathbf{0.194}_{\pm.003}$	−20.64
	Traffic	$0.307_{\pm.001}$	$\mathbf{0.250}_{\pm.001}$	−18.69	$0.292_{\pm.000}$	$\mathbf{0.252}_{\pm.001}$	−13.82	$0.305_{\pm.004}$	$\mathbf{0.258}_{\pm.008}$	−15.55
	Weather	$0.369_{\pm.002}$	$\mathbf{0.293}_{\pm.003}$	−20.54	$0.333_{\pm.001}$	$\mathbf{0.295}_{\pm.002}$	−11.28	$0.335_{\pm.002}$	$\mathbf{0.299}_{\pm.010}$	−10.87
	Wind	$0.610_{\pm.000}$	$\mathbf{0.591}_{\pm.002}$	−3.08	$0.611_{\pm.001}$	$\mathbf{0.591}_{\pm.003}$	−3.28	$0.615_{\pm.003}$	$\mathbf{0.602}_{\pm.008}$	−2.05

Evaluation metrics. We evaluate point forecasts using root mean squared error (**RMSE**), taking the predictive mean as the point forecast for all heads. For Gaussian and Student-t heads, this is $\mu_{c,h}$ (with $\nu_{c,h} > 1$ for Student-t). For the quantile head, we integrate the quantile function over $(0, 1)$, using piecewise-linear interpolation between predicted quantiles and exponential extrapolation in the tails. For CADRE, we weight each bin's mean by its learned probability: finite bins contribute their midpoints, while the two unbounded bins contribute the base distribution's conditional means over the respective tails.

We evaluate predictive distributions using continuous ranked probability score (**CRPS**) and mean weighted quantile loss (**Mean wQL**), defined in Section 3.2. Mean wQL averages wQL over the nine levels $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$. Gaussian and Student-$t$ CDFs follow directly from their predicted parameters; the quantile head's CDF is obtained by inverting its continuous quantile function. For CADRE, let F_0 denote the base CDF and use the bin boundaries b_k and cumulative probabilities $\widehat{F}_k$ defined in Section 4.3. The refined CDF is

$$\widehat{F}(u) = \begin{cases} K\widehat{F}_1 F_0(u), & u \le b_1, \\ \widehat{F}_{k-1} + (\widehat{F}_k - \widehat{F}_{k-1})\dfrac{u - b_{k-1}}{b_k - b_{k-1}}, & b_{k-1} < u \le b_k, \quad 2 \le k \le K-1, \\ 1 - K(1 - \widehat{F}_{K-1})(1 - F_0(u)), & u > b_{K-1}. \end{cases}$$

The tail terms rescale the base tails, each with original probability $1/K$, to their learned probabilities. We compute CRPS by substituting $\widehat{F}$ into (2), including both unbounded tails. Quantiles for mean wQL are obtained by inverting $\widehat{F}$: linear interpolation between bin boundaries within finite bins, and inversion of the rescaled base CDF in the tails.

For any metric S, we report the relative change from the base forecaster to CADRE as

$$\Delta = \frac{S_{\text{CADRE}} - S_{\text{Base}}}{S_{\text{Base}}} \times 100\%. \tag{8}$$

Lower values are better for all three metrics, and negative Δ indicates improvement over the base forecaster.

Table 3: Average scores over the six datasets (see Tables 2) for DLinear, PatchTST, and iTransformer with Student-t heads. Δ is the percentage change between the averaged scores. Blue (negative) values indicate improvement. Lower values (bold) are better.

	DLinear			PatchTST			iTransformer		
Metric	Base	CADRE	Δ (%)	Base	CADRE	Δ (%)	Base	CADRE	Δ (%)
RMSE	0.564	**0.528**	−6.41	0.555	**0.524**	−5.55	0.547	**0.542**	−0.82
CRPS	0.246	**0.215**	−12.86	0.222	**0.207**	−7.05	0.225	**0.223**	−0.96
Mean wQL	0.335	**0.293**	−12.36	0.303	**0.282**	−6.83	0.307	**0.305**	−0.76

Table 4: Forecasting performance under cross-dataset transfer to Weather using DLinear. Baseline uses the source pretrained model. Student-t finetunes only the base head f on Weather, whereas CADRE jointly finetunes f and the refinement head g. The backbone ϕ remains fixed. Lower values (bold) are better.

	Electricity → Weather			PEMS07 → Weather			Solar → Weather		
Metric	Baseline	Student-t	CADRE	Baseline	Student-t	CADRE	Baseline	Student-t	CADRE
RMSE	0.552	0.535	**0.506**	0.569	0.557	**0.495**	0.578	0.557	**0.508**
CRPS	0.270	0.263	**0.224**	0.268	0.257	**0.212**	0.312	0.263	**0.220**
Mean wQL	0.431	0.421	**0.360**	0.427	0.411	**0.340**	0.478	0.421	**0.352**

5.2 Experimental Results

Table 2 compares DLinear base forecasters with their CADRE counterparts across Gaussian, Student-t, and quantile heads. CADRE reduces CRPS and mean wQL in all 18 (dataset, head) settings and improves RMSE in 16 of them. Distributional gains are particularly pronounced on PEMS07, Solar, and Traffic, indicating that refinement can improve probabilistic forecasts under different assumptions about the base predictive distribution. The two RMSE exceptions occur with the Gaussian head on Traffic and Wind, where RMSE increases by 3.00% and 0.94%, respectively, despite improvements in both probabilistic scores. A possible explanation is that CBCE supervises cumulative probabilities without directly minimizing the error of the predictive mean. In practice, improvements in distributional quality therefore need not translate into lower RMSE. These results motivate further investigation of how to balance point accuracy and distributional quality during training and model selection.

The improvements for the quantile head are also informative. Unlike Gaussian and Student-t heads, it is not restricted to symmetric predictive distributions, yet it still benefits from refinement on every dataset. This suggests that CADRE can complement the flexibility of a finite-quantile representation as well as improve forecasts from symmetric parametric families.

To assess whether these benefits extend across forecasting architectures, Table 3 reports scores averaged over the six datasets for DLinear, PatchTST, and iTransformer with Student-t heads. CADRE improves all three averaged metrics for each backbone. The relative gains are largest for DLinear, followed by PatchTST and iTransformer. These results support CADRE's applicability across different architectures, with the magnitude of improvement depending on the backbone.

Figure 3 compares the empirical distribution of observed future data (left) with the predictive distributions of the base forecaster (middle) and CADRE (right), obtained by pooling samples across test windows. CADRE more closely captures this peak and the right-skewed shape, providing qualitative evidence of improved agreement between the predictive and observed distributions. Further details and more examples are provided in Appendix A.4.

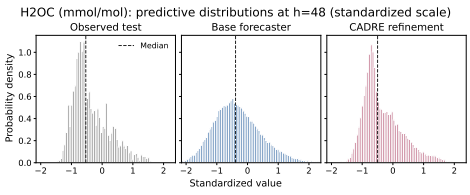

Figure 3: Observed and predictive distributions.

Transfer learning under domain shift. Table 4 examines transfer to Weather from DLinear models pretrained on Electricity, PEMS07, and Solar, with the source pretrained backbone ϕ frozen throughout adaptation. CADRE improves RMSE, CRPS, and mean wQL in all three transfers. In particular, it reduces CRPS by approximately 14.9%–17.6%

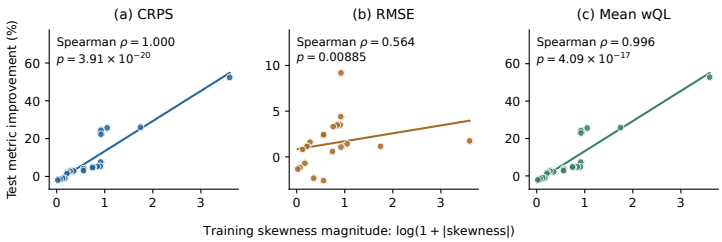

Figure 4: Training skewness and test set improvements across the 21 Weather channels. Positive improvement indicates better performance after refinement.

relative to the finetuned Student-t head. These gains show that refinement provides benefits beyond adapting the parameters of the base distribution alone. They also suggest that CADRE can improve forecasting under domain shift while reusing the source model's temporal representation.

5.3 When Does Refinement Help?

Using the Weather dataset, we further investigate the conditions under which CADRE provides the greatest benefit. The dataset contains 21 variables spanning temperature, humidity, wind, solar radiation, and rainfall, with each variable represented by a separate forecast channel. These variables differ in their distributional shapes, providing a setting to examine whether CADRE benefits more from refining forecasts for variables with greater asymmetry. We examine this relationship by comparing each variable's training set skewness s_c with its test set improvement in CRPS, RMSE, and mean wQL. We use $x_c = \log(1 + |s_c|)$ to measure asymmetry in either direction while compressing the large range of skewness values.Improvement is the percentage reduction in each metric relative to its corresponding base forecaster. Figure 4 reports Spearman rank correlations, which measure whether more skewed channels tend to improve more without assuming a linear relationship.

Asymmetry is strongly associated with improvements in CRPS ($\rho = 1.000$, $p = 3.91 \times 10^{-20}$) and mean wQL ($\rho = 0.996$, $p = 4.09 \times 10^{-17}$), and moderate for RMSE ($\rho = 0.564$, $p \approx 0.00885$). Thus, more skewed channels tend to exhibit larger gains across all three metrics, although this ordering is more consistent for the probabilistic scores. For example, the most skewed channel, *rain*, improves by 52.36% in CRPS and 52.80% in mean wQL, compared with 1.75% in RMSE. The distributional associations remain positive after excluding this channel ($\rho = 1.000$ and 0.995, respectively), so they are not solely a consequence of its extreme skewness.

These results suggest that CADRE is particularly beneficial for probabilistic forecasting of strongly asymmetric variables. However, training set skewness describes each channel's marginal distribution; this analysis therefore establishes an empirical association rather than a causal explanation of refinement gains.

6 Conclusion

We introduced CADRE, a distributional refinement framework for probabilistic time series forecasting. CADRE partitions the full support using base forecast quantiles, including two unbounded outer bins, and learns their probability masses. Joint training adapts the temporal representation, base forecast, and refined allocation through a shared backbone. Across six datasets, CADRE consistently improves CRPS and mean wQL for Gaussian, Student-t, and finite-quantile heads, while improving RMSE in most settings. Additional experiments show benefits across backbones and under cross-dataset transfer, while a channel-level analysis on Weather links greater marginal skewness to larger probabilistic gains. These findings support using base forecasts to guide flexible predictive distributions without fixing their final probability allocation.

Limitations. CADRE relies on the base forecast for its quantile partition and conditional tail shapes, which may constrain refinement under misspecification. Future work could develop more flexible, input-adaptive partitions and tails. Moreover, the refinement objective does not directly optimize point forecast accuracy, so probabilistic gains need not yield better point forecasts. Architectures and objectives that better balance point accuracy and distributional quality warrant further study.

AI USE STATEMENT

In this work, we used generative AI tools to search the relevant literature and polish the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. Transactions on Machine Learning Research, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

Benjamin Avanzi, Eric T. Dong, Patrick J. Laub, and Bernard Wong. Distributional refinement network: Distributional forecasting via deep learning. Insurance: Mathematics and Economics, 128:103246, 2026. doi: 10.1016/j.insmatheco.2026.103246. URL https://doi.org/10.1016/j.insmatheco.2026.103246.

Shane Bergsma, Tim Zeyl, Javad Rahimipour Anaraki, and Lei Guo. C2FAR: Coarse-to-fine autoregressive networks for precise probabilistic forecasting. In Advances in Neural Information Processing Systems, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/899511e37a8e01e1bd6f6f1d377cc250-Abstract.html.

John Bjørnar Bremnes. Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. Monthly Weather Review, 148(1):403–414, 2020. doi: 10.1175/MWR-D-19-0227.1. URL https://doi.org/10.1175/MWR-D-19-0227.1.

Lei Chen, Xinyu Su, Xiaohui Zhong, and Hao Li. QuantWeather: Quantile-aware probabilistic forecasting for subseasonal precipitation, 2026. URL https://arxiv.org/abs/2605.10297.

Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff G. Schneider. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. In Advances in Neural Information Processing Systems, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5b168fdba5ee5ea262cc2d4c0b457697-Abstract.html.

Biprateep Dey, David Zhao, Brett H. Andrews, Jeffrey A. Newman, Rafael Izbicki, and Ann B. Lee. Towards instance-wise calibration: Local amortized diagnostics and reshaping of conditional densities (LADaR). Machine Learning: Science and Technology, 6(4):045058, 2025. doi: 10.1088/2632-2153/ae1f05. URL https://doi.org/10.1088/2632-2153/ae1f05.

Victor Dheur and Souhaib Ben Taieb. A large-scale study of probabilistic calibration in neural network regression. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pp. 7813–7836. PMLR, 2023. URL https://proceedings.mlr.press/v202/dheur23a.html.

Victor Dheur and Souhaib Ben Taieb. Probabilistic calibration by design for neural network regression. In Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, volume 238 of Proceedings of Machine Learning Research, pp. 3133–3141. PMLR, 2024. URL https://proceedings.mlr.press/v238/dheur24a.html.

Elena Ehrlich, Laurent Callot, and François-Xavier Aubet. Spliced binned-pareto distribution for robust modeling of heavy-tailed time series. In ICLR 2021 Workshop on Robust and Reliable Machine Learning in the Real World, 2021. URL https://arxiv.org/abs/2106.10952.

Jan Gasthaus, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski. Probabilistic forecasting with spline quantile function RNNs. In Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, volume 89 of Proceedings of Machine Learning Research, pp. 1901–1910. PMLR, 2019. URL https://proceedings.mlr.press/v89/gasthaus19a.html.

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

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.

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

Alexander Jordan, Fabian Krüger, and Sebastian Lerch. Evaluating probabilistic forecasts with scoringRules. Journal of Statistical Software, 90(12):1–37, 2019. doi: 10.18637/jss.v090.i12.

Kelvin Kan, François-Xavier Aubet, Tim Januschowski, Youngsuk Park, Konstantinos Benidis, Lars Ruthotto, and Jan Gasthaus. Multivariate quantile function forecaster. arXiv preprint arXiv:2202.11316, 2022. URL https://arxiv.org/abs/2202.11316.

Volodymyr Kuleshov and Shachi Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 11683–11693. PMLR, 2022. URL https://proceedings.mlr.press/v162/kuleshov22a.html.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 2796–2804. PMLR, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

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. ACM, 2018. doi: 10.1145/3209978.3210006.

Rui Li, Brian J. Reich, and Howard D. Bondell. Deep distribution regression. Computational Statistics & Data Analysis, 159:107203, 2021. doi: 10.1016/j.csda.2021.107203. URL https://doi.org/10.1016/j.csda.2021.107203.

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

Yuqi Nie, Nam 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.

Youngsuk Park, Danielle Maddix, François-Xavier Aubet, Kelvin Kan, Jan Gasthaus, and Yuyang Wang. Learning quantile functions without quantile crossing for distribution-free time series forecasting. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, volume 151 of Proceedings of Machine Learning Research, pp. 8127–8150. PMLR, 2022. URL https://proceedings.mlr.press/v151/park22a.html.

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

Stephan Rasp and Sebastian Lerch. Neural networks for postprocessing ensemble weather forecasts. Monthly Weather Review, 146(11):3885–3900, 2018. doi: 10.1175/MWR-D-18-0187.1. URL https://doi.org/10.1175/MWR-D-18-0187.1.

Aaron Rumack, Ryan J. Tibshirani, and Roni Rosenfeld. Recalibrating probabilistic forecasts of epidemics. PLOS Computational Biology, 18(12):e1010771, 2022. doi: 10.1371/journal.pcbi.1010771. URL https://doi.org/10.1371/journal.pcbi.1010771.

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. doi: 10.1016/j.ijforecast.2019.07.001.

Benedikt Schulz and Sebastian Lerch. Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. Monthly Weather Review, 150(1):235–257, 2022. doi: 10.1175/MWR-D-21-0150.1. URL https://doi.org/10.1175/MWR-D-21-0150.1.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pp. 914–921, 2020. doi: 10.1609/aaai.v34i01.5438.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 5897–5906. PMLR, 2019. URL https://proceedings.mlr.press/v97/song19a.html.

Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. In Advances in Neural Information Processing Systems, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/73c03186765e199c116224b68adc5fa0-Abstract.html.

Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multi-horizon quantile recurrent forecaster. arXiv preprint arXiv:1711.11053, 2017. URL https://arxiv.org/abs/1711.11053.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

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, 2021a. doi: 10.1609/aaai.v35i12.17325.

Tianhui Zhou, Yitong Li, Yuan Wu, and David Carlson. Estimating uncertainty intervals from collaborating networks. Journal of Machine Learning Research, 22(257):1–47, 2021b. URL https://jmlr.org/papers/v22/20-1100.html.

A Appendix

A.1 Dataset Descriptions and Sources

The six multivariate time-series datasets 5 that span electricity consumption (Electricity), solar power generation (Solar), wind power generation (Wind), transportation (PEMS07 and Traffic), and weather forecasting. Electricity, Traffic, and Weather are widely used long-horizon forecasting benchmarks (Zhou et al., 2021a; Zeng et al., 2023), while Solar follows the multivariate forecasting benchmark introduced by Lai et al. (Lai et al., 2018). Wind is obtained from the TFB benchmark suite (Qiu et al., 2024). PEMS07 contains traffic-flow measurements collected by the California Department of Transportation's Performance Measurement System and is a standard graph/time-series traffic forecasting benchmark (Song et al., 2020).

A.2 Implementation and Evaluation Details

Optimization. For the main DLinear experiments, we initialize the backbone ϕ and base predictive head f from the pretrained base forecaster, and jointly optimize them with the refinement head g. We use $K = 64$ adaptive bins and the joint objective $\mathcal{L} = \mathcal{L}_{\text{CBCE}} + \lambda \mathcal{L}_{\text{Base}}$ with $\lambda = 0.01$. Optimization uses AdamW with separate learning rates for the refinement head and the pretrained forecasting components. The learning rate for refinement head is 0.001, the learning rate for backbone and base distributional forecaster is 10^{-5}. Experiments were conducted on NVIDIA H100 80GB GPUs on AWS.

Table 5: Summary of the multivariate forecasting datasets.

Dataset	Channels	Frequency
Electricity	321	hourly
PEMS07	883	5 min
Solar	137	10 min
Traffic	862	hourly
Weather	21	10 min
Wind	7	15 min

A.3 Additional Forecasting Backbones

We also evaluate CADRE with PatchTST (Nie et al., 2023) and iTransformer (Liu et al., 2024), using Student-t base heads.

For these larger backbones, we use a refinement-only initialization before the joint optimization described in Section 4.3. Starting from the pretrained base forecaster, we first freeze the backbone and Student-t head and optimize only the refinement head using CBCE. We then restore the refinement checkpoint with the lowest validation CBCE, unfreeze all components, and jointly optimize the backbone, Student-t head, and refinement head using the joint loss

$$\mathcal{L} = \mathcal{L}_{\text{CBCE}} + \lambda \mathcal{L}_{\text{NLL}}.$$

This schedule separates the initial fitting of the newly introduced probability-mass head from adaptation of the pretrained temporal representation. During refinement-only training, the temporal features and reference distribution remain fixed, allowing the refinement head to first learn a probability redistribution on a stationary representation. Joint training then allows the temporal representation, base distribution, and refined probabilities to adapt together. Table 6 reports the corresponding results for the pretrained base forecaster, refinement-only training, and joint CADRE optimization.

A.4 Predictive Distribution Visualizations

To provide a qualitative view of how CADRE changes the shape of the predictive distribution, we compare the empirical distribution of held-out observations with the predictive marginals produced by the base forecaster and the corresponding CADRE model. We visualize Gaussian, Student-t, and quantile-based forecasters separately.

For a fixed channel and forecast horizon, each test window contributes one observed target value. For the corresponding base and CADRE forecasters, we draw $S = 4$ samples from the complete conditional predictive distribution for that test window and pool the samples across all test windows. The resulting histograms therefore summarize the predictive marginal distributions induced by the collection of conditional forecasts over the test set.

These plots are intended to illustrate changes in distributional shape, including asymmetry, concentration, and tail behavior, rather than to represent the predictive distribution of a single test instance. All panels are shown on the standardized scale used for evaluation.

Following figures provide a qualitative comparison between the empirical distribution of held-out observations and the predictive marginals induced by the base and refined forecasters. Across the shown examples, the pretrained base forecasters produce predictive marginals that are relatively regular and smooth compared with the observed test distribution. For Gaussian and Student-t heads, this is consistent with the fact that the base predictive family is symmetric. For the quantile head, the predictive distribution is more flexible, but its pooled marginal can still smooth out local asymmetry and irregular structure in the observed data. After refinement, CADRE reallocates probability mass across adaptive bins and produces predictive marginals whose overall shape more closely matches the empirical pattern of the held-out observations, including changes in concentration, asymmetry, and tail behavior.

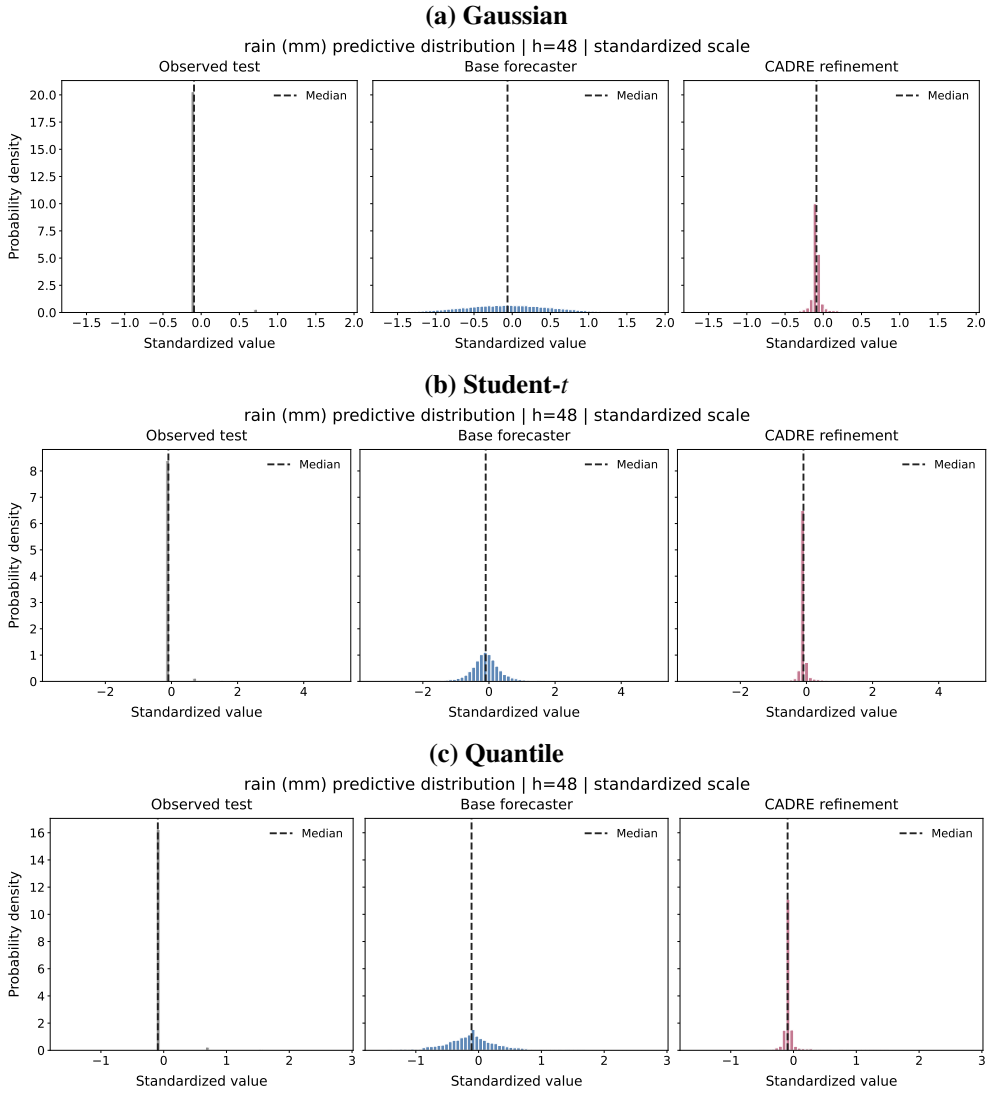

Figure 5: **Observed and predictive marginal distributions on Weather at forecast horizon $h = 48$.** Results are shown for (a) Gaussian, (b) Student-t, and (c) quantile base forecasters. In each row, the left panel shows the empirical distribution of the held-out observations, the middle panel shows the predictive marginal induced by the pretrained base forecaster, and the right panel shows the predictive marginal after CADRE refinement. Each test window contributes one observed target and four samples from each corresponding conditional predictive distribution; predictive samples are pooled across test windows. Dashed vertical lines denote the median.

Table 6: Student-t forecasting results with PatchTST and iTransformer for seed 42. Base denotes the pretrained base forecaster; Ref.-only trains only the refinement head with the base forecaster frozen; Joint jointly trains the backbone and both heads from the refinement-only checkpoint. $\Delta = (M_{\text{Joint}} - M_{\text{Base}})/M_{\text{Base}} \times 100\%$ is computed from the three-decimal scores shown and reported to two decimal places. Blue (negative) values indicate improvement, red (positive) values indicate deterioration, and zero changes are shown in black. Bold marks the best configuration based on unrounded metrics.

		PatchTST				iTransformer			
	Dataset	Base	Ref.-only	Joint	Δ (%)	Base	Ref.-only	Joint	Δ (%)
CRPS	Electricity	0.169	0.167	**0.165**	−2.37	0.175	0.175	**0.175**	0.00
	PEMS07	0.149	0.147	**0.137**	−8.05	0.147	0.146	**0.145**	−1.36
	Solar	0.201	0.169	**0.145**	−27.86	0.175	0.181	**0.171**	−2.29
	Traffic	0.201	0.193	**0.186**	−7.46	0.242	0.239	**0.238**	−1.65
	Weather	0.175	0.173	**0.172**	−1.71	0.179	0.178	**0.178**	−0.56
	Wind	0.439	0.438	**0.435**	−0.91	0.432	0.432	**0.430**	−0.46
Mean wQL	Electricity	0.221	0.219	**0.217**	−1.81	0.229	0.229	**0.229**	0.00
	PEMS07	0.190	0.187	**0.174**	−8.42	0.187	0.185	**0.185**	−1.07
	Solar	0.275	0.237	**0.201**	−26.91	0.244	0.254	**0.240**	−1.64
	Traffic	0.271	0.261	**0.250**	−7.75	0.327	0.323	**0.322**	−1.53
	Weather	0.280	0.278	**0.275**	−1.79	0.287	0.286	**0.286**	−0.35
	Wind	0.578	0.577	**0.574**	−0.69	0.569	0.569	**0.567**	−0.35
RMSE	Electricity	0.372	0.368	**0.365**	−1.88	0.376	0.375	**0.375**	−0.27
	PEMS07	0.354	0.349	**0.324**	−8.47	0.339	0.334	**0.332**	−2.06
	Solar	0.523	0.484	**0.428**	−18.16	**0.482**	0.504	0.489	+1.45
	Traffic	0.746	0.715	**0.707**	−5.23	0.765	0.743	**0.742**	−3.01
	Weather	0.453	0.449	**0.444**	−1.99	0.448	**0.447**	0.448	0.00
	Wind	0.883	0.881	**0.878**	−0.57	0.871	0.870	**0.868**	−0.34