CONFORMAL PREDICTION FOR TIME SERIES WITH DEEP SEQUENCE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in deep learning for time series prediction have amplified the need for reliable uncertainty quantification. Conformal prediction has gained attention as a distribution-free framework for constructing prediction intervals with coverage guarantees. However, its coverage guarantees rely on data exchangeability, an assumption generally violated in time series. Active research has focused on developing conformal prediction methods for time series that overcome this limitation. While deep sequence models, such as recurrent neural networks and Transformers, have often been used in conformal prediction for time series, limited work has systematically studied how deep sequence models can be utilized in conformal prediction for time series. In this work, we systematically investigate the use of deep sequence models in conformal prediction for time series through three approaches: conditional quantile regression, conditional quantile function estimation, and localized conformal prediction. We provide a theoretical analysis establishing asymptotic conditional coverage guarantees for all three approaches under suitable assumptions. Through comprehensive experiments on real-world datasets, we demonstrate the effectiveness of leveraging deep sequence models into conformal prediction for time series.

1 INTRODUCTION

Time series prediction aims to forecast future outcomes from past sequential observations of features and outcomes (Box et al., 2015), and has been widely applied across a broad range of downstream tasks, such as climate modeling, biological science, and finance (Lim & Zohren, 2021). Recent advances in deep learning have led to the development of various foundation models for time series prediction (Miller et al., 2024; Kim et al., 2025; Wen et al., 2023; Kong et al., 2025). The growing use of such foundation models highlights the need for reliable uncertainty quantification in time series prediction. Conformal prediction has been actively adopted in time series prediction as a reliable and distribution-free framework for uncertainty quantification, constructing prediction sets with desired coverage guarantees (Zhou et al., 2025).

Conformal prediction provides prediction intervals with coverage guarantees under the assumption that the data are exchangeable, without requiring additional distributional assumptions (Vovk et al., 2005; Vovk, 2012). Despite these attractive properties for uncertainty quantification, applying conformal prediction to time series forecasting remains challenging as the temporal dependence present in time series generally violates the exchangeability assumption. Consequently, substantial research has focused on extending conformal prediction to time series by developing methods that account for temporal dependence and the resulting violations of exchangeability (Xu & Xie, 2021; 2023a; Auer et al., 2024; Lee et al., 2024a; Menadjiev et al., 2026). Within this line of research, deep sequence models, such as recurrent neural networks (Elman, 1990; Hochreiter & Schmidhuber, 1997; Cho et al., 2014) and Transformers (Vaswani et al., 2017), have been increasingly adopted due to their ability to capture complex temporal dependencies (Lee et al., 2024b; Chen et al., 2024; Li & Rodríguez, 2025; Cini et al., 2025; Neglia et al., 2026; Lee et al., 2026). However, a systematic understanding of how deep sequence models can be leveraged and their practical effectiveness in conformal prediction for time series remains limited.

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.

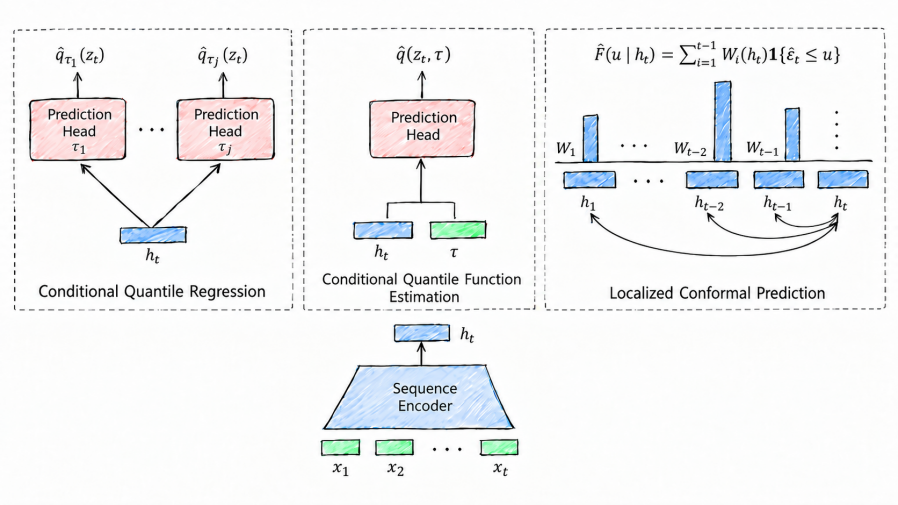

Figure 1: Three approaches for leveraging deep sequence models in conformal prediction for time series. All methods use a deep sequence model to encode the time-series context, and the resulting representation is used differently within each method: (1) quantile regression with quantile-specific prediction heads conditioned on the representation; (2) quantile function estimation for arbitrary quantile levels conditioned on the representation; and (3) similarity-based weighting computed from the representations for estimating the conditional cumulative distribution function.

To address this gap, we make the following contributions:

- We propose three methods that leverage deep sequence models for conformal prediction for time series: conditional quantile regression, conditional quantile function estimation, and localized conformal prediction.
- We establish asymptotic conditional coverage guarantees for all three methods under suitable assumptions.
- We conduct comprehensive experiments on real-world datasets, demonstrating the competitive performance of the proposed methods and providing practical insights into their use.

1.1 Related Works

Conformal prediction beyond the exchangeability assumption. A line of work on conformal prediction under distribution shift focuses on adaptively updating the calibration procedure online. Gibbs & Candes (2021) address distribution shifts in sequential data by adaptively updating a coverage level based on past coverage errors, and Zaffran et al. (2022) extend this framework to time series by proposing an aggregation-based approach to improve its adaptivity. More recent work along this line includes Bhatnagar et al. (2023); Angelopoulos et al. (2023; 2024); Gibbs & Candès (2024); Areces et al. (2025). Another line of work on conformal prediction under distribution shift between the calibration and test set is by reweighting the data points. Tibshirani et al. (2019) address covariate shift by reweighting calibration samples according to the likelihood ratio between the test and training covariate distributions, requiring this ratio to be known or accurately estimated. Barber et al. (2023) uses predefined weights, such as exponentially decaying weights that assign greater importance to more recent observations. However, determining appropriate weights remains challenging when the underlying shift is unknown, and inaccurate weight estimates or poorly chosen weighting schemes may lead to coverage degradation or inefficient prediction intervals. Although not specifically designed to address non-exchangeability, localized conformal prediction (Guan, 2023) uses similarity-based weights to emphasize calibration samples near the test point and adapt prediction intervals to local distributional characteristics.

Conformal prediction for time series. Early approaches combine ensemble-based predictors with sequential conformal calibration. Xu & Xie (2021; 2023a) use bootstrap ensembles with a

rolling window of prediction residuals, and Jensen et al. (2022) combine ensemble quantile regression with conformal calibration to adapt prediction interval widths. Related to the idea of localized conformal prediction (Guan, 2023), a growing body of work reweights calibration observations based on their relevance to the current prediction. Auer et al. (2024) use modern Hopfield networks to assign greater weight to observations from similar temporal regimes, and Lee et al. (2024a) employ a reweighted Nadaraya–Watson estimator to obtain data-adaptive weights for quantile estimation of prediction residuals. Chen et al. (2024) use deep sequence models to learn latent representations for attention-based reweighting of calibration samples. Neglia et al. (2026) use similarities between reservoir states to reweight historical residuals, capturing local temporal dynamics without training a calibration network.

Another important body of work estimates conditional quantiles of future prediction residuals given historical observations. Xu & Xie (2023b) use quantile random forests to exploit dependencies among residuals, and Lee et al. (2024b) extend this approach using a Transformer to leverage the past residuals and covariates. Cini et al. (2025) combine sequence models with graph neural networks to estimate residual quantiles across correlated time series while learning the relational structure from data. Menadjiev et al. (2026) introduce DistMatch, which groups residual histories into distributionally similar bins using Kolmogorov–Smirnov statistics and performs bin-specific quantile estimation.

1.2 Problem Setup

Consider a sequence of observations $\{(X_t, Y_t)\}_{t\geq 1}$, where $X_t \in \mathbb{R}^d$ denotes a d-dimensional feature vector and $Y_t \in \mathbb{R}$ denotes a continuous scalar outcome. Given a forecast horizon $\ell \geq 1$, let Z_t denote the contextual information available at time t, used to construct a prediction interval for $Y_{t+\ell}$. For instance, if the most recent w features are used as context, then $Z_t = [X_{t-(w-1)}, \ldots, X_t]$.

We assume that a point prediction model $\widehat{f}$ is available to produce

$$\widehat{Y}_{t+\ell} = \widehat{f}(Z'_t), \quad \widehat{\epsilon}_{t+\ell} = Y_{t+\ell} - \widehat{f}(Z'_t). \tag{1}$$

Here, Z'_t denotes the information used by the point prediction model and need not coincide with Z_t. The choice of Z'_t depends on the prediction model $\widehat{f}$ and can include any information available to the predictor at the time of prediction. For example, a standard regression model may use $Z'_t = X_t$, whereas a time-series foundation model uses a longer historical context. We use the signed prediction residual $\widehat{\epsilon}_{t+\ell}$ as the nonconformity score to construct prediction intervals.

Write T for the number of observed context–residual pairs used to fit a quantile model or to calibrate a fixed representation. All the methods we propose estimate residual quantiles, denoted by $\widehat{q}_{T,\tau}(z)$. For fixed levels $0 < \alpha_l < \alpha_h < 1$ with $\alpha_h - \alpha_l = 1 - \alpha$, define

$$\widehat{C}_T^{\alpha,\ell}(z, \widehat{y}) = [\widehat{y} + \widehat{q}_{T,\alpha_l}(z), \widehat{y} + \widehat{q}_{T,\alpha_h}(z)]. \tag{2}$$

The observed pairs are $\{(Z_i, \widehat{\epsilon}_{i+\ell})\}_{i=1}^T$, so the last calibration outcome is available at time $T+\ell$. The prediction interval at time $t \geq T+\ell$ is evaluated at $(Z_t, \widehat{Y}_{t+\ell})$. Throughout the analysis, $t_T \geq T+\ell$ is an arbitrary deterministic sequence of forecast origins.

We seek narrow intervals with coverage conditional on the observed context, a stronger criterion than average marginal coverage (Vovk, 2012). However, exact distribution-free conditional coverage with finite-length intervals is generally impossible (Lei & Wasserman, 2014). Our theoretical target is therefore

$$\mathbb{P}\left(Y_{t_T+\ell} \in \widehat{C}_T^{\alpha,\ell}(Z_{t_T}, \widehat{Y}_{t_T+\ell}) \mid Z_{t_T}\right) = 1 - \alpha + o_P(1). \tag{3}$$

Since the coverage error is bounded by one, (3) also implies marginal coverage converging to $1-\alpha$.

2 Methods

In this section, we propose three approaches to leverage deep sequence models in conformal prediction for time series. Figure 1 provides an overview of the three proposed methods. The three methods uses a sequence encoder but use its encoded representation in different ways: as a finite collection of quantile heads, a head that also takes the quantile level as input, or kernel weights for

observed calibration residuals. Given the context z_t at time t, an encoder $\mathrm{Enc}_{\theta'}(\cdot)$ maps z_t to a latent representation h_t that summarizes the relevant contextual information, as $h_t = \mathrm{Enc}_{\theta'}(z_t)$, where θ' denotes the parameters of the encoder. We write $H_t = \mathrm{Enc}_{\theta'}(Z_t)$ for the corresponding random representation.

2.1 Conditional Quantile Regression using Deep Sequence Models

Conditional quantile regression with deep sequence models has been explored in probabilistic time series forecasting (Wen et al., 2017; Chen et al., 2020; Lim et al., 2021). Building on this idea, a straightforward way to leverage deep sequence models in conformal prediction for time series is to model the conditional quantiles of prediction residuals given their historical context. Specifically, a deep sequence model encodes the historical time series context into a latent representation, which is then used as input to multiple prediction heads, one for each target quantile level. The resulting quantile estimates are used to construct prediction intervals.

Let $\mathcal{T} = \{\tau_1 < \ldots < \tau_J\} \subset (0, 1)$ denote the set of quantile levels at which the conditional quantiles of the prediction residuals are estimated. We assume that the number of quantile levels J is fixed throughout training and inference. For notational simplicity, we omit the subscript when referring to a generic quantile level $\tau \in \mathcal{T}$. In an unconstrained multi-head parameterization, for each $\tau \in \mathcal{T}$ the corresponding quantile-specific prediction head $g_{\theta^\tau}(\cdot)$ takes h_t as input and produces an estimate of the conditional τ-quantile of the prediction residual $\widehat{q}_{T,\tau}$, given by

$$\widehat{q}_{T,\tau}(z_t) = g_{\theta^\tau}(h_t) = g_{\theta^\tau}\left(\mathrm{Enc}_{\theta'}(z_t)\right),$$

where θ^τ denotes the parameters of the prediction head associated with quantile level τ.

Rather than employing a separate encoder for each quantile level in $\mathcal{T}$, a common architecture uses a single shared encoder together with multiple quantile-specific prediction heads (Wen et al., 2017; Lim et al., 2021). Transformer (Vaswani et al., 2017) and recurrent neural networks (Hochreiter & Schmidhuber, 1997; Cho et al., 2014) are common choices for the encoder.

Although the model is trained using quantile loss, quantile crossing may occur due to estimation, approximation, or optimization errors (Bondell et al., 2010). To prevent quantile crossing, we parameterize the prediction heads using nonnegative increments. The head corresponding to the lowest quantile level predicts its quantile value directly, while each subsequent head produces a nonnegative increment that is added to the preceding quantile estimate. Specifically, we define

$$\begin{aligned}
&\widehat{q}_{T,\tau_1}(z_t) = g_{\theta^{\tau_1}}(h_t), \\
&\widehat{q}_{T,\tau_j}(z_t) = \widehat{q}_{T,\tau_{j-1}}(z_t) + \sigma(g_{\theta^{\tau_j}}(h_t)), \quad j = 2, \ldots, J,
\end{aligned}$$

where $\sigma : \mathbb{R} \to [0, \infty)$ is a nonnegative activation function, such as ReLU or softplus. This construction guarantees

$$\widehat{q}_{T,\tau_1}(z_t) \leq \cdots \leq \widehat{q}_{T,\tau_J}(z_t),$$

for every context z_t. We refer to this conditional quantile regression approach using a deep sequence encoder as DSCP-QR throughout the paper.

2.2 Conditional Quantile Function Estimation using Deep Sequence Models

Conditional quantile regression with deep sequence models provides a direct way to estimate prediction residual quantiles given historical context. However, the multi-head formulation requires a separate prediction head for each target quantile level, restricting estimation to a predefined set of levels. To address this limitation, we consider conditional quantile function estimation, in which the quantile level is provided as an additional model input alongside the encoded historical context. This allows a single prediction head to estimate arbitrary quantile levels. Similar formulations have been explored in distributional reinforcement learning (Dabney et al., 2018; Yang et al., 2019), uncertainty estimation (Tagasovska & Lopez-Paz, 2019), and probabilistic time series forecasting (Gouttes et al., 2021).

Rather than specifying a finite set of target quantile levels, we allow the quantile level τ to vary continuously over $(0, 1)$. The prediction head combines h_t with the requested scalar level τ to estimate the corresponding conditional quantile of the prediction residual:

$$\widehat{q}_{T,\tau}(z_t) = g_{\theta''}(h_t, \tau) = g_{\theta''}\left(\mathrm{Enc}_{\theta'}(z_t), \tau\right),$$

where θ'' denotes the prediction-head parameters.

Unlike the finite-head approach, this construction does not restrict the endpoints to quantile levels specified during model construction. By varying the requested levels, the same trained model can produce intervals with different nominal coverage levels and different allocations of nominal miscoverage between the lower and upper tails, without additional prediction heads or retraining.

Similar to conditional quantile regression, quantile crossing may occur when estimating quantiles using the learned conditional quantile function. However, the nonnegative-increment parameterization used in § 2.1 is not applicable here, since the conditional quantile function is designed to estimate quantiles at arbitrary levels rather than at a finite set of prespecified quantile levels. To prevent quantile crossing, we therefore construct the prediction head as a neural network that is partially monotonic with respect to the quantile level τ, while remaining unrestricted with respect to the context representation h_t. Specifically, the prediction head is constrained to satisfy

$$g_{\theta''}(h_t, \tau_1) \leq g_{\theta''}(h_t, \tau_2), \quad \text{for any } \tau_1 < \tau_2.$$

Consequently, the estimated conditional quantile function is nondecreasing in the quantile level, thereby preventing quantile crossing. The details of the partially monotonic prediction head are provided in Appendix B.3. We refer to this conditional quantile function estimation approach using a deep sequence encoder as DSCP-QFE throughout the paper.

2.3 Localized Conformal Prediction using Deep Sequence Models

Localized conformal prediction (Guan, 2023) assigns greater weight to calibration samples that are similar to the test sample, rather than weighting all calibration samples equally. Building on this idea for time series prediction setup, we use a sequence model to encode historical contexts into latent representations and compute similarity-based calibration weights in the learned representation space. These weights are then used to estimate the conditional cumulative distribution function (CDF) of the prediction residuals.

We train the encoder using quantile losses at multiple fixed quantile levels. This encourages the learned representation to retain information about the context for conditional distribution of the residuals. Consequently, contexts with similar conditional distribution are expected to exhibit similar learned representations, making distances in the representation space informative for identifying relevant calibration samples. The resulting representations therefore can be used to construct similarity-based calibration weights, assigning greater weight to calibration contexts that are closer to the test context in the learned representation space.

For this method, the latent representation serves as a learned state variable for local calibration. For an ℓ-step ahead prediction, our goal is to estimate the true conditional CDF of the prediction residual,

$$F^H_{\widehat{\epsilon}}(u \mid h) \coloneqq \mathbb{P}\left(\widehat{\epsilon}_{t+\ell} \leq u \mid H_t = h\right).$$

Given a calibration set $\{(H_i, \widehat{\epsilon}_{i+\ell})\}_{i=1}^T$, nonnegative kernel K, bandwidth $b_T > 0$, and representation metric d_H, we define the weighted empirical CDF $\widehat{F}^H_{T,\hat{\epsilon}}(\cdot \mid h)$ at a query representation h as

$$K_{T,i}(h) = K(d_H(H_i, h)/b_T), \qquad S_T(h) = \sum_{i=1}^T K_{T,i}(h),$$

$$W_{T,i}(h) = K_{T,i}(h)/S_T(h), \qquad \widehat{F}^H_{T,\hat{\epsilon}}(u \mid h) = \sum_{i=1}^T W_{T,i}(h)\mathbb{1}\{\widehat{\epsilon}_{i+\ell} \leq u\},$$

when $S_T(h) > 0$; otherwise set $\widehat{F}^H_{T,\hat{\epsilon}}(u \mid h) = \mathbb{1}\{u \geq 0\}$.

We obtain quantiles by inversion of this weighted empirical CDF:

$$\widehat{q}^H_{T,\tau}(h) = \inf\{u : \widehat{F}^H_{T,\hat{\epsilon}}(u \mid h) \geq \tau\},$$

$$\widehat{q}_{T,\tau}(z) = \widehat{q}^H_{T,\tau}(\mathrm{Enc}_{\theta'}(z)), \qquad \widehat{F}_{T,\hat{\epsilon}}(u \mid z) = \widehat{F}^H_{T,\hat{\epsilon}}(u \mid \mathrm{Enc}_{\theta'}(z)).$$

Here, the inverse is simply computed by sorting the calibration residuals and accumulating their weights. It is nondecreasing in τ, so the resulting interval (2) has ordered endpoints and quantile crossing will not occur. We refer to this localized conformal prediction using a deep sequence encoder as DSCP-Local throughout the paper.

3 THEORETICAL ANALYSIS

Our analysis establishes asymptotic conditional coverage through quantile consistency for DSCP-QR and DSCP-QFE, and conditional CDF consistency for DSCP-Local. Throughout, $\widehat{f}$ is fixed, $t_T \geq T + \ell$ is deterministic, and α_l, α_h are fixed with $\alpha_h - \alpha_l = 1 - \alpha$. Write $W_t = (Z_t, \widehat{\epsilon}_{t+\ell})$, $\mathcal{Z} = \mathrm{supp}(P_Z)$, and $q^\star_\tau(z)$ for the true conditional residual quantile. The full assumptions and proofs are provided in Appendix A.

3.1 ASYMPTOTIC CONDITIONAL COVERAGE FOR DSCP-QR

DSCP-QR fits all target quantiles through a shared encoder. Therefore, we require a single encoder–head configuration to approximate the target quantiles jointly and training to attain the joint empirical minimum up to an $o_P(1)$ error. To relate this training criterion to population loss under temporal dependence, we assume W_t is stationary and α-mixing. These requirements and the remaining regularity conditions are stated precisely in Assumptions A.1–A.5.

Theorem 3.1 (Conditional coverage for DSCP-QR)**.** *Under Assumptions A.1–A.5, the DSCP-QR interval* (2)*, with $\alpha_l, \alpha_h \in \mathcal{T}$, satisfies*

$$\mathbb{P}\left(Y_{t_T+\ell} \in \widehat{C}_T^{\alpha,\ell}(Z_{t_T}, \widehat{Y}_{t_T+\ell}) \mid Z_{t_T}\right) = 1 - \alpha + o_P(1). \tag{4}$$

Proof sketch. For the quantile loss $\rho_\tau(u) = u(\tau - \mathbb{1}\{u < 0\})$, fixed weights $w_\tau > 0$ with $\sum_{\tau\in\mathcal{T}} w_\tau = 1$, and $\mathbf{q} = (q_\tau)_{\tau\in\mathcal{T}}$, define the expected quantile losses

$$R_\tau(q) = \mathbb{E}\left[\rho_\tau\{\widehat{\epsilon}_{t+\ell} - q(Z_t)\}\right], \quad \mathcal{R}(\mathbf{q}) = \sum_{\tau\in\mathcal{T}} w_\tau R_\tau(q_\tau).$$

Let $\widehat{\mathcal{R}}_T$ be the empirical risk obtained by replacing the expectations in R_τ with averages over $\{(Z_i, \widehat{\epsilon}_{i+\ell})\}_{i=1}^T$, and let r_T be the optimization error in Assumption A.5. Write $\widehat{\mathbf{q}}_T = (\widehat{q}_{T,\tau})_{\tau\in\mathcal{T}}$, $\mathbf{q}^\star = (q^\star_\tau)_{\tau\in\mathcal{T}}$, and let $\mathfrak{Q}_T$ be the jointly feasible model class. Comparing the training loss $\widehat{\mathcal{R}}_T$ with $\mathcal{R}$ gives

$$\begin{aligned} 0 &\leq \mathcal{R}(\widehat{\mathbf{q}}_T) - \mathcal{R}(\mathbf{q}^\star) \\ &\leq \underbrace{2 \sup_{\mathbf{q}\in\mathfrak{Q}_T} |\widehat{\mathcal{R}}_T(\mathbf{q}) - \mathcal{R}(\mathbf{q})|}_{\text{empirical–population error}} + \underbrace{\inf_{\mathbf{q}\in\mathfrak{Q}_T} \mathcal{R}(\mathbf{q}) - \mathcal{R}(\mathbf{q}^\star)}_{\text{joint approximation error}} + \underbrace{r_T}_{\text{optimization error}} \xrightarrow{p} 0. \end{aligned}$$

The first term vanishes by mixing and compactness; the remaining terms vanish by assumption. Thus the fitted quantiles achieve the same limiting expected loss as the true quantiles. The density and compactness conditions ensure that any fixed uniform quantile error entails a strictly positive loss gap, so

$$\max_{\tau\in\mathcal{T}} \sup_{z\in\mathcal{Z}} |\widehat{q}_{T,\tau}(z) - q^\star_\tau(z)| \xrightarrow{p} 0.$$

Let $\widehat{A}_T$ and $A^\star_T$ be the coverage events for the fitted and true quantile endpoints. Stationarity and CDF continuity at the true endpoints give $\mathbb{P}(\widehat{A}_T \triangle A^\star_T) \to 0$. Since $\mathbb{P}(A^\star_T \mid Z_{t_T}) = 1 - \alpha$,

$$\mathbb{E}\left|\mathbb{P}(\widehat{A}_T \mid Z_{t_T}) - (1-\alpha)\right| \leq \mathbb{P}(\widehat{A}_T \triangle A^\star_T) \longrightarrow 0,$$

which implies (4) (see Appendix A.1).

3.2 ASYMPTOTIC CONDITIONAL COVERAGE FOR DSCP-QFE

DSCP-QFE learns one quantile function across levels sampled from ν. As in DSCP-QR, we require accurate approximation and optimization, now for the loss averaged over these levels, under stationary α-mixing data. The sampling distribution ν has full support on a fixed compact interval $\mathcal{I} \subset (0, 1)$, so every neighborhood of a target level contributes to the objective. The full conditions are stated in Assumption A.1 and Assumptions A.12–A.16.

Theorem 3.2 (Conditional coverage for DSCP-QFE)**.** *Under Assumption A.1 and Assumptions A.12–A.16, the DSCP-QFE interval* (2)*, with $\alpha_l, \alpha_h \in \mathcal{I}$, satisfies* (4)*.*

Proof sketch. Write $\widehat{q}_T(z,\tau) = \widehat{q}_{T,\tau}(z)$, $q^\star(z,\tau) = q^\star_\tau(z)$ and $\mathcal{R}(q) = \int_{\mathcal{I}} R_\tau(q(\cdot,\tau))\nu(d\tau)$, with R_τ defined above. The same decomposition as DSCP-QR gives

$$0 \le \mathcal{R}(\widehat{q}_T) - \mathcal{R}(q^\star) = \int_{\mathcal{I}} \underbrace{\left[R_\tau(\widehat{q}_T(\cdot,\tau)) - R_\tau(q^\star_\tau)\right]}_{\text{additional expected loss at level } \tau \ge 0} \nu(d\tau) \xrightarrow{p} 0.$$

The local density condition identifies the true quantile at each level. Joint continuity and full support ensure that an error at any (z,τ) contributes positive loss on a neighborhood of that pair. Compactness therefore converts the displayed loss convergence into

$$\sup_{z\in\mathcal{Z},\tau\in\mathcal{I}} |\widehat{q}_T(z,\tau) - q^\star(z,\tau)| \xrightarrow{p} 0.$$

At the two fixed endpoint levels, the coverage-event comparison from DSCP-QR yields (4) (see Appendix A.2).

3.3 Asymptotic Conditional Coverage for DSCP-Local

For DSCP-Local, the encoder is fixed during calibration and inference, and the weighted residual CDF targets $F^H_{\widehat{\epsilon}}(u \mid h) = \mathbb{P}(\widehat{\epsilon}_{t+\ell} \le u \mid H_t = h)$. The primary coverage guarantee is therefore conditional on the representation H_t.

We assume $V_t = (H_t, \widehat{\epsilon}_{t+\ell})$ is strictly stationary and β-mixing. Local estimation requires continuous conditional residual CDFs that vary little among nearby representations, and sufficient probability mass in these neighborhoods. With a bounded, compactly supported kernel, the bandwidth shrinks to eliminate smoothing bias while satisfying the dependence-adjusted variance condition. The precise conditions are stated in Assumptions A.21–A.24.

Theorem 3.3 (Representation-conditional consistency and coverage)**.** *Under Assumptions A.21–A.24, the DSCP-Local estimator with a fixed encoder satisfies*

$$\sup_u |\widehat{F}^H_{T,\hat{\epsilon}}(u \mid H_{t_T}) - F^H_{\widehat{\epsilon}}(u \mid H_{t_T})| \xrightarrow{p} 0, \tag{5}$$

$$\mathbb{P}\left(Y_{t_T+\ell} \in \widehat{C}^{\alpha,\ell}_T(Z_{t_T}, \widehat{Y}_{t_T+\ell}) \mid H_{t_T}\right) = 1 - \alpha + o_P(1). \tag{6}$$

Proof sketch. Local mass, CDF regularity, and the bandwidth condition yield uniform CDF consistency for P_H-almost every fixed representation. A β-mixing coupling argument extends consistency to the dependent test representation H_{t_T}. The estimated quantiles are then bracketed by oracle quantiles at neighboring probability levels, which gives (6) (see Appendix A.3.1).

Extension to context-conditional coverage requires residual sufficiency: the conditional residual distribution given Z_t depends on Z_t only through H_t. This condition permits dimension reduction while preserving the conditional distribution relevant to coverage. Detailed discussion can be found in Appendix A.3.2.

Corollary 3.4 (Context-conditional coverage for DSCP-Local)**.** *Under the conditions of Theorem 3.3, additionally suppose W_t is strictly stationary and, for every u and P_Z-almost every z,*

$$\mathbb{P}(\widehat{\epsilon}_{t+\ell} \le u \mid Z_t = z) = F^H_{\widehat{\epsilon}}(u \mid \mathrm{Enc}_{\theta'}(z)). \tag{7}$$

Then the DSCP-Local interval also satisfies (4)*.*

Proof sketch. Let $D_T = \sup_u |\widehat{F}^H_{T,\hat{\epsilon}}(u \mid H_{t_T}) - F^H_{\widehat{\epsilon}}(u \mid H_{t_T})|$ and let $\widehat{A}_T$ denote the DSCP-Local coverage event. By (7) and CDF continuity, an oracle interval between levels $a < b$ has probability $b - a$ conditional on Z_{t_T}. On $\{D_T \le \eta\}$, each fitted endpoint at level τ lies between the oracle quantiles at levels $\tau - \eta$ and $\tau + \eta$. Thus, for any sufficiently small fixed $\eta > 0$,

$$\mathbb{E}|\mathbb{P}(\widehat{A}_T \mid Z_{t_T}) - (1-\alpha)| \le 2\eta + \mathbb{P}(D_T > \eta).$$

Since $D_T \xrightarrow{p} 0$ by (5), taking $T \to \infty$ and then $\eta \downarrow 0$ gives convergence in mean, hence in probability (see Appendix A.3.2).

Table 1: Coverage gap (ΔCov), prediction interval width (PI-W), and Winkler score (WS), averaged across sequences within each dataset, for all methods. LR, LSTM, and Chronos indicate the base predictor. The target coverage is $1-\alpha = 0.9$. Results with undercoverage are color-coded according to the magnitude of the coverage gap: results with $-0.5\alpha \leq \Delta\text{Cov} < -0.25\alpha$ are shown in red, while those with $\Delta\text{Cov} < -0.5\alpha$ are greyed out. The best three results except the undercoveraged results based on Winkler score are underlined.

Method	Air								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
Split CP	$+0.028_{\pm.011}$	$100.5_{\pm11.1}$	$150.0_{\pm14.6}$	$-0.005_{\pm.014}$	$73.9_{\pm8.21}$	$137.2_{\pm15.1}$	$+0.011_{\pm.010}$	$187.7_{\pm17.2}$	$315.6_{\pm21.8}$
NexCP	$-0.001_{\pm.001}$	$86.8_{\pm10.6}$	$141.9_{\pm15.4}$	$-0.004_{\pm.001}$	$76.1_{\pm9.10}$	$133.6_{\pm14.9}$	$-0.005_{\pm.001}$	$180.2_{\pm14.2}$	$286.0_{\pm18.9}$
HopCPT	$-0.027_{\pm.032}$	$79.1_{\pm14.9}$	$148.5_{\pm22.7}$	$-0.030_{\pm.028}$	$67.1_{\pm15.7}$	$134.1_{\pm15.6}$	$-0.010_{\pm.038}$	$221.0_{\pm84.5}$	$347.3_{\pm101.2}$
SPCI	$-0.110_{\pm.006}$	$75.9_{\pm8.86}$	$160.7_{\pm16.9}$	$-0.110_{\pm.006}$	$64.2_{\pm7.89}$	$140.5_{\pm16.5}$	$-0.088_{\pm.009}$	$81.9_{\pm7.31}$	$160.5_{\pm13.3}$
KOWCPI	$+0.012_{\pm.009}$	$92.8_{\pm10.1}$	$150.2_{\pm14.6}$	$-0.003_{\pm.010}$	$73.3_{\pm7.81}$	$137.1_{\pm14.5}$	$+0.005_{\pm.009}$	$181.2_{\pm15.5}$	$316.0_{\pm22.0}$
ResCP	$-0.012_{\pm.005}$	$80.6_{\pm10.3}$	$135.3_{\pm14.3}$	$-0.019_{\pm.006}$	$67.6_{\pm9.86}$	$119.8_{\pm13.4}$	$+0.033_{\pm.003}$	$126.6_{\pm10.4}$	$171.4_{\pm12.6}$
DistMatch	$-0.004_{\pm.006}$	$83.8_{\pm9.26}$	$139.1_{\pm14.9}$	$-0.005_{\pm.005}$	$71.3_{\pm7.68}$	$125.3_{\pm13.4}$	$+0.039_{\pm.004}$	$120.5_{\pm10.3}$	$169.0_{\pm13.4}$
DSCP-QR	$+0.012_{\pm.016}$	$85.3_{\pm10.5}$	$121.3_{\pm12.5}$	$-0.021_{\pm.017}$	$61.1_{\pm9.32}$	$103.4_{\pm12.1}$	$+0.014_{\pm.014}$	$87.1_{\pm10.7}$	$125.5_{\pm10.6}$
DSCP-QFE	$-0.064_{\pm.015}$	$61.3_{\pm6.40}$	$132.7_{\pm15.6}$	$-0.126_{\pm.026}$	$37.9_{\pm4.73}$	$124.4_{\pm16.3}$	$-0.091_{\pm.032}$	$52.5_{\pm7.39}$	$170.1_{\pm13.1}$
DSCP-Local	$-0.015_{\pm.003}$	$81.0_{\pm10.0}$	$133.2_{\pm14.4}$	$-0.013_{\pm.005}$	$67.8_{\pm8.67}$	$116.0_{\pm13.3}$	$+0.027_{\pm.006}$	$117.4_{\pm9.14}$	$160.7_{\pm11.4}$
Method	Solar								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
Split CP	$-0.001_{\pm.024}$	$115.4_{\pm85.3}$	$202.6_{\pm81.1}$	$-0.002_{\pm.020}$	$88.1_{\pm30.2}$	$176.3_{\pm33.0}$	$+0.008_{\pm.010}$	$165.9_{\pm23.0}$	$270.4_{\pm26.9}$
NexCP	$+0.001_{\pm.002}$	$111.8_{\pm22.6}$	$181.0_{\pm32.0}$	$-0.002_{\pm.002}$	$89.6_{\pm21.5}$	$161.5_{\pm29.3}$	$-0.001_{\pm.001}$	$158.2_{\pm24.7}$	$249.5_{\pm30.4}$
HopCPT	$+0.001_{\pm.031}$	$106.1_{\pm25.6}$	$148.4_{\pm37.7}$	$+0.007_{\pm.035}$	$76.7_{\pm24.0}$	$123.6_{\pm40.5}$	$+0.004_{\pm.035}$	$114.1_{\pm25.9}$	$179.2_{\pm49.3}$
SPCI	$-0.106_{\pm.028}$	$77.5_{\pm13.2}$	$155.3_{\pm24.1}$	$-0.087_{\pm.067}$	$56.0_{\pm12.8}$	$120.7_{\pm23.6}$	$-0.077_{\pm.016}$	$60.0_{\pm8.97}$	$119.8_{\pm16.5}$
KOWCPI	$+0.004_{\pm.010}$	$114.9_{\pm48.3}$	$196.3_{\pm59.2}$	$+0.005_{\pm.011}$	$91.2_{\pm22.5}$	$172.6_{\pm29.9}$	$-0.010_{\pm.014}$	$154.4_{\pm28.5}$	$263.9_{\pm30.4}$
ResCP	$-0.026_{\pm.007}$	$85.4_{\pm19.0}$	$158.2_{\pm31.4}$	$+0.008_{\pm.005}$	$67.0_{\pm14.3}$	$107.4_{\pm21.5}$	$+0.011_{\pm.004}$	$86.5_{\pm13.7}$	$130.3_{\pm17.0}$
DistMatch	$+0.007_{\pm.008}$	$100.3_{\pm20.1}$	$168.6_{\pm31.4}$	$+0.014_{\pm.009}$	$81.1_{\pm17.3}$	$145.2_{\pm25.7}$	$+0.024_{\pm.006}$	$98.7_{\pm15.9}$	$153.2_{\pm20.9}$
DSCP-QR	$-0.019_{\pm.107}$	$81.9_{\pm14.6}$	$121.6_{\pm82.9}$	$-0.006_{\pm.120}$	$55.4_{\pm10.8}$	$84.0_{\pm84.0}$	$+0.007_{\pm.020}$	$62.9_{\pm9.97}$	$82.4_{\pm11.5}$
DSCP-QFE	$-0.114_{\pm.019}$	$55.9_{\pm9.50}$	$159.7_{\pm41.8}$	$-0.120_{\pm.061}$	$33.4_{\pm8.74}$	$122.8_{\pm37.2}$	$-0.103_{\pm.035}$	$42.1_{\pm8.71}$	$123.3_{\pm15.6}$
DSCP-Local	$-0.001_{\pm.008}$	$88.6_{\pm18.4}$	$123.8_{\pm27.6}$	$+0.016_{\pm.010}$	$64.0_{\pm15.3}$	$87.0_{\pm17.6}$	$+0.011_{\pm.007}$	$83.1_{\pm14.3}$	$109.9_{\pm13.2}$
Method	Sapflow								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
Split CP	$+0.032_{\pm.027}$	$803.5_{\pm577.6}$	$1197.0_{\pm839.7}$	$+0.022_{\pm.022}$	$1318.4_{\pm2199.0}$	$2045.5_{\pm3441.5}$	$+0.004_{\pm.035}$	$2380.8_{\pm1929.1}$	$3854.7_{\pm2892.2}$
NexCP	$+0.002_{\pm.005}$	$656.4_{\pm427.6}$	$1067.9_{\pm731.7}$	$-0.002_{\pm.005}$	$1179.7_{\pm1774.7}$	$1722.0_{\pm2626.9}$	$-0.002_{\pm.005}$	$2154.9_{\pm1774.7}$	$3602.8_{\pm2810.5}$
HopCPT	$+0.015_{\pm.059}$	$790.8_{\pm619.4}$	$1101.1_{\pm777.9}$	$-0.016_{\pm.075}$	$886.7_{\pm927.5}$	$1311.9_{\pm1127.4}$	$-0.001_{\pm.069}$	$2313.5_{\pm1832.1}$	$3088.5_{\pm1985.2}$
SPCI	$+0.044_{\pm.022}$	$704.6_{\pm484.2}$	$979.4_{\pm667.9}$	$+0.031_{\pm.026}$	$650.5_{\pm466.9}$	$913.4_{\pm642.3}$	$-0.032_{\pm.057}$	$793.0_{\pm585.1}$	$1351.1_{\pm942.4}$
KOWCPI	$+0.021_{\pm.058}$	$875.2_{\pm705.2}$	$1277.3_{\pm895.5}$	$+0.009_{\pm.065}$	$977.2_{\pm789.7}$	$1429.3_{\pm1029.9}$	$-0.033_{\pm.076}$	$2024.9_{\pm1730.2}$	$2773.7_{\pm2044.1}$
ResCP	$-0.004_{\pm.024}$	$534.0_{\pm361.2}$	$905.4_{\pm634.6}$	$-0.002_{\pm.024}$	$593.1_{\pm527.6}$	$962.9_{\pm777.3}$	$+0.021_{\pm.026}$	$1373.2_{\pm1111.2}$	$1854.1_{\pm1309.8}$
DistMatch	$+0.024_{\pm.017}$	$655.2_{\pm454.0}$	$988.3_{\pm688.5}$	$+0.031_{\pm.016}$	$706.4_{\pm528.5}$	$1039.6_{\pm754.9}$	$+0.046_{\pm.019}$	$1422.1_{\pm1021.1}$	$1868.7_{\pm1257.6}$
DSCP-QR	$+0.005_{\pm.064}$	$708.5_{\pm511.9}$	$994.5_{\pm687.8}$	$-0.008_{\pm.111}$	$839.3_{\pm917.8}$	$1142.3_{\pm1023.3}$	$-0.061_{\pm.065}$	$864.8_{\pm549.1}$	$1450.2_{\pm953.0}$
DSCP-QFE	$-0.046_{\pm.116}$	$609.6_{\pm421.5}$	$1209.8_{\pm829.9}$	$-0.115_{\pm.128}$	$527.0_{\pm461.3}$	$1188.3_{\pm884.6}$	$-0.137_{\pm.101}$	$563.4_{\pm442.8}$	$1885.1_{\pm1216.4}$
DSCP-Local	$+0.001_{\pm.012}$	$558.2_{\pm375.5}$	$864.2_{\pm619.6}$	$+0.007_{\pm.013}$	$632.1_{\pm550.7}$	$911.4_{\pm743.6}$	$+0.017_{\pm.010}$	$1400.5_{\pm1102.3}$	$1860.4_{\pm1382.5}$

4 EXPERIMENTS

4.1 SETUP

Datasets. We evaluate the proposed methods and baselines on three real-world datasets that are commonly used as benchmarks in conformal prediction for time-series. The Air Quality dataset (Zhang et al., 2017) contains air quality measurements and 10 associated features collected from 12 locations in Beijing, China. The Solar dataset (Sengupta et al., 2018) contains half-hourly measurements of solar radiation along with eight environmental features at 50 different locations. The Sapflow dataset, obtained from the SAPFLUXNET database (Poyatos et al., 2020), contains measurements of water movement within plants. Since the individual sequences in the Sapflow dataset vary in length, we use a subset of 24 time series, each containing between 15,000 and 20,000 observations with varying sampling frequencies, following Auer et al. (2024). Additional details of the datasets are provided in Appendix B.1.

Base predictors. We consider three base predictors: linear regression, a recurrent neural network (RNN), and Chronos (Ansari et al., 2024; 2025). For linear regression, we use a sliding window of the past k time steps of features and outcomes to predict the outcome at the next time step, and train a separate model for each individual time-series sequence. For the RNN, we use LSTM cells to perform the same one-step-ahead prediction using the past k time steps of features and outcomes. Unlike linear regression, a single RNN model is trained jointly using the training splits of all time series sequences. Since Chronos is a pretrained foundation model for time series forecasting, we use its zero-shot forecasts without fine tuning. Additional details on the base predictors are provided in Appendix B.2.

Proposed and baseline methods. For the three proposed methods, we consider LSTM and Transformer as the sequence encoder. Based on the literature review in § 1.1, we consider several representative conformal prediction methods as baseline methods, with an emphasis on methods developed for time series or distribution shift: NexCP (Barber et al., 2023), SPCI (Xu & Xie, 2023b), HopCPT (Auer et al., 2024), KOWCPI (Lee et al., 2024a), ResCP (Neglia et al., 2026), and DistMatch (Menadjiev et al., 2026). For each time-series sequence, the first 33% is used as the training set. The first 67% of the remaining sequence are used for calibration, and the rest are used for testing, resulting in an approximate 33:45:22 split. The training set is used to train the base predictor, the calibration set is used to calibrate the prediction intervals, and the test set is used for evaluation. We consider a one-step-ahead prediction setting where the prediction interval for time $t + 1$ is constructed using the historical context at time t. Implementation details and hyperparameter selection procedures for the proposed methods and baselines are provided in Appendices B.3 and B.4, respectively.

Evaluation metrics. We evaluate all methods using three metrics: coverage gap, prediction interval width, and Winkler score (Winkler, 1972). The coverage gap measures the difference between empirical and target coverage. Prediction interval width measures the size of the prediction interval. The Winkler score jointly accounts for interval width and coverage violations. Each metric is computed over the test set and then averaged across sequences within each dataset. Detailed definitions are provided in Appendix B.5.

4.2 Results

Deep sequence models provide efficient prediction intervals for time series. Table 1 reports the average coverage gap, interval width, and Winkler score for all methods across three datasets at a target coverage level of 0.9. The results for Split CP and NexCP suggest that relying on exchangeability or simply assigning greater weights to recent observations may be insufficient to construct efficient prediction intervals for time series. For each proposed method, we report the best results obtained using either an RNN or a Transformer encoder. DSCP-Local achieves the best overall performance while maintaining stable coverage across all three datasets. DSCP-QR also performs well, often outperforming the strongest baselines, although it occasionally exhibits undercoverage. Similar trends hold at higher and lower target coverage levels of 0.95 and 0.8, as shown in Tables 15 and 16, respectively.

We observe that the RNN encoder generally outperforms the Transformer encoder in our experiments. One possible explanation is that perhaps relatively short-term temporal dependencies are sufficient for modeling the conditional residual distributions in these settings, although further analysis is needed to assess this hypothesis. DSCP-QFE exhibits undercoverage, possibly because its partially monotonic prediction heads limit model expressiveness. The ablation study below, in which this architectural constraint is removed, provides empirical support for this interpretation.

The effect of architectural constraints for preventing quantile crossing on prediction interval efficiency. Table 17 reports ablation results for DSCP-QR and DSCP-QFE without their respective architectural constraints for preventing quantile crossing. For DSCP-QR, removing the incremental prediction heads has marginal effect on performance but leads to quantile crossing. In contrast, removing the partially monotonic heads substantially improves the performance of DSCP-QFE.

5 Conclusion

In this study, we propose three methods that leverage deep sequence models for conformal prediction in time series. We establish asymptotic conditional coverage guarantees for all three methods under suitable assumptions. Experiments on three real-world datasets against a broad set of baseline methods demonstrate the effectiveness of deep sequence models for conformal prediction for time series. We believe that our theoretical analysis and empirical findings provide useful insights and practical guidance for future research. Future work should address the limitations of this study by developing experimental protocols to assess conditional coverage empirically and designing architectures that prevent quantile crossing without compromising prediction interval efficiency. Another promising direction is to develop novel methods that more effectively leverage deep sequence models for time-series conformal prediction.

AI USE STATEMENT

In this work, we used generative AI tools for

1. Supporting the literature review.
2. Assisting in verifying the mathematical soundness of theoretical claims.
3. Assisting in verifying and generating code.
4. Improving the clarity, grammar, and wording of the manuscript.

We have carefully reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by at least two authors.

ETHICS STATEMENT

We confirm that our study complies with the ICLR Code of Ethics and does not present additional ethical issues.

REPRODUCIBILITY STATEMENT

The source code for all experiments is available at `anonymous_url`. Complete proofs and additional experimental details are provided in the Appendix.

REFERENCES

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal pid control for time series prediction. *Advances in neural information processing systems*, 36:23047–23074, 2023.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *Forty-first International Conference on Machine Learning*, 2024.

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.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Felipe Areces, Christopher Mohri, Tatsunori Hashimoto, and John Duchi. Online conformal prediction via online optimization. In *Forty-second International Conference on Machine Learning*, 2025.

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 2337–2363. PMLR, 2023.

Howard D Bondell, Brian J Reich, and Huixia Wang. Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838, 2010.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2, January 2005. ISSN 1549-5787.

Baiting Chen, Zhimei Ren, and Lu Cheng. Conformalized time series with semantic features. *Advances in Neural Information Processing Systems*, 37:121449–121474, 2024.

Yitian Chen, Yanfei Kang, Yixiong Chen, and Zizhuo Wang. Probabilistic forecasting with temporal convolutional neural network. *Neurocomputing*, 399:491–501, 2020.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014.

Andrea Cini, Alexander Jenkins, Danilo Mandic, Cesare Alippi, and Filippo Maria Bianchi. Relational conformal prediction for correlated time series. In *Forty-second International Conference on Machine Learning*, 2025.

R. Dennis Cook. Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, 22(1):1 – 26, 2007.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018.

Jérôme Dedecker, Clémentine Prieur, and Paul Raynaud De Fitte. Parametrized kantorovich-rubinštein theorem and application to the coupling of random variables. In *Dependence in probability and statistics*, pp. 105–121. Springer, 2006.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Adèle Gouttes, Kashif Rasul, Mateusz Koren, Johannes Stephan, and Tofigh Naghibi. Probabilistic time series forecasting with implicit quantile networks. *arXiv preprint arXiv:2107.03743*, 2021.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Julien Herzen, Francesco Lässig, Samuele Giuliano Piazzetta, Thomas Neuer, Léo Tafti, Guillaume Raille, Tomas Van Pottelbergh, Marek Pasieka, Andrzej Skrodzki, Nicolas Huguenin, et al. Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research*, 23(124):1–6, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Vilde Jensen, Filippo Maria Bianchi, and Stian Normann Anfinsen. Ensemble conformalized quantile regression for probabilistic time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9014–9025, 2022.

Jongseon Kim, Hyungjoon Kim, HyunGi Kim, Dongjun Lee, and Sungroh Yoon. A comprehensive survey of deep learning for time series forecasting: architectural diversity and open challenges. *Artificial Intelligence Review*, 58(7):216, 2025.

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

Xiangjie Kong, Zhenghao Chen, Weiyao Liu, Kaili Ning, Lechao Zhang, Syauqie Muhammad Marier, Yichen Liu, Yuhao Chen, and Feng Xia. Deep learning for time series forecasting: a survey. *International Journal of Machine Learning and Cybernetics*, 16(7):5079–5112, 2025.

Jonghyeok Lee, Chen Xu, and Yao Xie. Kernel-based optimally weighted conformal prediction intervals. *arXiv preprint arXiv:2405.16828*, 2024a.

Junghwan Lee, Chen Xu, and Yao Xie. Transformer conformal prediction for time series. *arXiv preprint arXiv:2406.05332*, 2024b.

Junghwan Lee, Chen Xu, and Yao Xie. Flow-based conformal prediction for multi-dimensional time series. In *International Conference on Learning Representations*, volume 2026, pp. 113989–114018, 2026.

Kuang-Yao Lee, Bing Li, and Francesca Chiaromonte. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics*, pp. 221–249, 2013.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

Ruipu Li and Alexander Rodríguez. Neural conformal control for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18439–18447, 2025.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194), 2021.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Enver Menadjiev, Jihyeon Seong, Jisu Yeo, and Jaesik Choi. Distmatch: Adaptive binning via distribution matching for robust sequential conformal prediction. In *Forty-third International Conference on Machine Learning*, 2026.

John A Miller, Mohammed Aldosari, Farah Saeed, Nasid Habib Barna, Subas Rana, I Budak Arpinar, and Ninghao Liu. A survey of deep learning and foundation models for time series forecasting. *arXiv preprint arXiv:2401.13912*, 2024.

Roberto Neglia, Andrea Cini, Michael Bronstein, and Filippo Maria Bianchi. Rescp: Reservoir conformal prediction for time series forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 108607–108636, 2026.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Rafael Poyatos, Víctor Granda, Víctor Flo, Mark A Adams, Balázs Adorján, David Aguadé, Marcos PM Aidar, Scott Allen, M Susana Alvarado-Barrientos, Kristina J Anderson-Teixeira, et al. Global transpiration data from sap flow measurements: the sapfluxnet database. *Earth System Science Data Discussions*, 2020:1–57, 2020.

Emmanuel Rio. Covariance inequalities for strongly mixing processes. In *Annales de l'IHP Probabilités et statistiques*, volume 29, pp. 587–597, 1993.

Emmanuel Rio. *Asymptotic theory of weakly dependent random processes*, volume 80. Springer, 2017.

Manajit Sengupta, Yu Xie, Anthony Lopez, Aron Habte, Galen Maclaurin, and James Shelby. The national solar radiation data base (nsrdb). *Renewable and sustainable energy reviews*, 89:51–60, 2018.

Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. *Advances in neural information processing systems*, 32, 2019.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: a survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 6778–6786, 2023.

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

Robert L Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021.

Chen Xu and Yao Xie. Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023a.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pp. 38707–38727. PMLR, 2023b.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International conference on machine learning*, pp. 25834–25866. PMLR, 2022.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings. Mathematical, physical, and engineering sciences*, 473(2205):20170457, 2017.

Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng. Conformal prediction: A data perspective. *ACM computing surveys*, 58(2):1–37, 2025.

A Detailed Theoretical Analysis and Proofs

This appendix gives the full assumptions and intermediate results underlying Section 3.

Throughout this section, the point predictor $\widehat{f}$ is fixed. The encoder is learned jointly with the quantile heads in DSCP-QR and DSCP-QFE, and is fixed in DSCP-Local. For the stationary pair $(Z_t, \widehat{\epsilon}_{t+\ell})$ and a quantile level $\tau \in (0,1)$, write

$$F_{\widehat{\epsilon}}(u \mid z) := \mathbb{P}\left(\widehat{\epsilon}_{t+\ell} \le u \mid Z_t = z\right), \quad q^\star_\tau(z) := \inf\left\{u \in \mathbb{R} : F_{\widehat{\epsilon}}(u \mid z) \ge \tau\right\},$$

as the conditional distribution of the prediction residual, and the corresponding true conditional τ-quantile function. The forecast origins $t_T \ge T + \ell$ and endpoint levels α_l, α_h are as in Section 1.2.

A.1 DSCP-QR

We first analyze DSCP-QR, which utilizes the finite collection of quantile heads trained with a shared encoder. The argument proceeds from uniform convergence of the training risk to uniform quantile consistency, and then to conditional coverage.

Let $\mathcal{T} = \{\tau_1, \ldots, \tau_J\} \subset (0,1)$ be fixed. The shared encoder and its heads induce a deterministic, nonempty joint class. We collect the estimated conditional quantile functions into the vector $\widehat{\mathbf{q}}_T = (\widehat{q}_{T,\tau})_{\tau\in\mathcal{T}}$, and denote the corresponding vector of true conditional quantile functions by $\mathbf{q}^\star = (q^\star_\tau)_{\tau\in\mathcal{T}}$. For each $\tau \in \mathcal{T}$, let $\mathcal{Q}_{T,\tau}$ denote the candidate class of conditional quantile functions at level τ. The shared-encoder architecture induces a joint class of vector-valued quantile functions

$$\mathfrak{Q}_T \subseteq \prod_{\tau\in\mathcal{T}} \mathcal{Q}_{T,\tau}.$$

By construction, $\widehat{\mathbf{q}}_T \in \mathfrak{Q}_T$.

The conditional quantile functions are estimated using the quantile loss $\rho_\tau(u) = u\,(\tau - \mathbb{1}\{u < 0\})$, where $\mathbb{1}\{\cdot\}$ denotes the indicator function. For a candidate function $q \in \mathcal{Q}_{T,\tau}$, define the empirical quantile risk and the corresponding population risk by

$$\widehat{R}_{T,\tau}(q) = \frac{1}{T}\sum_{i=1}^{T} \rho_\tau(\widehat{\epsilon}_{i+\ell} - q(Z_i)), \quad R_\tau(q) = \mathbb{E}\left[\rho_\tau(\widehat{\epsilon}_{t+\ell} - q(Z_t))\right]. \tag{8}$$

For a generic vector $\mathbf{q} = (q_\tau)_{\tau\in\mathcal{T}}$, define the joint empirical risk and the corresponding joint population risk by

$$\widehat{\mathcal{R}}_T(\mathbf{q}) = \sum_{\tau\in\mathcal{T}} w_\tau \widehat{R}_{T,\tau}(q_\tau), \quad \mathcal{R}(\mathbf{q}) = \sum_{\tau\in\mathcal{T}} w_\tau R_\tau(q_\tau),$$

where the fixed weights satisfy $w_\tau > 0$ and $\sum_{\tau\in\mathcal{T}} w_\tau = 1$. The model is trained using the joint empirical risk $\widehat{\mathcal{R}}_T$ over the class $\mathfrak{Q}_T$.

Assumption A.1 (Stationarity and strong mixing)**.** The process $\{W_t = (Z_t, \widehat{\epsilon}_{t+\ell})\}_{t\geq 1}$ is strictly stationary and strongly mixing, with mixing coefficients $\alpha_W(k)$ satisfying

$$\alpha_W(k) \longrightarrow 0.$$

Assumption A.2 (Residuals and conditional quantile regularity)**.** The following conditions hold.

(i) (Bounded residuals) There exists a constant $B_{\widehat{\epsilon}} < \infty$ such that $|\widehat{\epsilon}_{t+\ell}| \leq B_{\widehat{\epsilon}}$ almost surely for all $t \geq 1$.

(ii) (Regularity around the conditional quantiles) For each $\tau \in \mathcal{T}$, the true conditional τ-quantile $q^\star_\tau(z)$ is unique and continuous on $\mathcal{Z} := \mathrm{supp}(P_Z)$. Moreover, there exist constants $c_f > 0$ and $\delta > 0$ such that, for every $\tau \in \mathcal{T}$ and every $z \in \mathcal{Z}$, the conditional distribution of $\widehat{\epsilon}_{t+\ell}$ given $Z_t = z$ admits a density $f_{\widehat{\epsilon}}(\cdot \mid z)$ on $[q^\star_\tau(z) - \delta,\, q^\star_\tau(z) + \delta]$, satisfying

$$f_{\widehat{\epsilon}}(u \mid z) \geq c_f \;\text{ if }\; |u - q^\star_\tau(z)| \leq \delta, \quad z \in \mathcal{Z},\ \tau \in \mathcal{T}.$$

Bounded residuals provide an integrable envelope for the quantile loss, and local absolute continuity excludes atoms at the target quantiles.

Assumption A.3 (Model classes)**.** For each $\tau \in \mathcal{T}$, there exists a deterministic function class $\overline{\mathcal{Q}}_\tau = \{q_\theta : \theta \in \overline{\Theta}_\tau\}$ such that $\mathcal{Q}_{T,\tau} \subseteq \overline{\mathcal{Q}}_\tau$ for all T. The parameter space $\overline{\Theta}_\tau$ and $\mathcal{Z} = \mathrm{supp}(P_Z)$ are compact. The mapping $(\theta, z) \mapsto q_\theta(z)$ is jointly continuous. Moreover, there exists $B_Q < \infty$ bounding all candidate functions:

$$\sup_{\tau\in\mathcal{T}} \sup_{\eta\in\overline{\Theta}_\tau} \sup_{z\in\mathcal{Z}} |q_\eta(z)| \leq B_Q.$$

Assumption A.3 holds, for example, when the context space is compact and all candidate models belong to a fixed finite neural architecture whose component maps are jointly continuous in their inputs and parameters, with parameters restricted to a common compact set. Under Assumption A.3, each $\overline{\mathcal{Q}}_\tau$ is compact under $\|q\|_\infty = \sup_{z\in\mathcal{Z}} |q(z)|$. Moreover, since $\rho_\tau(u) \leq |u|$, Assumptions A.2 and A.3 imply that, uniformly over $\tau \in \mathcal{T}$ and $q \in \overline{\mathcal{Q}}_\tau$,

$$\rho_\tau(\widehat{\epsilon}_{t+\ell} - q(Z_t)) \leq |\widehat{\epsilon}_{t+\ell}| + |q(Z_t)| \leq B_{\widehat{\epsilon}} + B_Q.$$

Assumption A.4 (Joint approximation)**.** The shared-encoder classes satisfy, as $T \to \infty$,

$$a_T^{\text{joint}} := \inf_{\mathbf{q} \in \mathfrak{Q}_T} \{\mathcal{R}(\mathbf{q}) - \mathcal{R}(\mathbf{q}^\star)\} \longrightarrow 0.$$

Assumption A.4 does not require $\mathbf{q}^\star \in \mathfrak{Q}_T$ for any finite T. It requires only that the model class contain jointly feasible quantile-function vectors whose joint population risk approaches that of the true conditional quantile vector as $T \to \infty$.

Assumption A.5 (Approximate empirical risk minimization)**.** The jointly trained estimator $\widehat{\mathbf{q}}_T \in \mathfrak{Q}_T$ satisfies

$$\widehat{\mathcal{R}}_T\left(\widehat{\mathbf{q}}_T\right) \leq \inf_{\mathbf{q} \in \mathfrak{Q}_T} \widehat{\mathcal{R}}_T(\mathbf{q}) + r_T,$$

where $r_T \geq 0$ and $r_T \xrightarrow{p} 0$.

Assumption A.5 is the standard approximate-extremum condition (Newey & McFadden, 1994). This requirement concerns the achieved objective value relative to the global empirical infimum, and the quantity r_T allows for optimization error in the model training.

Lemma A.6 (Uniform convergence of the empirical risk)**.** *Under Assumptions A.1, A.2(i), and A.3, for every $\tau \in \mathcal{T}$,*

$$\sup_{q \in \mathcal{Q}_{T,\tau}} \left| \widehat{R}_{T,\tau}(q) - R_\tau(q) \right| \xrightarrow{p} 0.$$

Proof of Lemma A.6. Fix $\tau \in \mathcal{T}$. Since $\mathcal{Q}_{T,\tau} \subseteq \overline{\mathcal{Q}}_\tau$, it suffices to establish uniform convergence over the fixed class $\overline{\mathcal{Q}}_\tau$. For $q \in \overline{\mathcal{Q}}_\tau$, write

$$L_{\tau,q}(W_i) = \rho_\tau\{\widehat{\epsilon}_{i+\ell} - q(Z_i)\}, \quad B_X = B_{\widehat{\epsilon}} + B_Q.$$

Then,

$$\widehat{R}_{T,\tau}(q) = \frac{1}{T}\sum_{i=1}^{T} L_{\tau,q}(W_i), \quad R_\tau(q) = \mathbb{E}\left[L_{\tau,q}(W_i)\right].$$

Since the quantile loss satisfies $0 \leq \rho_\tau(u) \leq |u|$, Assumptions A.2(i) and A.3 give

$$0 \leq L_{\tau,q}(W_i) \leq |\widehat{\epsilon}_{i+\ell}| + |q(Z_i)| \leq B_X.$$

By stationarity, the centered variables defined as

$$X_{i,q} = L_{\tau,q}(W_i) - R_\tau(q), \tag{9}$$

so that $\sum_{i=1}^{T} X_{i,q}/T = \widehat{R}_{T,\tau}(q) - R_\tau(q)$, satisfy $\mathbb{E}X_{i,q} = 0$ and $|X_{i,q}| \leq B_X$ almost surely. Since $X_{t,q}$ is a measurable transformation of W_t, by Assumption A.1, the process $\{X_{t,q}\}_{t\geq 1}$ is also strictly stationary and strongly mixing, with mixing coefficients bounded by $\alpha_W(k)$. Thus,

$$\sum_{i=1}^{T} \mathrm{Var}(X_{i,q}) \leq TB_X^2, \tag{10}$$

and the covariance inequality for strong mixing processes (Rio, 1993; 2017) gives

$$|\mathrm{Cov}(X_{i,q}, X_{i+k,q})| \leq 4B_X^2\alpha_W(k). \tag{11}$$

Consequently,

$$\begin{aligned} 2\sum_{1\leq i<j\leq T} \mathrm{Cov}(X_{i,q}, X_{j,q}) &= 2\sum_{k=1}^{T-1}\sum_{i=1}^{T-k} \mathrm{Cov}(X_{i,q}, X_{i+k,q}) \\ &\leq 8B_X^2 \sum_{k=1}^{T-1}(T-k)\alpha_W(k) \\ &\leq 8B_X^2 T\sum_{k=1}^{T-1}\alpha_W(k). \end{aligned} \tag{12}$$

Combining (10)–(12), since $\alpha_W(k) \to 0$ in Assumption A.1,

$$\begin{aligned}\mathrm{Var}\{\widehat{R}_{T,\tau}(q) - R_\tau(q)\} &= \frac{1}{T^2}\left\{\sum_{i=1}^{T}\mathrm{Var}(X_{i,q}) + 2\sum_{1\le i<j\le T}\mathrm{Cov}(X_{i,q}, X_{j,q})\right\} \\ &\le \frac{B_X^2}{T} + \frac{8B_X^2}{T}\sum_{k=1}^{T-1}\alpha_W(k) \longrightarrow 0.\end{aligned}$$

Indeed, for every $a > 0$, choose k_a such that $\alpha_W(k) \le a$ for $k \ge k_a$. Then, for $T > k_a$,

$$0 \le \frac{1}{T}\sum_{k=1}^{T-1}\alpha_W(k) \le \frac{1}{T}\sum_{k=1}^{k_a-1}\alpha_W(k) + a, \qquad \limsup_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T-1}\alpha_W(k) \le a.$$

Letting $a \to 0$ proves the required Cesàro convergence. Since $\mathbb{E}[\widehat{R}_{T,\tau}(q) - R_\tau(q)] = 0$, Chebyshev's inequality implies, for every fixed $q \in \overline{\mathcal{Q}}_\tau$,

$$\widehat{R}_{T,\tau}(q) - R_\tau(q) \xrightarrow{p} 0. \tag{13}$$

We now use this pointwise convergence to establish uniform convergence over $\overline{\mathcal{Q}}_\tau$. The quantile loss is 1-Lipschitz, meaning that

$$|\rho_\tau(u) - \rho_\tau(v)| \le |u - v|, \quad \forall u, v \in \mathbb{R}. \tag{14}$$

Hence, for $q, q' \in \overline{\mathcal{Q}}_\tau$,

$$\begin{aligned}|\widehat{R}_{T,\tau}(q) - \widehat{R}_{T,\tau}(q')| &\le \frac{1}{T}\sum_{i=1}^{T}|\rho_\tau\{\widehat{\epsilon}_{i+\ell} - q(Z_i)\} - \rho_\tau\{\widehat{\epsilon}_{i+\ell} - q'(Z_i)\}| \\ &\le \frac{1}{T}\sum_{i=1}^{T}|q(Z_i) - q'(Z_i)| \le \|q - q'\|_\infty, \qquad (15)\\ |R_\tau(q) - R_\tau(q')| &\le \mathbb{E}|q(Z_i) - q'(Z_i)| \le \|q - q'\|_\infty. \qquad (16)\end{aligned}$$

Under Assumption A.3, joint continuity on the compact set $\overline{\Theta}_\tau \times \mathcal{Z}$ implies

$$\theta_n \to \theta \quad \Longrightarrow \quad \sup_{z\in\mathcal{Z}}|q_{\theta_n}(z) - q_\theta(z)| \to 0.$$

Thus $\overline{\mathcal{Q}}_\tau$ is the continuous image of the compact parameter set in the uniform norm, and is compact. Given $\varepsilon > 0$, choose a finite net $q_1, \ldots, q_N \in \overline{\mathcal{Q}}_\tau$ such that for every $q \in \overline{\mathcal{Q}}_\tau$, $\|q - q_j\|_\infty < \varepsilon/4$ for some $j \in \{1, \ldots, N\}$. The triangle inequality and (15)–(16) yield

$$\sup_{q\in\overline{\mathcal{Q}}_\tau}|\widehat{R}_{T,\tau}(q) - R_\tau(q)| \le \frac{\varepsilon}{2} + \max_{1\le j\le N}|\widehat{R}_{T,\tau}(q_j) - R_\tau(q_j)|.$$

By (13) and the union bound,

$$\mathbb{P}\left(\sup_{q\in\overline{\mathcal{Q}}_\tau}|\widehat{R}_{T,\tau}(q) - R_\tau(q)| > \varepsilon\right) \le \sum_{j=1}^{N}\mathbb{P}\left(|\widehat{R}_{T,\tau}(q_j) - R_\tau(q_j)| > \frac{\varepsilon}{2}\right) \longrightarrow 0.$$

Taking the supremum over the smaller class $\mathcal{Q}_{T,\tau}$ proves the claim.

□

Corollary A.7 (Uniform convergence of the joint risk)**.** *Under the conditions of Lemma A.6,*

$$\sup_{\mathbf{q}\in\mathfrak{Q}_T}\left|\widehat{\mathcal{R}}_T(\mathbf{q}) - \mathcal{R}(\mathbf{q})\right| \xrightarrow{p} 0.$$

Proof of Corollary A.7. For every $\mathbf{q} \in \mathfrak{Q}_T$, each coordinate belongs to $\mathcal{Q}_{T,\tau}$. Therefore,

$$\begin{aligned}\sup_{\mathbf{q}\in\mathfrak{Q}_T} |\widehat{\mathcal{R}}_T(\mathbf{q}) - \mathcal{R}(\mathbf{q})| &= \sup_{\mathbf{q}\in\mathfrak{Q}_T} \left| \sum_{\tau\in\mathcal{T}} w_\tau \{\widehat{R}_{T,\tau}(q_\tau) - R_\tau(q_\tau)\} \right| \\ &\le \sum_{\tau\in\mathcal{T}} w_\tau \sup_{q\in\mathcal{Q}_{T,\tau}} |\widehat{R}_{T,\tau}(q) - R_\tau(q)| \xrightarrow{p} 0,\end{aligned}$$

where the last convergence follows from Lemma A.6 since $|\mathcal{T}| < \infty$ and the weights are fixed. □

Lemma A.8 (Population excess risk convergence)**.** *Under the conditions of Corollary A.7 and Assumptions A.4 and A.5,*

$$\mathcal{R}\left(\widehat{\mathbf{q}}_T\right) - \mathcal{R}(\mathbf{q}^\star) \xrightarrow{p} 0.$$

Proof of Lemma A.8. Write

$$\Delta_T = \sup_{\mathbf{q}\in\mathfrak{Q}_T} |\widehat{\mathcal{R}}_T(\mathbf{q}) - \mathcal{R}(\mathbf{q})| \xrightarrow{p} 0,$$

by Corollary A.7. By Assumption A.4, we can choose deterministic $\mathbf{q}_T^\circ \in \mathfrak{Q}_T$ such that

$$\mathcal{R}(\mathbf{q}_T^\circ) - \mathcal{R}(\mathbf{q}^\star) \le a_T^{\text{joint}} + T^{-1}. \tag{17}$$

The role of $\mathbf{q}_T^\circ$ is to provide a near-optimal population risk approximation to the true quantile vector $\mathbf{q}^\star$ within the model class $\mathfrak{Q}_T$. Decomposing the risk term gives

$$\begin{aligned}0 &\le \mathcal{R}(\widehat{\mathbf{q}}_T) - \mathcal{R}(\mathbf{q}^\star) \\ &= [\mathcal{R}(\widehat{\mathbf{q}}_T) - \widehat{\mathcal{R}}_T(\widehat{\mathbf{q}}_T)] + [\widehat{\mathcal{R}}_T(\widehat{\mathbf{q}}_T) - \widehat{\mathcal{R}}_T(\mathbf{q}_T^\circ)] + [\widehat{\mathcal{R}}_T(\mathbf{q}_T^\circ) - \mathcal{R}(\mathbf{q}_T^\circ)] + [\mathcal{R}(\mathbf{q}_T^\circ) - \mathcal{R}(\mathbf{q}^\star)] \\ &\le 2\Delta_T + r_T + a_T^{\text{joint}} + T^{-1} \xrightarrow{p} 0.\end{aligned}$$

The second term is at most r_T by Assumption A.5. The last convergence follows from Assumptions A.4 and A.5. □

Corollary A.9 (Coordinatewise population excess risk convergence)**.** *Under the conditions of Lemma A.8, for every $\tau \in \mathcal{T}$,*

$$R_\tau\left(\widehat{q}_{T,\tau}\right) - R_\tau\left(q_\tau^\star\right) \xrightarrow{p} 0.$$

Proof of Corollary A.9. For every $\tau \in \mathcal{T}$, by definition,

$$R_\tau(\widehat{q}_{T,\tau}) - R_\tau(q_\tau^\star) \ge 0.$$

Since $w_\tau > 0$ is fixed,

$$\begin{aligned}0 \le R_\tau(\widehat{q}_{T,\tau}) - R_\tau(q_\tau^\star) &\le \frac{1}{w_\tau} \sum_{\tau'\in\mathcal{T}} w_{\tau'} \{R_{\tau'}(\widehat{q}_{T,\tau'}) - R_{\tau'}(q_{\tau'}^\star)\} \\ &= \frac{\mathcal{R}(\widehat{\mathbf{q}}_T) - \mathcal{R}(\mathbf{q}^\star)}{w_\tau} \xrightarrow{p} 0,\end{aligned}$$

by Lemma A.8. □

Lemma A.10 (Uniform consistency of the estimated conditional quantile functions)**.** *Under the conditions of Corollary A.9 and Assumptions A.2–A.3,*

$$\max_{\tau\in\mathcal{T}} \sup_{z\in\mathcal{Z}} |\widehat{q}_{T,\tau}(z) - q_\tau^\star(z)| \xrightarrow{p} 0.$$

Proof of Lemma A.10. Fix $\tau \in \mathcal{T}$. For $q \in \overline{\mathcal{Q}}_\tau$, write

$$D_\tau(q) = R_\tau(q) - R_\tau(q_\tau^\star).$$

For $a \in \mathbb{R}$ and $z \in \mathcal{Z}$, the conditional excess quantile loss is

$$\begin{aligned}\Gamma_\tau(a, z) &= \mathbb{E}\left[\rho_\tau(\widehat{\epsilon}_{t+\ell} - a) - \rho_\tau\{\widehat{\epsilon}_{t+\ell} - q_\tau^\star(z)\} \mid Z_t = z\right] \\ &= \int_{q_\tau^\star(z)}^{a} \{F_{\widehat{\epsilon}}(u \mid z) - \tau\}\, du.\end{aligned}$$

Local absolute continuity in Assumption A.2(ii) gives

$$F_{\widehat{\epsilon}}\left(q_\tau^\star(z) \mid z\right) = \tau.$$

Suppose first $a > q_\tau^\star(z)$ and let $d = \min\{a - q_\tau^\star(z), \delta\}$, where δ is the neighborhood radius in Assumption A.2(ii). Monotonicity of the CDF and its local density lower bound imply

$$\begin{aligned}\Gamma_\tau(a,z) &\geq \int_{q_\tau^\star(z)}^{q_\tau^\star(z)+d} \left[F_{\widehat{\epsilon}}(u \mid z) - \tau\right] du \\ &= \int_{q_\tau^\star(z)}^{q_\tau^\star(z)+d} \int_{q_\tau^\star(z)}^{u} f_{\widehat{\epsilon}}(v \mid z) dv du \\ &\geq \int_{q_\tau^\star(z)}^{q_\tau^\star(z)+d} c_f \left(u - q_\tau^\star(z)\right) du = \frac{c_f d^2}{2}.\end{aligned}$$

If $a < q_\tau^\star(z)$, put $d = \min\{q_\tau^\star(z) - a, \delta\}$. The same argument gives

$$\begin{aligned}\Gamma_\tau(a,z) &= \int_{a}^{q_\tau^\star(z)} \left[\tau - F_{\widehat{\epsilon}}(u \mid z)\right] du \\ &\geq \int_{q_\tau^\star(z)-d}^{q_\tau^\star(z)} \int_{u}^{q_\tau^\star(z)} f_{\widehat{\epsilon}}(v \mid z) dv du \\ &\geq \int_{q_\tau^\star(z)-d}^{q_\tau^\star(z)} c_f \left(q_\tau^\star(z) - u\right) du = \frac{c_f d^2}{2}.\end{aligned}$$

The case $a = q_\tau^\star(z)$ gives zero on both sides. Thus, for $a \in \mathbb{R}$ and $z \in \mathcal{Z}$,

$$\Gamma_\tau(a,z) \geq \frac{c_f}{2} \min\left\{|a - q_\tau^\star(z)|, \delta\right\}^2. \tag{18}$$

In particular,

$$D_\tau(q) = \mathbb{E}\Gamma_\tau(q(Z_t), Z_t) \geq \frac{c_f}{2} \mathbb{E} \min\{|q(Z_t) - q_\tau^\star(Z_t)|, \delta\}^2 \geq 0,$$

and therefore

$$D_\tau(q) = 0 \quad \Longrightarrow \quad q(Z_t) = q_\tau^\star(Z_t) \ \text{ a.s.}$$

For $q \in \overline{\mathcal{Q}}_\tau$, both q and $q_\tau^\star$ are continuous by Assumptions A.3 and A.2(ii), respectively. If $q(z_0) \neq q_\tau^\star(z_0)$ at some $z_0 \in \mathcal{Z}$, continuity provides a relative neighborhood N of z_0 and a constant $c > 0$ with $|q(z) - q_\tau^\star(z)| \geq c$ on N. Since $\mathcal{Z} = \operatorname{supp}(P_Z)$, $P_Z(N) > 0$, and

$$D_\tau(q) \geq \frac{c_f}{2} \min\{c, \delta\}^2 P_Z(N) > 0.$$

Consequently,

$$D_\tau(q) = 0 \quad \Longrightarrow \quad q = q_\tau^\star \ \text{ on } \mathcal{Z}.$$

To translate population excess-risk convergence into uniform convergence of the estimated quantile function, we show that any candidate function that remains a fixed distance away from $q_\tau^\star$ in the uniform norm must incur a strictly positive excess population risk. Given $\varepsilon > 0$, define the compact set

$$\mathcal{A}_{\tau,\varepsilon} = \left\{q \in \overline{\mathcal{Q}}_\tau : \|q - q_\tau^\star\|_\infty \geq \varepsilon\right\}.$$

This set contains all candidate functions in $\overline{\mathcal{Q}}_\tau$ that are not uniformly close to $q_\tau^\star$. If $\mathcal{A}_{\tau,\varepsilon}$ is empty, then $\mathbb{P}\left(\|\hat{q}_{T,\tau} - q_\tau^\star\|_\infty \geq \varepsilon\right) = 0$. Otherwise, the inequality

$$|D_\tau(q) - D_\tau(q')| = |R_\tau(q) - R_\tau(q')| \leq \|q - q'\|_\infty$$

shows that D_τ is continuous. Since $D_\tau(q) > 0$ for every $q \in \mathcal{A}_{\tau,\varepsilon}$, continuity and compactness imply

$$c_{\tau,\varepsilon} = \inf_{q \in \mathcal{A}_{\tau,\varepsilon}} D_\tau(q) > 0.$$

Since $\widehat{q}_{T,\tau} \in \overline{\mathcal{Q}}_\tau$,

$$\mathbb{P}(\|\widehat{q}_{T,\tau} - q_\tau^\star\|_\infty \geq \varepsilon) \leq \mathbb{P}\{D_\tau(\widehat{q}_{T,\tau}) \geq c_{\tau,\varepsilon}\} \longrightarrow 0,$$

by Corollary A.9. Since $|\mathcal{T}| < \infty$, the union bound proves the claim. □

Theorem A.11 (Asymptotic conditional coverage)**.** *Suppose Assumptions A.1–A.5 hold. Fix $\alpha_l, \alpha_h \in \mathcal{T}$ with $\alpha_h - \alpha_l = 1 - \alpha$. Then, for any deterministic sequence of $t_T \geq T + \ell$, as $T \to \infty$,*

$$\left|\mathbb{P}\left(Y_{t_T+\ell} \in \widehat{C}_T^{\alpha,\ell}(Z_{t_T}, \widehat{f}(Z'_{t_T})) \mid Z_{t_T}\right) - (1-\alpha)\right| \xrightarrow{p} 0.$$

Proof of Theorem A.11. For the fixed levels α_l, α_h in the theorem, define the oracle interval

$$C^{\star,\alpha,\ell}(z, \widehat{y}) = [\widehat{y} + q^\star_{\alpha_l}(z), \widehat{y} + q^\star_{\alpha_h}(z)].$$

Local absolute continuity in Assumption A.2 at the two quantiles implies

$$F_{\widehat{\epsilon}}(q^\star_\tau(z) \mid z) = F_{\widehat{\epsilon}}(q^\star_\tau(z)- \mid z) = \tau, \quad \tau \in \{\alpha_l, \alpha_h\}. \tag{19}$$

With the same point prediction $\widehat{Y}_{t_T+\ell} = \widehat{f}(Z'_{t_T})$ in both estimated and oracle intervals, define the coverage events

$$\widehat{A}_T = \left\{Y_{t_T+\ell} \in \widehat{C}_T^{\alpha,\ell}(Z_{t_T}, \widehat{Y}_{t_T+\ell})\right\} = \{\widehat{q}_{T,\alpha_l}(Z_{t_T}) \leq \widehat{\epsilon}_{t_T+\ell} \leq \widehat{q}_{T,\alpha_h}(Z_{t_T})\},$$
$$A^\star_T = \left\{Y_{t_T+\ell} \in C^{\star,\alpha,\ell}(Z_{t_T}, \widehat{Y}_{t_T+\ell})\right\} = \{q^\star_{\alpha_l}(Z_{t_T}) \leq \widehat{\epsilon}_{t_T+\ell} \leq q^\star_{\alpha_h}(Z_{t_T})\}.$$

Stationarity and (19) give

$$\begin{aligned}\mathbb{P}(A^\star_T \mid Z_{t_T}) &= F_{\widehat{\epsilon}}(q^\star_{\alpha_h}(Z_{t_T}) \mid Z_{t_T}) - F_{\widehat{\epsilon}}(q^\star_{\alpha_l}(Z_{t_T})- \mid Z_{t_T}) \\ &= \alpha_h - \alpha_l = 1 - \alpha \quad \text{a.s.}\end{aligned}$$

Set

$$d_T = \max_{\tau \in \{\alpha_l, \alpha_h\}} \|\widehat{q}_{T,\tau} - q^\star_\tau\|_\infty \xrightarrow{p} 0$$

by Lemma A.10. Given $\eta \in (0, \delta)$, on the event $\{d_T \leq \eta\}$, both estimated quantile endpoints differ from their oracle counterparts by at most η. Hence, the estimated and oracle coverage events can differ only if the residual lies within η of one of the two oracle endpoints. Therefore,

$$\widehat{A}_T \triangle A^\star_T \subseteq \{d_T > \eta\} \cup \bigcup_{\tau \in \{\alpha_l, \alpha_h\}} \{|\widehat{\epsilon}_{t_T+\ell} - q^\star_\tau(Z_{t_T})| \leq \eta\},$$

where $\triangle$ denotes the symmetric difference. Thus union bound gives

$$\mathbb{P}(\widehat{A}_T \triangle A^\star_T) \leq \mathbb{P}(d_T > \eta) + \sum_{\tau \in \{\alpha_l, \alpha_h\}} \mathbb{P}\{|\widehat{\epsilon}_{t_T+\ell} - q^\star_\tau(Z_{t_T})| \leq \eta\}. \tag{20}$$

For either endpoint, stationarity gives the bound independent of T as

$$\mathbb{P}\{|\widehat{\epsilon}_{t_T+\ell} - q^\star_\tau(Z_{t_T})| \leq \eta\} = \mathbb{P}\{|\widehat{\epsilon}_{1+\ell} - q^\star_\tau(Z_1)| \leq \eta\} =: b_\tau(\eta). \tag{21}$$

By continuity from above of probability and local absolute continuity in Assumption A.2(ii),

$$\lim_{\eta \downarrow 0} b_\tau(\eta) = \mathbb{P}\{\widehat{\epsilon}_{1+\ell} = q^\star_\tau(Z_1)\} = \mathbb{E}\left[F_{\widehat{\epsilon}}(q^\star_\tau(Z_1) \mid Z_1) - F_{\widehat{\epsilon}}(q^\star_\tau(Z_1)- \mid Z_1)\right] = 0.$$

Combining (20) and (21),

$$\limsup_{T\to\infty} \mathbb{P}(\widehat{A}_T \triangle A^\star_T) \leq \limsup_{T\to\infty} \mathbb{P}(d_T > \eta) + \sum_{\tau \in \{\alpha_l, \alpha_h\}} b_\tau(\eta) = \sum_{\tau \in \{\alpha_l, \alpha_h\}} b_\tau(\eta). \tag{22}$$

Letting $\eta \downarrow 0$ yields $\mathbb{P}(\widehat{A}_T \triangle A^\star_T) \to 0$. The exact oracle coverage now gives

$$\begin{aligned}|\mathbb{P}(\widehat{A}_T \mid Z_{t_T}) - (1-\alpha)| &= |\mathbb{E}[\mathbb{1}\{\widehat{A}_T\} - \mathbb{1}\{A^\star_T\} \mid Z_{t_T}]| \\ &\leq \mathbb{E}[|\mathbb{1}\{\widehat{A}_T\} - \mathbb{1}\{A^\star_T\}| \mid Z_{t_T}] \\ &= \mathbb{P}(\widehat{A}_T \triangle A^\star_T \mid Z_{t_T}).\end{aligned}$$

Taking expectations and applying Markov's inequality gives

$$\begin{aligned}&\mathbb{P}\left(|\mathbb{P}(\widehat{A}_T \mid Z_{t_T}) - (1-\alpha)| > \varepsilon\right) \\ &\leq \frac{1}{\varepsilon}\mathbb{E}|\mathbb{P}(\widehat{A}_T \mid Z_{t_T}) - (1-\alpha)| \leq \frac{\mathbb{P}(\widehat{A}_T \triangle A^\star_T)}{\varepsilon} \longrightarrow 0,\end{aligned}$$

for every $\varepsilon > 0$. This proves the claim. □

A.2 DSCP-QFE

DSCP-QFE estimates one function of the context and the quantile level. Unless otherwise mentioned, we retain the notation of the preceding subsection.

Let $\mathcal{I} = [\underline{\tau}, \overline{\tau}] \subset (0,1)$ be a compact interval of admissible quantile levels. While the inference can be conducted at any quantile levels in $(0,1)$, we introduce $\mathcal{I}$ for theoretical analysis.

For $(z,\tau) \in \mathcal{Z} \times \mathcal{I}$, define the true quantile function

$$q^\star(z,\tau) := q^\star_\tau(z) = \inf\{u \in \mathbb{R} : F_{\hat{\epsilon}}(u \mid z) \geq \tau\}.$$

For a generic quantile function, $q : \mathcal{Z} \times \mathcal{I} \to \mathbb{R}$, we write $q(z,\tau) = q_\tau(z)$. The estimated function is denoted by $\widehat{q}_T$, and $\widehat{q}_{T,\tau}(z) = \widehat{q}_T(z,\tau)$. Thus, the notation for each estimated conditional quantile agrees with that used for conditional quantile regression. The architecture is represented by

$$q_\theta(z,\tau) = g_{\theta''}\left(\mathrm{Enc}_{\theta'}(z), \tau\right), \quad \theta = (\theta', \theta''),$$

Let $\mathfrak{Q}_T$ denote the class of jointly feasible candidate functions, so that $\widehat{q}_T \in \mathfrak{Q}_T$.

Let ν be a probability distribution on $\mathcal{I}$. The integrated population risk is

$$\mathcal{R}(q) = \int_{\mathcal{I}} R_\tau(q(\cdot,\tau))\nu(d\tau) = \mathbb{E}\left[\rho_U\{\hat{\epsilon}_{t+\ell} - q(Z_t, U)\}\right], \tag{23}$$

where $U \sim \nu$ is independent of the process $\{(Z_t, \hat{\epsilon}_{t+\ell})\}_{t\geq 1}$. The true conditional quantile function minimizes this risk over measurable candidates for which the risk is finite.

For independently sampled quantile levels $U_1, \ldots, U_T$, the randomized empirical risk is

$$\widehat{\mathcal{R}}_T(q) = \frac{1}{T}\sum_{t=1}^{T} \rho_{U_t}\{\widehat{\epsilon}_{t+\ell} - q(Z_t, U_t)\}, \tag{24}$$

which uses one sampled quantile level per sample. A fixed number of levels per pair can be treated by averaging their losses. If levels are resampled during optimization, the corresponding integrated empirical objective is

$$\widetilde{\mathcal{R}}_T(q) = \frac{1}{T}\sum_{t=1}^{T} \int_{\mathcal{I}} \rho_\tau\{\widehat{\epsilon}_{t+\ell} - q(Z_t, \tau)\}\nu(d\tau). \tag{25}$$

Assumption A.12 (Quantile-level sampling)**.** The variables U_t, $t \geq 1$, are independent and identically distributed with law ν, independent of the process $\{W_t\}_{t\geq 1}$. The probability law ν has full support on $\mathcal{I}$, i.e., $\mathrm{supp}(\nu) = \mathcal{I}$.

Full support is used to identify the target at every level by continuity.

Assumption A.13 (Residuals and conditional quantile regularity)**.** The following conditions hold.

(i) (Bounded residuals) There exists a constant $B_{\widehat{\epsilon}} < \infty$ such that $|\widehat{\epsilon}_{t+\ell}| \leq B_{\widehat{\epsilon}}$ almost surely for all $t \geq 1$.

(ii) (Regularity around the conditional quantiles) For each $(z,\tau) \in \mathcal{Z} \times \mathcal{I}$, the true conditional τ-quantile $q^\star(z,\tau)$ is unique and jointly continuous on $\mathcal{Z} \times \mathcal{I}$. Moreover, there exist constants $c_f > 0$ and $\delta > 0$ such that, for every $\tau \in \mathcal{I}$ and every $z \in \mathcal{Z}$, the conditional distribution of $\widehat{\epsilon}_{t+\ell}$ given $Z_t = z$ admits a density $f_{\widehat{\epsilon}}(\cdot \mid z)$ on $[q^\star_\tau(z) - \delta,\, q^\star_\tau(z) + \delta]$, satisfying

$$f_{\widehat{\epsilon}}(u \mid z) \geq c_f \;\text{ if }\; |u - q^\star_\tau(z)| \leq \delta, \quad z \in \mathcal{Z},\ \tau \in \mathcal{I}.$$

Assumption A.14 (Model classes)**.** The classes $\mathfrak{Q}_T$ are deterministic, nonempty, and contained in a fixed class

$$\overline{\mathfrak{Q}} = \{q_\theta : \theta \in \overline{\Theta}\} \subseteq C(\mathcal{Z} \times \mathcal{I}).$$

The parameter space $\overline{\Theta}$ and $\mathcal{Z} = \mathrm{supp}(P_Z)$ are compact. The mapping $(\theta, z, \tau) \mapsto q_\theta(z,\tau)$ is jointly continuous. Moreover, there exists $B_Q < \infty$ such that

$$\sup_{\theta\in\overline{\Theta}} \sup_{(z,\tau)\in\mathcal{Z}\times\mathcal{I}} |q_\theta(z,\tau)| \leq B_Q.$$

Assumption A.15 (Joint approximation)**.** The integrated approximation error satisfies, as $T \to \infty$,

$$a_T := \inf_{q \in \mathfrak{Q}_T} \{\mathcal{R}(q) - \mathcal{R}(q^\star)\} \longrightarrow 0.$$

Assumption A.16 (Approximate empirical risk minimization)**.** The estimator $\widehat{q}_T \in \mathfrak{Q}_T$ satisfies

$$\widehat{\mathcal{R}}_T(\widehat{q}_T) \leq \inf_{q \in \mathfrak{Q}_T} \widehat{\mathcal{R}}_T(q) + r_T,$$

where $r_T \geq 0$ and $r_T \xrightarrow{p} 0$.

The error r_T accommodates optimization and algorithmic randomness. Exact empirical-risk minimization is the special case $r_T = 0$. If the integrated empirical objective (25) is optimized instead, impose this same inequality with $\widetilde{\mathcal{R}}_T$ in place of $\widehat{\mathcal{R}}_T$. Then, the subsequent results hold under that version of the assumption.

Lemma A.17 (Uniform convergence of the empirical risk)**.** *Under Assumptions A.1, A.12, A.13(i), and A.14,*

$$\sup_{q \in \mathfrak{Q}_T} \left| \widehat{\mathcal{R}}_T(q) - \mathcal{R}(q) \right| \xrightarrow{p} 0.$$

The same conclusion holds for $\widetilde{\mathcal{R}}_T$.

Proof of Lemma A.17. Define the concatenated process

$$W_t^U := (W_t, U_t) = (Z_t, \widehat{\epsilon}_{t+\ell}, U_t). \tag{26}$$

By Assumption A.1, $\{W_t\}_{t \geq 1}$ is strictly stationary and strongly mixing. Moreover, by Assumption A.12, $\{U_t\}_{t \geq 1}$ is i.i.d. and independent of the process $\{W_t\}_{t \geq 1}$. Therefore, $\{W_t^U\}_{t \geq 1}$ is strictly stationary and strongly mixing, and by Bradley (2005, Theorem 5.1(a)), its mixing coefficients satisfy

$$\alpha_{W^U}(k) \leq \alpha_W(k), \quad k \geq 1. \tag{27}$$

In particular,

$$0 \leq \alpha_{W^U}(k) \leq \alpha_W(k) \longrightarrow 0.$$

For $q \in \overline{\mathfrak{Q}}$, set $L_q(W_t^U) = \rho_{U_t}\{\widehat{\epsilon}_{t+\ell} - q(Z_t, U_t)\}$ and $B_X = B_{\widehat{\epsilon}} + B_Q$. Throughout the proof, the uniform norm is taken over $\mathcal{Z} \times \mathcal{I}$. Independence of U_t and W_t, Assumptions A.13(i) and A.14 and (14) give

$$\mathbb{E}L_q(W_t^U) = \int_{\mathcal{I}} R_\tau(q(\cdot, \tau))\nu(d\tau) = \mathcal{R}(q), \quad 0 \leq L_q(W_t^U) \leq B_X,$$
$$|L_q(W_t^U) - L_{q'}(W_t^U)| \leq |q(Z_t, U_t) - q'(Z_t, U_t)| \leq \|q - q'\|_\infty.$$

Joint continuity in Assumption A.14 makes $\theta \mapsto q_\theta$ continuous in this norm. Thus, $\overline{\mathfrak{Q}}$ is compact as the continuous image. Applying the arguments of Lemma A.6 with W_t^U, L_q, and $\overline{\mathfrak{Q}}$ in place of W_t, $L_{\tau,q}$, and $\overline{\mathcal{Q}}_\tau$, respectively, gives the conclusion.

For the integrated objective, instead set

$$\widetilde{L}_q(W_t) = \int_{\mathcal{I}} \rho_\tau\{\widehat{\epsilon}_{t+\ell} - q(Z_t, \tau)\}\nu(d\tau), \quad \widetilde{\mathcal{R}}_T(q) = \frac{1}{T}\sum_{t=1}^{T} \widetilde{L}_q(W_t).$$

Since ν is a probability law, Tonelli's theorem and the same loss bounds imply

$$\mathbb{E}\widetilde{L}_q(W_t) = \mathcal{R}(q), \quad 0 \leq \widetilde{L}_q(W_t) \leq B_X, \quad |\widetilde{L}_q(W_t) - \widetilde{L}_{q'}(W_t)| \leq \|q - q'\|_\infty.$$

The same proof of Lemma A.6 now applies directly to W_t and $\widetilde{L}_q$, and yields the uniform convergence for $\widetilde{\mathcal{R}}_T$. □

Lemma A.18 (Population excess risk convergence)**.** *Under the conditions of Lemma A.17 and Assumptions A.15 and A.16,*

$$\mathcal{R}(\widehat{q}_T) - \mathcal{R}(q^\star) \xrightarrow{p} 0.$$

Proof of Lemma A.18. By Lemma A.17,

$$\Delta_T \coloneqq \sup_{q \in \mathfrak{Q}_T} |\widehat{\mathcal{R}}_T(q) - \mathcal{R}(q)| \xrightarrow{p} 0.$$

Choose a deterministic $q_T^\circ \in \mathfrak{Q}_T$ with $\mathcal{R}(q_T^\circ) - \mathcal{R}(q^\star) \le a_T + T^{-1}$, which exists by Assumption A.15. The decomposition argument in the proof of Lemma A.8 applies with the present integrated risks and a_T in place of a_T^{joint}. Consequently,

$$0 \le \mathcal{R}(\widehat{q}_T) - \mathcal{R}(q^\star) \le 2\Delta_T + r_T + a_T + T^{-1} \xrightarrow{p} 0,$$

by Assumptions A.15 and A.16. For the integrated empirical objective, replace $\widehat{\mathcal{R}}_T$ by $\widetilde{\mathcal{R}}_T$ throughout, using the corresponding version of Assumption A.16. □

Lemma A.19 (Uniform consistency of the estimated conditional quantile function)**.** *Under the conditions of Lemma A.18 and Assumptions A.12–A.14,*

$$\sup_{z \in \mathcal{Z}} \sup_{\tau \in \mathcal{I}} |\widehat{q}_T(z,\tau) - q^\star(z,\tau)| \xrightarrow{p} 0.$$

Proof of Lemma A.19. Write $D(q) = \mathcal{R}(q) - \mathcal{R}(q^\star)$ for $q \in \overline{\mathfrak{Q}}$. Assumption A.13(ii) allows the lower bound (18) from the proof of Lemma A.10 to be applied at every $\tau \in \mathcal{I}$, with $q_\tau^\star(z) = q^\star(z,\tau)$. Integrating that bound gives

$$D(q) \ge \frac{c_f}{2} \int_{\mathcal{I}} \int_{\mathcal{Z}} \min\{|q(z,\tau) - q^\star(z,\tau)|, \delta\}^2 \, P_Z(dz)\nu(d\tau) \ge 0. \tag{28}$$

If $q(z_0,\tau_0) \ne q^\star(z_0,\tau_0)$, joint continuity gives relative open neighborhoods N_z and N_τ of z_0 and τ_0, and a constant $c > 0$, such that $|q - q^\star| \ge c$ on $N_z \times N_\tau$. Since $\mathcal{Z} = \operatorname{supp}(P_Z)$ and $\mathcal{I} = \operatorname{supp}(\nu)$,

$$P_Z(N_z) > 0, \quad \nu(N_\tau) > 0, \quad D(q) \ge \frac{c_f}{2} \min\{c,\delta\}^2 P_Z(N_z)\nu(N_\tau) > 0.$$

Consequently,

$$D(q) = 0 \implies q = q^\star \text{ on } \mathcal{Z} \times \mathcal{I}.$$

The class $\overline{\mathfrak{Q}}$ is compact, as established in the proof of Lemma A.17, and D is continuous because

$$|D(q) - D(q')| = |\mathcal{R}(q) - \mathcal{R}(q')| \le \|q - q'\|_\infty.$$

The compact-separation argument in the proof of Lemma A.10 therefore applies with $\mathcal{Z} \times \mathcal{I}$, $\overline{\mathfrak{Q}}$, D, and $q^\star$ in place of $\mathcal{Z}$, $\overline{\mathcal{Q}}_\tau$, D_τ, and $q_\tau^\star$. Together with $D(\widehat{q}_T) \xrightarrow{p} 0$ from Lemma A.18, this gives $\|\widehat{q}_T - q^\star\|_\infty \xrightarrow{p} 0$ on the product domain. □

Theorem A.20 (Conditional coverage for quantile function estimation)**.** *Suppose Assumption A.1 and Assumptions A.12–A.16 hold. Fix $\alpha_l, \alpha_h \in \mathcal{I}$ with $\alpha_h - \alpha_l = 1 - \alpha$. Then, for any deterministic sequence of $t_T \ge T + \ell$, as $T \to \infty$,*

$$\left| \mathbb{P}\left(Y_{t_T+\ell} \in \widehat{C}_T^{\alpha,\ell}(Z_{t_T}, \widehat{f}(Z'_{t_T})) \mid Z_{t_T} \right) - (1-\alpha) \right| \xrightarrow{p} 0.$$

Proof of Theorem A.20. For the fixed levels $\tau \in \{\alpha_l, \alpha_h\}$ in the theorem, set $\widehat{q}_{T,\tau}(z) = \widehat{q}_T(z,\tau)$ and $q_\tau^\star(z) = q^\star(z,\tau)$. Lemma A.19 gives

$$d_T \coloneqq \max_{\tau \in \{\alpha_l,\alpha_h\}} \sup_{z \in \mathcal{Z}} |\widehat{q}_{T,\tau}(z) - q_\tau^\star(z)| \le \|\widehat{q}_T - q^\star\|_\infty \xrightarrow{p} 0.$$

Local absolute continuity in Assumption A.13(ii) at the two quantiles implies

$$F_{\widehat{\epsilon}}(q_\tau^\star(z) \mid z) = F_{\widehat{\epsilon}}(q_\tau^\star(z) - \mid z) = \tau, \quad \tau \in \{\alpha_l, \alpha_h\}.$$

Together with stationarity of W_t, these verify all ingredients of the coverage argument in Theorem A.11. Applying that proof to the present endpoint functions and the same point prediction $\widehat{f}(Z'_{t_T})$ yields the claimed conclusion. □

A.3 DSCP-LOCAL

The localized estimator first targets the residual distribution conditional on $H_t = \mathrm{Enc}_{\theta'}(Z_t)$. We establish consistency at a fixed representation value, transfer it to the observed representation, and then obtain conditional coverage. Coverage conditional on Z_t is obtained under an additional residual-sufficiency assumption. Throughout this subsection, the encoder is fixed, and its representation space is a Borel subset of a Euclidean space equipped with a jointly measurable metric d_H.

A.3.1 COVERAGE CONDITIONAL ON THE REPRESENTATION H_t

Assumption A.21 (Stationarity and absolute regularity)**.** The process $V_t = (H_t, \hat{\epsilon}_{t+\ell})$, $t \geq 1$, is strictly stationary and β-mixing with its mixing coefficient

$$\beta_V(k) \longrightarrow 0.$$

Assumption A.22 (Kernel and local mass)**.** The measurable kernel satisfies $0 \leq K \leq K_{\max} < \infty$, $K(v) = 0$ for $v > 1$, and $K(v) \geq K_{\min} > 0$ for $0 \leq v \leq a_0$, where $a_0 \in (0, 1]$. There is a common $d_0 > 0$ such that, for P_H-almost every h, constants $0 < c_h < C_h < \infty$ satisfy

$$\mathbb{P}(d_H(H, h) \leq a_0 b_T) \geq c_h b_T^{d_0}, \quad \mathbb{P}(d_H(H, h) \leq b_T) \leq C_h b_T^{d_0}, \tag{29}$$

for all sufficiently large T, where $H \sim P_H$ is a generic representation.

Several kernels, including uniform, triangular, Epanechnikov, as well as the truncated Gaussian kernel satisfy the kernel conditions in Assumption A.22.

Assumption A.23 (Local CDF regularity)**.** For P_H-almost every h, $F^H_{\widehat{\epsilon}}(\cdot \mid h)$ is a continuous proper CDF and

$$\omega_T(h) \coloneqq \operatorname*{ess\,sup}_{h' :\, d_H(h', h) \leq b_T} \sup_{u \in \mathbb{R}} |F^H_{\widehat{\epsilon}}(u \mid h') - F^H_{\widehat{\epsilon}}(u \mid h)| \to 0. \tag{30}$$

The essential supremum is with respect to P_H.

The quantity $\omega_T(h)$ controls the bias from pooling nearby representations. A sufficient condition is local Hölder continuity of the conditional CDF in the representation metric, uniformly in u, which gives $\omega_T(h) \leq L_h b_T^{\gamma}$ for some $\gamma > 0$. Continuity in u excludes atoms and permits inversion at neighboring probability levels.

For example, let $\widehat{\epsilon}_{t+\ell} = m(H_t) + \xi_{t+\ell}$, where $\xi_{t+\ell}$ is independent of H_t and has a CDF G with density bounded by L_G. Suppose $|m(h') - m(h)| \leq L_{m,h} d_H(h', h)^{\gamma}$ locally at P_H-almost every h, for some $\gamma > 0$. Then $F^H_{\widehat{\epsilon}}(u \mid h) = G(u - m(h))$ is a continuous proper CDF and

$$\sup_u |F^H_{\widehat{\epsilon}}(u \mid h') - F^H_{\widehat{\epsilon}}(u \mid h)| \leq L_G L_{m,h} d_H(h', h)^{\gamma},$$

so $\omega_T(h) \leq L_G L_{m,h} b_T^{\gamma} \to 0$ whenever $b_T \to 0$.

Assumption A.24 (Bandwidth and dependence structure)**.** Let $p_T = b_T^{d_0}$ and

$$\Lambda_T = p_T + \sum_{k=1}^{T-1} \min\{p_T, \beta_V(k)\}. \tag{31}$$

Assume $b_T \to 0$ and

$$\kappa_T^{\beta} \coloneqq \frac{\Lambda_T}{T p_T^2} \to 0. \tag{32}$$

Lemma A.25 (Conditional CDF consistency at a fixed representation)**.** *Assume A.21–A.23, $b_T \to 0$, and the variance condition* (32)*. Let M_T be deterministic integers with $1 \leq M_T \leq T$ and $M_T/T \to 1$. Write*

$$S_{T,M}(h) = \sum_{i=1}^{M} K_{T,i}(h), \qquad B_{T,M}(u, h) = \sum_{i=1}^{M} K_{T,i}(h) \mathbb{1}\{\widehat{\epsilon}_{i+\ell} \leq u\},$$
$$\mu_T(h) = \mathbb{E}K(d_H(H, h)/b_T), \qquad \widehat{F}^H_{T,M}(u \mid h) = B_{T,M}(u, h)/S_{T,M}(h),$$

if $S_{T,M}(h) > 0$. When $S_{T,M}(h) = 0$, set $\widehat{F}^H_{T,M}(u \mid h) = \mathbb{1}\{u \geq 0\}$. Then, for P_H-almost every fixed h,

$$\frac{S_{T,M_T}(h)}{M_T \mu_T(h)} \xrightarrow{p} 1, \qquad \sup_u |\widehat{F}^H_{T,M_T}(u \mid h) - F^H_{\widehat{\epsilon}}(u \mid h)| \xrightarrow{p} 0. \tag{33}$$

Proof of Lemma A.25. Fix h for which (29) and (30) hold. In this proof, C denotes a finite positive constant that may change depending on h but not on T, M, or u. For all sufficiently large T, the kernel bounds (29) give

$$K_{\min} c_h p_T \leq K_{\min} \mathbb{P}(d_H(H,h) \leq a_0 b_T) \leq \mu_T(h) \leq K_{\max} \mathbb{P}(d_H(H,h) \leq b_T) \leq K_{\max} C_h p_T. \tag{34}$$

For either $X_i = K_{T,i}(h)$ or $X_i = K_{T,i}(h)\mathbb{1}\{\widehat{\epsilon}_{i+\ell} \leq u\}$,

$$\begin{aligned} \mathrm{Var}(X_i) &\leq \mathbb{E}X_i^2 \leq K^2_{\max} \mathbb{P}(d_H(H,h) \leq b_T) \leq K^2_{\max} C_h p_T, \\ |\mathrm{Cov}(X_i, X_{i+k})| &\leq \mathbb{E}(X_i X_{i+k}) + \mathbb{E}X_i \mathbb{E}X_{i+k} \leq 2K^2_{\max} \mathbb{P}(d_H(H,h) \leq b_T) \leq 2K^2_{\max} C_h p_T. \end{aligned}$$

Also, from the coupling argument of Dedecker et al. (2006, Corollary 1), we have $|\mathrm{Cov}(X_i, X_{i+k})| \leq K^2_{\max} \beta_V(k)$. Therefore,

$$|\mathrm{Cov}(X_i, X_{i+k})| \leq K^2_{\max} \min\{2C_h p_T, \beta_V(k)\} \leq C \min\{p_T, \beta_V(k)\},$$

and hence we can bound the variance term as

$$\begin{aligned} \mathrm{Var}\left(\sum_{i=1}^M X_i\right) &= M\mathrm{Var}(X_1) + 2\sum_{k=1}^{M-1} (M-k)\mathrm{Cov}(X_1, X_{1+k}) \\ &\leq CM\left[p_T + \sum_{k=1}^{T-1} \min\{p_T, \beta_V(k)\}\right] = CM\Lambda_T. \end{aligned} \tag{35}$$

To bound the bias term, define the population localized CDF by

$$\overline{F}^H_T(u \mid h) = \frac{\mathbb{E}[K(d_H(H,h)/b_T) F^H_{\widehat{\epsilon}}(u \mid H)]}{\mu_T(h)}.$$

Assumption A.23 gives

$$\left|\overline{F}^H_T(u \mid h) - F^H_{\widehat{\epsilon}}(u \mid h)\right| = \frac{\left|\mathbb{E}[K(d_H(H,h)/b_T)\{F^H_{\widehat{\epsilon}}(u \mid H) - F^H_{\widehat{\epsilon}}(u \mid h)\}]\right|}{\mu_T(h)} \leq \omega_T(h).$$

In particular,

$$\sup_u \left|\overline{F}^H_T(u \mid h) - F^H_{\widehat{\epsilon}}(u \mid h)\right| \leq \omega_T(h) \to 0. \tag{36}$$

Since $\mathbb{E}S_{T,M}(h) = M\mu_T(h)$ and

$$\mathbb{E}B_{T,M}(u,h) = \sum_{i=1}^M \mathbb{E}\left[K_{T,i}(h)\mathbb{E}\{\mathbb{1}(\widehat{\epsilon}_{i+\ell} \leq u) \mid H_i\}\right] = M\mu_T(h)\overline{F}^H_T(u \mid h),$$

Chebyshev's inequality gives, for every $v > 0$,

$$\begin{aligned} \mathbb{P}\left(\left|\frac{S_{T,M}(h)}{M\mu_T(h)} - 1\right| > v\right) &= \mathbb{P}(|S_{T,M}(h) - M\mu_T(h)| > vM\mu_T(h)) \\ &\leq \frac{CM\Lambda_T}{v^2 M^2 \mu_T(h)^2} \leq \frac{C\kappa_T^\beta}{v^2} \to 0, \\ \mathbb{P}\left(\left|B_{T,M}(u,h) - M\mu_T(h)\overline{F}^H_T(u \mid h)\right| > vM\mu_T(h)\right) & \\ \leq \frac{CM\Lambda_T}{v^2 M^2 \mu_T(h)^2} \leq \frac{C\kappa_T^\beta}{v^2} \to 0. & \end{aligned}$$

The first bound proves the first conclusion of (33). For $0 < v < 1$, on the event where the following inequality,

$$\max\left\{\left|S_{T,M}(h) - M\mu_T(h)\right|, \left|B_{T,M}(u,h) - M\mu_T(h)\overline{F}_T^H(u \mid h)\right|\right\} \le \frac{vM\mu_T(h)}{4},$$

holds, we have $S_{T,M}(h) > 0$ and

$$\begin{aligned}\left|\widehat{F}_{T,M}^H(u \mid h) - \overline{F}_T^H(u \mid h)\right| &= \frac{\left|B_{T,M}(u,h) - S_{T,M}(h)\overline{F}_T^H(u \mid h)\right|}{S_{T,M}(h)} \\ &\le \frac{2}{M\mu_T(h)}\left|B_{T,M}(u,h) - M\mu_T(h)\overline{F}_T^H(u \mid h)\right| \\ &\quad + \frac{2}{M\mu_T(h)}\left|M\mu_T(h) - S_{T,M}(h)\right| \\ &\le v.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{P}\left(\left|\widehat{F}_{T,M}^H(u \mid h) - F_{\widehat{\epsilon}}^H(u \mid h)\right| > \omega_T(h) + v\right) \\ &\le \mathbb{P}\left(\left|S_{T,M}(h) - M\mu_T(h)\right| > \frac{vM\mu_T(h)}{4}\right) \\ &\quad + \mathbb{P}\left(\left|B_{T,M}(u,h) - M\mu_T(h)\overline{F}_T^H(u \mid h)\right| > \frac{vM\mu_T(h)}{4}\right) \\ &\le C\kappa_T^\beta / v^2. \end{aligned} \tag{37}$$

Now we use the finite-grid argument of Van der Vaart (2000, Lemma 2.11) for the uniformity result. Fix $m \ge 2$ and abbreviate $F = F_{\widehat{\epsilon}}^H(\cdot \mid h)$. Take $x_j = F^{-1}(j/m)$ for $j = 1, \dots, m-1$, and $x_0 = -\infty, x_m = +\infty$. For any proper CDF G and $x_j \le u < x_{j+1}$,

$$\begin{aligned}G(u) - F(u) &\le G(x_{j+1}) - F(x_j) \le \max_{0\le k\le m}|G(x_k) - F(x_k)| + m^{-1}, \\ F(u) - G(u) &\le F(x_{j+1}) - G(x_j) \le \max_{0\le k\le m}|G(x_k) - F(x_k)| + m^{-1}, \\ \|G - F\|_\infty &\le \max_{0\le k\le m}|G(x_k) - F(x_k)| + m^{-1}.\end{aligned}$$

Applying this to $G = \widehat{F}_{T,M}^H(\cdot \mid h)$ and using (37),

$$\begin{aligned}&\mathbb{P}\left(\left\|\widehat{F}_{T,M}^H(\cdot \mid h) - F_{\widehat{\epsilon}}^H(\cdot \mid h)\right\|_\infty > \omega_T(h) + v + m^{-1}\right) \\ &\le \sum_{j=1}^{m-1}\mathbb{P}\left(|G(x_j) - F(x_j)| > \omega_T(h) + v\right) \\ &\le Cm\kappa_T^\beta / v^2. \end{aligned} \tag{38}$$

Given $0 < \varepsilon < 1$, choose $m \ge 4/\varepsilon$ and $v = \varepsilon/4$. For all sufficiently large T, $\omega_T(h) \le \varepsilon/4$, so

$$\mathbb{P}\left(\left\|\widehat{F}_{T,M}^H(\cdot \mid h) - F_{\widehat{\epsilon}}^H(\cdot \mid h)\right\|_\infty > \varepsilon\right) \le \frac{16Cm\kappa_T^\beta}{\varepsilon^2} \to 0.$$

This proves the second conclusion of (33). □

Theorem A.26 (H-conditional CDF consistency)**.** *Under Assumptions A.21–A.24,*

$$\sup_u |\widehat{F}_{T,\hat{\epsilon}}^H(u \mid H_{t_T}) - F_{\widehat{\epsilon}}^H(u \mid H_{t_T})| \xrightarrow{p} 0 \tag{39}$$

Proof of Theorem A.26. Set $s_T = \lfloor\sqrt{Tp_T}\rfloor$ and $M_T = T - s_T$, and retain the prefix $\mathcal{V}_T = (V_1, \dots, V_{M_T})$. Since $\Lambda_T \ge p_T$, Assumption A.24 gives $(Tp_T)^{-1} \le \kappa_T^\beta \to 0$. Consequently,

$s_T \to \infty$, $s_T/(Tp_T) \le (Tp_T)^{-1/2} \to 0$, $M_T/T \to 1$, and $\beta_V(s_T) \to 0$. On $S_{T,M_T}(h) > 0$,

$$S_T(h)\left\{\widehat{F}^H_{T,\hat{\epsilon}}(u \mid h) - \widehat{F}^H_{T,M_T}(u \mid h)\right\} = \sum_{i=1}^{T} K_{T,i}(h)\mathbb{1}\{\widehat{\epsilon}_{i+\ell} \le u\} - S_T(h)\widehat{F}^H_{T,M_T}(u \mid h)$$
$$= \sum_{i=M_T+1}^{T} K_{T,i}(h)\left\{\mathbb{1}\{\widehat{\epsilon}_{i+\ell} \le u\} - \widehat{F}^H_{T,M_T}(u \mid h)\right\}.$$

Since $0 \le \widehat{F}^H_{T,M_T} \le 1$ and $S_T(h) \ge S_{T,M_T}(h)$,

$$\left\|\widehat{F}^H_{T,\hat{\epsilon}}(\cdot \mid h) - \widehat{F}^H_{T,M_T}(\cdot \mid h)\right\|_\infty \le \frac{\sum_{i=M_T+1}^{T} K_{T,i}(h)}{S_T(h)} \le \frac{K_{\max}\lfloor\sqrt{Tp_T}\rfloor}{S_{T,M_T}(h)}.$$

Now, the triangle inequality gives

$$\left\|\widehat{F}^H_{T,\hat{\epsilon}}(\cdot \mid h) - F^H_{\widehat{\epsilon}}(\cdot \mid h)\right\|_\infty \le \left\|\widehat{F}^H_{T,M_T}(\cdot \mid h) - F^H_{\widehat{\epsilon}}(\cdot \mid h)\right\|_\infty + \frac{K_{\max}\lfloor\sqrt{Tp_T}\rfloor}{S_{T,M_T}(h)}$$

Define the bound g_T, which is the function of only $\mathcal{V}_T$ and h:

$$g_T(\mathcal{V}_T, h) = \begin{cases} \min\left\{1, \left\|\widehat{F}^H_{T,M_T}(\cdot \mid h) - F^H_{\widehat{\epsilon}}(\cdot \mid h)\right\|_\infty + \frac{K_{\max}s_T}{S_{T,M_T}(h)}\right\}, & S_{T,M_T}(h) > 0, \\ 1, & S_{T,M_T}(h) = 0, \end{cases}$$

so that

$$\left\|\widehat{F}^H_{T,\hat{\epsilon}}(\cdot \mid h) - F^H_{\widehat{\epsilon}}(\cdot \mid h)\right\|_\infty \le g_T(\mathcal{V}_T, h) \le 1. \tag{40}$$

For almost every fixed h, Lemma A.25 and (34) imply

$$\left\|\widehat{F}^H_{T,M_T}(\cdot \mid h) - F^H_{\widehat{\epsilon}}(\cdot \mid h)\right\|_\infty \xrightarrow{p} 0, \quad \frac{S_{T,M_T}(h)}{M_T\mu_T(h)} \xrightarrow{p} 1, \quad \frac{K_{\max}s_T}{M_T\mu_T(h)} \le \frac{K_{\max}}{K_{\min}c_h}\frac{T}{M_T}\frac{s_T}{Tp_T} \to 0.$$

On $\{S_{T,M_T}(h) \ge M_T\mu_T(h)/2\}$,

$$\frac{K_{\max}s_T}{S_{T,M_T}(h)} \le \frac{2K_{\max}s_T}{M_T\mu_T(h)}.$$

Therefore, given $\varepsilon > 0$,

$$\begin{aligned} &\mathbb{P}\{g_T(\mathcal{V}_T, h) > \varepsilon\} \\ &\le \mathbb{P}\left\{\left\|\widehat{F}^H_{T,M_T}(\cdot \mid h) - F^H_{\widehat{\epsilon}}(\cdot \mid h)\right\|_\infty > \frac{\varepsilon}{2}\right\} \\ &\quad + \mathbb{P}\left\{\left|\frac{S_{T,M_T}(h)}{M_T\mu_T(h)} - 1\right| > \frac{1}{2}\right\} + \mathbb{1}\left\{\frac{4K_{\max}s_T}{M_T\mu_T(h)} > \varepsilon\right\} \\ &\to 0. \end{aligned}$$

Since $0 \le g_T(\mathcal{V}_T, h) \le 1$ and $g_T(\mathcal{V}_T, h) \xrightarrow{p} 0$, we have $\mathbb{E}g_T(\mathcal{V}_T, h) \to 0$. Dominated convergence therefore gives

$$\int \mathbb{E}g_T(\mathcal{V}_T, h)P_H(dh) \to 0. \tag{41}$$

Applying Dedecker et al. (2006, Corollary 1) gives

$$\left|\mathbb{E}g_T(\mathcal{V}_T, H_{t_T}) - \int \mathbb{E}g_T(\mathcal{V}_T, h)P_H(dh)\right| \le \beta_V(\lfloor\sqrt{Tp_T}\rfloor).$$

Consequently, combining (40) and (41),

$$\mathbb{E}\left[\sup_u \left|\widehat{F}^H_{T,\hat{\epsilon}}(u \mid H_{t_T}) - F^H_{\widehat{\epsilon}}(u \mid H_{t_T})\right|\right] \le \mathbb{E}g_T(\mathcal{V}_T, H_{t_T})$$
$$\le \int \mathbb{E}g_T(\mathcal{V}_T, h)P_H(dh) + \beta_V(\lfloor\sqrt{Tp_T}\rfloor) \to 0,$$

and Markov's inequality gives the conclusion. □

Corollary A.27 (H-conditional coverage)**.** *Under Assumptions A.21–A.24,*

$$\left|\mathbb{P}\left(Y_{t_T+\ell} \in \widehat{C}_T^{\alpha,\ell}(Z_{t_T}, \widehat{Y}_{t_T+\ell}) \mid H_{t_T}\right) - (1-\alpha)\right| \xrightarrow{p} 0. \tag{42}$$

Proof of Corollary A.27. For $0 < a < 1$, define the true generalized inverse

$$q_a^{H,\star}(h) = \inf\{u : F_{\widehat{\epsilon}}^H(u \mid h) \geq a\}.$$

For every "good" h of Assumption A.23, properness makes $q_a^{H,\star}(h)$ finite, and continuity gives

$$F_{\widehat{\epsilon}}^H(q_a^{H,\star}(h)- \mid h) = F_{\widehat{\epsilon}}^H(q_a^{H,\star}(h) \mid h) = a. \tag{43}$$

Indeed, $u < q_a^{H,\star}(h)$ implies $F_{\widehat{\epsilon}}^H(u \mid h) < a$, and right-continuity implies $F_{\widehat{\epsilon}}^H(q_a^{H,\star}(h) \mid h) \geq a$. The quantile is measurable in h, since for every $r \in \mathbb{R}$,

$$\{h : q_a^{H,\star}(h) < r\} = \bigcup_{v \in \mathbb{Q}:\, v<r} \{h : F_{\widehat{\epsilon}}^H(v \mid h) \geq a\}.$$

Fix

$$0 < \eta < \min\{\alpha_l, 1-\alpha_h, (\alpha_h - \alpha_l)/2\}, \quad D_T = \sup_u |\widehat{F}_{T,\hat{\epsilon}}^H(u \mid H_{t_T}) - F_{\widehat{\epsilon}}^H(u \mid H_{t_T})|.$$

Whenever $\|\widehat{F}_{T,\hat{\epsilon}}^H(\cdot \mid h) - F_{\widehat{\epsilon}}^H(\cdot \mid h)\|_\infty \leq \eta$ and $\tau \in \{\alpha_l, \alpha_h\}$,

$$\begin{aligned} u < q_{\tau-\eta}^{H,\star}(h) \implies \widehat{F}_{T,\hat{\epsilon}}^H(u \mid h) \leq F_{\widehat{\epsilon}}^H(u \mid h) + \eta < \tau, \\ \widehat{F}_{T,\hat{\epsilon}}^H(q_{\tau+\eta}^{H,\star}(h) \mid h) \geq F_{\widehat{\epsilon}}^H(q_{\tau+\eta}^{H,\star}(h) \mid h) - \eta = \tau. \end{aligned}$$

Taking the infimum in the definition of $\widehat{q}_{T,\tau}^H$ gives

$$q_{\tau-\eta}^{H,\star}(h) \leq \widehat{q}_{T,\tau}^H(h) \leq q_{\tau+\eta}^{H,\star}(h). \tag{44}$$

Write

$$\begin{aligned} A_T &= \{Y_{t_T+\ell} \in \widehat{C}_T^{\alpha,\ell}(Z_{t_T}, \widehat{Y}_{t_T+\ell})\} \\ &= \{\widehat{q}_{T,\alpha_l}^H(H_{t_T}) \leq \widehat{\epsilon}_{t_T+\ell} \leq \widehat{q}_{T,\alpha_h}^H(H_{t_T})\}. \end{aligned}$$

By (44), on $\{D_T \leq \eta\}$,

$$\{q_{\alpha_l+\eta}^{H,\star}(H_{t_T}) \leq \widehat{\epsilon}_{t_T+\ell} \leq q_{\alpha_h-\eta}^{H,\star}(H_{t_T})\} \subseteq A_T \subseteq \{q_{\alpha_l-\eta}^{H,\star}(H_{t_T}) \leq \widehat{\epsilon}_{t_T+\ell} \leq q_{\alpha_h+\eta}^{H,\star}(H_{t_T})\}.$$

For any fixed $0 < a < b < 1$, stationarity and (43) imply, almost surely,

$$\begin{aligned} &\mathbb{P}\{q_a^{H,\star}(H_{t_T}) \leq \widehat{\epsilon}_{t_T+\ell} \leq q_b^{H,\star}(H_{t_T}) \mid H_{t_T}\} \\ &= F_{\widehat{\epsilon}}^H(q_b^{H,\star}(H_{t_T}) \mid H_{t_T}) - F_{\widehat{\epsilon}}^H(q_a^{H,\star}(H_{t_T})- \mid H_{t_T}) = b - a. \end{aligned} \tag{45}$$

The left limit accounts for the closed lower endpoint. Conditioning the event inclusions on H_{t_T} therefore gives

$$1 - \alpha - 2\eta - \mathbb{P}(D_T > \eta \mid H_{t_T}) \leq \mathbb{P}(A_T \mid H_{t_T}) \leq 1 - \alpha + 2\eta + \mathbb{P}(D_T > \eta \mid H_{t_T}),$$

and hence

$$|\mathbb{P}(A_T \mid H_{t_T}) - (1-\alpha)| \leq 2\eta + \mathbb{P}(D_T > \eta \mid H_{t_T}). \tag{46}$$

Given $\delta > 0$, choose the preceding η also to satisfy $2\eta < \delta$. Conditional expectation and Markov's inequality yield

$$\begin{aligned} &\mathbb{P}\{|\mathbb{P}(A_T \mid H_{t_T}) - (1-\alpha)| > \delta\} \\ &\leq \mathbb{P}\{\mathbb{P}(D_T > \eta \mid H_{t_T}) > \delta - 2\eta\} \\ &\leq \frac{\mathbb{E}[\mathbb{P}(D_T > \eta \mid H_{t_T})]}{\delta - 2\eta} = \frac{\mathbb{P}(D_T > \eta)}{\delta - 2\eta} \longrightarrow 0, \end{aligned}$$

where the last step is Theorem A.26. □

A.3.2 Extension to coverage conditional on the context Z_t

Assumption A.28 (Stationarity and residual sufficiency)**.** The process $W_t = (Z_t, \widehat{\epsilon}_{t+\ell})$ is strictly stationary, and for every $u \in \mathbb{R}$ and $\mathbb{P}_Z$-almost every z,

$$F_{\widehat{\epsilon}}(u \mid z) = F^H_{\widehat{\epsilon}}(u \mid \mathrm{Enc}_{\theta'}(z)). \tag{47}$$

Interpretation of residual sufficiency. Because $H_t = \mathrm{Enc}_{\theta'}(Z_t)$ is a function of Z_t, (47) is equivalent to $\widehat{\epsilon}_{t+\ell} \perp\!\!\!\perp Z_t \mid H_t$. This is the sufficient dimension reduction condition for the residual (Cook, 2007; Lee et al., 2013). The encoder may discard features that do not affect its conditional distribution. For example, if $Z_t = (S_t, N_t)$, $H_t = S_t$, and $\widehat{\epsilon}_{t+\ell} = m(S_t) + \sigma(S_t)\xi_{t+\ell}$, where $\sigma > 0$ and $\xi_{t+\ell}$ is independent of Z_t with CDF G, then

$$F_{\widehat{\epsilon}}(u \mid s, n) = G\left(\frac{u - m(s)}{\sigma(s)}\right) = F^H_{\widehat{\epsilon}}(u \mid s),$$

although the nuisance component N_t cannot in general be recovered from H_t. This preserves the residual distribution rather than all information in Z_t.

Theorem A.29 (Z-conditional CDF consistency and Z-conditional coverage)**.** *Under Assumptions A.21–A.24 and A.28,*

$$\sup_u |\widehat{F}_{T,\widehat{\epsilon}}(u \mid Z_{t_T}) - F_{\widehat{\epsilon}}(u \mid Z_{t_T})| \xrightarrow{p} 0, \tag{48}$$

and hence,

$$\left|\mathbb{P}\left(Y_{t_T+\ell} \in \widehat{C}^{\alpha,\ell}_T(Z_{t_T}, \widehat{Y}_{t_T+\ell}) \mid Z_{t_T}\right) - (1-\alpha)\right| \xrightarrow{p} 0. \tag{49}$$

Proof of Theorem A.29. Take regular conditional CDF versions in Assumption A.28. For each rational u, the equality in (47) holds outside a P_Z-null set. The countable union of these null sets, together with the preimage of the exceptional set in Assumption A.23, is still P_Z-null. For every z outside this union and every real u, choose rationals $u_n \downarrow u$. Right-continuity gives

$$\begin{aligned} F_{\widehat{\epsilon}}(u \mid z) &= \lim_{n\to\infty} F_{\widehat{\epsilon}}(u_n \mid z) = \lim_{n\to\infty} F^H_{\widehat{\epsilon}}(u_n \mid \mathrm{Enc}_{\theta'}(z)) \\ &= F^H_{\widehat{\epsilon}}(u \mid \mathrm{Enc}_{\theta'}(z)). \end{aligned}$$

Thus the equality holds simultaneously for all u, on one full-measure set of contexts. By stationarity and the definition of $\widehat{F}_{T,\widehat{\epsilon}}$,

$$\begin{aligned} &\sup_u |\widehat{F}_{T,\widehat{\epsilon}}(u \mid Z_{t_T}) - F_{\widehat{\epsilon}}(u \mid Z_{t_T})| \\ &= \sup_u |\widehat{F}^H_{T,\widehat{\epsilon}}(u \mid H_{t_T}) - F^H_{\widehat{\epsilon}}(u \mid H_{t_T})| \xrightarrow{p} 0, \end{aligned}$$

which proves (48).

Use $q^{H,\star}_a$, D_T, and A_T from the proof of Corollary A.27. On the preceding "good" full-measure set, the identity of conditional CDFs and (43) imply, for $0 < a < b < 1$,

$$\begin{aligned} &\mathbb{P}\{q^{H,\star}_a(H_{t_T}) \le \widehat{\epsilon}_{t_T+\ell} \le q^{H,\star}_b(H_{t_T}) \mid Z_{t_T}\} \\ &= F_{\widehat{\epsilon}}(q^{H,\star}_b(H_{t_T}) \mid Z_{t_T}) - F_{\widehat{\epsilon}}(q^{H,\star}_a(H_{t_T})- \mid Z_{t_T}) \\ &= F^H_{\widehat{\epsilon}}(q^{H,\star}_b(H_{t_T}) \mid H_{t_T}) - F^H_{\widehat{\epsilon}}(q^{H,\star}_a(H_{t_T})- \mid H_{t_T}) = b - a. \end{aligned}$$

For $0 < \eta < \min\{\alpha_l, 1-\alpha_h, (\alpha_h - \alpha_l)/2\}$, condition the same event inclusions directly on Z_{t_T} to obtain

$$\begin{aligned} 1 - \alpha - 2\eta - \mathbb{P}(D_T > \eta \mid Z_{t_T}) &\le \mathbb{P}(A_T \mid Z_{t_T}) \\ &\le 1 - \alpha + 2\eta + \mathbb{P}(D_T > \eta \mid Z_{t_T}). \end{aligned}$$

For any $\delta > 0$, choose η also with $2\eta < \delta$. Then

$$\begin{aligned} &\mathbb{P}\{|\mathbb{P}(A_T \mid Z_{t_T}) - (1-\alpha)| > \delta\} \\ &\le \mathbb{P}\{\mathbb{P}(D_T > \eta \mid Z_{t_T}) > \delta - 2\eta\} \le \frac{\mathbb{P}(D_T > \eta)}{\delta - 2\eta} \longrightarrow 0. \end{aligned}$$

This proves the coverage assertion by conditioning the oracle events on Z_{t_T}. □

Table 2: Summary statistics of the three real-world datasets, including the number of features, sequence length, and number of time series.

	Number of features	Sequence length	Number of sequences
Air Quality	10	35064	12
Solar	8	26304	50
Sapflow	10	15007–19934	24

B Additional Details of Experiments

B.1 Datasets

We evaluate the proposed methods and baselines on three real-world datasets that are commonly used as benchmarks in conformal prediction for time series. This section provides details of the datasets, and Table 2 summarizes their key characteristics.

Air quality dataset. The Air Quality dataset (Zhang et al., 2017) comprises 35,064 air quality measurements collected from 12 locations in Beijing, China. It includes two air quality variables, PM10 and PM2.5. We use PM10 as the outcome variable. PM2.5 is excluded from the feature set. The dataset also contains ten additional environmental variables that are used as features. This dataset has also been used in prior work on conformal prediction for time series (Auer et al., 2024; Neglia et al., 2026).

Solar dataset. The Solar dataset is obtained from the U.S. National Solar Radiation Database (Sengupta et al., 2018) and contains half-hourly measurements of solar radiation, along with eight environmental features, at 50 locations. We use diffuse horizontal irradiance (DHI) as the outcome variable and the eight environmental variables as input features. We consider three years of data and resample the original half-hourly measurements to an hourly frequency. This results in 50 time series, each consisting of 26,304 hourly observations resampled from 52,609 half-hourly measurements. This dataset has also been used in prior works (Xu & Xie, 2021; 2023a; Auer et al., 2024; Neglia et al., 2026).

Sapflow dataset. Sap flow refers to the movement of water within a plant and is commonly used as a proxy for plant transpiration in environmental science. The Sapflow dataset is obtained from the SAPFLUXNET database (Poyatos et al., 2020). Following Auer et al. (2024), we use a subset of 24 time series, each containing between 15,000 and 20,000 observations. The sequences have varying sampling frequencies, and each sequence contains 10 features.

B.2 Base predictors

Linear regression. We use `LinearRegressionModel` from the Python Darts (Herzen et al., 2022) library to implement the linear regression base predictor. Each time series sequence is chronologically split into training and held-out sets according to the specified training ratio. A separate model is trained for each sequence. Predictions are made sequentially using a rolling window. At each prediction step, the model uses the most recent `past_window` outcomes and covariates as input features to predict the next `prediction_step` outcomes. The window is then advanced by `prediction_step` time steps, incorporating the new true outcomes and covariates to make the next prediction. Covariates at the forecasting timestamps are not provided. `prediction_step` corresponds to `output_chunk_length` hyperparameter in `LinearRegressionModel`. If `prediction_step` > 1, Darts fits a separate linear regressor for each prediction step. Training is conducted by minimizing the mean squared error between the predicted and true outcomes on the training set. After training, the model forecasts the held-out set of the sequence. We make the true covariates and outcomes available for use as windowed features when forecasting the held-out set. MSE and MAE are computed over each sequence's held-out set and then averaged across sequences. In all experiments, we set `past_window` to 100 and `prediction_step` to 1. Table 3 reports the MAE and RMSE of the linear regression base predictor on all datasets.

Table 3: Mean absolute errors (MAE) and root mean squared errors (RMSE) of three base predictors on all datasets.

	Air		Solar		Sapflow	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Linear Regression	22.1	34.8	26.3	140.8	161.2	375.7
LSTM	17.2	30.3	14.5	33.3	230.7	761.1
Chronos	39.7	65.0	24.6	51.4	420.2	1007.9

Recurrent neural networks. The recurrent neural network (RNN) base predictor is implemented as a global model trained jointly across all time series sequences in each dataset. Each time series sequence is chronologically split into training and held-out sets according to the specified training ratio, with the last 10% of the training set is used as the validation set for model selection. The covariates and outcomes are then normalized using the training sets across all sequences in the dataset. Predictions are made sequentially using a rolling window. At each prediction step, the model uses the most recent `past_window` outcomes and covariates as input features to predict the next `prediction_step` outcomes. The window is then advanced by `prediction_step` time steps, incorporating the new true outcomes and covariates to make the next prediction. Covariates at the forecasting timestamps are not provided. The model consists of a linear input embedding layer, followed by a multi-layer LSTM (Hochreiter & Schmidhuber, 1997), and a final fully connected layer that maps the hidden representation to the predicted outcomes. During training, the model is trained by minimizing the mean squared error using the Adam optimizer (Kingma, 2014). After training, the model forecasts the held-out set of each sequence, and the predictions are transformed back to the original scale. We make the true covariates and outcomes available for use as windowed features when forecasting the held-out set. MSE and MAE are computed over each sequence's held-out set and then averaged across sequences. In all experiments, we set `window_length` to 100 and `prediction_step` to 1. Table 3 reports the MAE and RMSE of the LSTM base predictor on all datasets.

Chronos. We implement the Chronos base predictor using the pretrained `amazon/chronos-2` checkpoint* through the `Chronos2Pipeline`†. We use the model in a zero-shot setting without fine-tuning it on our datasets. Therefore, unlike the linear regression and RNN base predictors, entire sequence is used as the held-out set. Predictions are made sequentially using a rolling window. At each prediction step, the model uses the most recent `past_window` outcomes and covariates as input features to predict the next `prediction_step` outcomes. The window is then advanced by `prediction_step` time steps, incorporating the new true outcomes and covariates to make the next prediction. Covariates at the forecasting timestamps are not provided. Although Chronos-2 can produce forecasts at flexible quantile levels, we use the median prediction, corresponding to the 0.5 quantile, as the predicted outcome. MSE and MAE are computed over each sequence's held-out set and then averaged across sequences. In all experiments, we set `window_length` to 100 and `prediction_length` to 10. Table 3 reports the MAE and RMSE of the Chronos base predictor on all datasets.

B.3 Implementation Details of Proposed Methods

DSCP-QR DSCP-QR consists of a linear projection layer that maps the input to the model dimension, a backbone sequence encoder, and quantile-specific prediction heads. Each head $g_{\theta^{\tau_j}}$ is a fully connected neural network with one or more hidden layers and ReLU activations. To prevent quantile crossing across the ordered target levels $0 < \tau_1 < \cdots < \tau_J < 1$, we parameterize successive quantiles using nonnegative increments:

$$\hat{q}_{T,\tau_1}(z_t) = g_{\theta^{\tau_1}}(h_t),$$
$$\hat{q}_{T,\tau_j}(z_t) = \hat{q}_{T,\tau_{j-1}} + \sigma(g_{\theta^{\tau_j}}(h_t)), \quad j = 2, \ldots, J,$$

*`https://huggingface.co/amazon/chronos-2`

†`https://github.com/amazon-science/chronos-forecasting`

Table 4: Hyperparameter search spaces used for the proposed methods. "Use of current covariate" indicates whether the covariate at the prediction time is included. For context features, `xr` denotes the concatenation of covariates and residuals, while `xry` denotes the concatenation of covariates, residuals, and observed outcomes.

Method	Hyperparameter	Search space
DSCP-QR-RNN	Model dimension	32, 64, 128
	Number of layers	1, 2, 4
	Use of current covariate	True, False
	Context window size	100, 200
	Prediction head hidden layers	1, 2
	Context features	`xr`, `xry`
DSCP-QR-Transformer	Model dimension	32, 64, 128
	Number of heads	2, 4
	Number of layers	2, 4
	Use of current covariate	True, False
	Context window size	100, 200
	Prediction head hidden layers	1, 2
	Context features	`xr`, `xry`
DSCP-QFE-RNN	Model dimension	32, 64, 128
	Number of layers	1, 2, 4
	Use of current covariate	True, False
	Context window size	100, 200
	Prediction head hidden layers	1, 2
	Context features	`xr`, `xry`
DSCP-QFE-Transformer	Model dimension	32, 64, 128
	Number of heads	2, 4
	Number of layers	2, 4
	Use of current covariate	True, False
	Context window size	100, 200
	Prediction head hidden layers	1, 2
	Context features	`xr`, `xry`
DSCP-Local-RNN	Model dimension	32, 64, 128
	Number of layers	1, 2, 4
	Use of current covariate	True, False
	Context window size	100, 200
	Prediction head hidden layers	1, 2
	Context features	`xr`, `xry`
	Calibration window size	100, 500
	Similarity function	`dot`, `cos`, `Euclidean`
	Temperature	0.1, 1.0, 10.0
DSCP-Local-Transformer	Model dimension	32, 64, 128
	Number of heads	2, 4
	Number of layers	2, 4
	Use of current covariate	True, False
	Context window size	100, 200
	Prediction head hidden layers	1, 2
	Context features	`xr`, `xry`
	Calibration window size	100, 500
	Similarity function	`dot`, `cos`, `Euclidean`
	Temperature	0.1, 1.0, 10.0

where $\sigma : \mathbb{R} \to [0, \infty)$ is a nonnegative activation function, such as ReLU or softplus. This construction guarantees

$$\widehat{q}_{T,\tau_1}(z_t) \leq \cdots \leq \widehat{q}_{T,\tau_J}(z_t)$$

for every context z_t. Without this constraint to prevent quantile crossing, each prediction head directly estimates its corresponding quantile:

$$\widehat{q}_{T,\tau_j}(z_t) = g_{\theta^{\tau_j}}(h_t), \quad j = 1, \ldots, J.$$

We use the first 75% of the calibration set for training and the remaining 25% for validation. Training is performed for at most 50 epochs with early stopping using a patience of three epochs, and the checkpoint with the lowest validation loss is selected. The validation loss is the same quantile loss used for training, evaluated at the target quantile levels. We use AdamW (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and a weight decay of 0.01. The initial learning rate is set to 10^{-3} for the RNN encoder and 10^{-4} for the Transformer encoder. To construct prediction intervals with target coverage levels of 0.8, 0.9, and 0.95, we use the target quantile sets $\{0.10, 0.90\}$, $\{0.05, 0.95\}$, and $\{0.025, 0.975\}$, respectively.

The best hyperparameter configuration is selected through a grid search over the search space specified in Table 4. Each configuration is trained and evaluated separately using a subset of sequences, with 20% of the sequences in each dataset used for hyperparameter selection. For each selected sequence, we use a chronological split, with the first 85% of the training data used for model fitting and the remaining 15% used for validation.

Because we prioritize coverage, we adopt a two-stage hyperparameter selection procedure. First, we filter configurations based on empirical coverage. A configuration is considered eligible only if the difference between its empirical and nominal coverage is at least -0.01 for every target quantile pair and every selected sequence. Among the eligible configurations, we select the one with the lowest average Winkler score. The Winkler score is first averaged over evaluation observations within each sequence and then averaged equally across sequences. This coverage-first selection procedure is similar to the hyperparameter selection criterion used by Auer et al. (2024). The best hyperparameter configurations selected for each dataset and base predictor are presented in Tables 5–7.

DSCP-QFE DSCP-QFE consists of a linear projection layer that maps the input to the model dimension, a backbone sequence encoder, and a prediction head. The prediction head is a fully connected neural network with one or more hidden layers and ReLU activations. To prevent quantile crossing in conditional quantile function estimation, we use a partially monotonic prediction head. As defined in § 2.2, let

$$h_t = \mathrm{Enc}_{\theta'}(z_t) \in \mathbb{R}^{d_h}$$

denote the latent representation of the historical context z_t. The encoder does not take the quantile level as an input, so h_t remains fixed when evaluating different quantile levels for the same context. The scalar quantile level $\tau \in (0, 1)$ is provided directly to the prediction head, yielding

$$\widehat{q}_T(z_t, \tau) = g_{\theta''}(h_t, \tau) = g_{\theta''}(\mathrm{Enc}_{\theta'}(z_t), \tau)\,. \tag{50}$$

Let $K \geq 1$ denote the number of hidden layers in the prediction head, and let d_k denote the width of the k-th hidden layer. The hidden representation from the first hidden layer $r_t^{(1)}(\tau) \in \mathbb{R}^{d_1}$ is defined by

$$r_t^{(1)}(\tau) = \phi_1\big(A_1 h_t + b_1^+ \tau + c_1\big)\,, \tag{51}$$

where $A_1 \in \mathbb{R}^{d_1 \times d_h}$, $b_1^+ \in [0, \infty)^{d_1}$, and $c_1 \in \mathbb{R}^{d_1}$. For $k = 2, \ldots, K$, the hidden representations from the subsequent hidden layers $r_t^{(k)}(\tau) \in \mathbb{R}^{d_k}$ are defined recursively by

$$r_t^{(k)}(\tau) = \phi_k\Big(A_k h_t + B_k^+ r_t^{(k-1)}(\tau) + c_k\Big)\,, \qquad k = 2, \ldots, K, \tag{52}$$

where $A_k \in \mathbb{R}^{d_k \times d_h}$, $B_k^+ \in [0, \infty)^{d_k \times d_{k-1}}$, and $c_k \in \mathbb{R}^{d_k}$. When $K = 1$, the recursive layers in (52) are omitted.

The output of the prediction head is

$$g_{\theta''}(h_t, \tau) = a(h_t) + (v^+)^\top r_t^{(K)}(\tau), \tag{53}$$

where $v^+ \in [0, \infty)^{d_K}$ and $a : \mathbb{R}^{d_h} \to \mathbb{R}$ is a continuous scalar-valued function that does not depend on τ. We use a linear mapping,

$$a(h_t) = w_a^\top h_t + b_a, \qquad w_a \in \mathbb{R}^{d_h}, \quad b_a \in \mathbb{R}. \tag{54}$$

Each $\phi_k : \mathbb{R} \to \mathbb{R}$ is a continuous, nondecreasing activation function applied componentwise, such as $\tanh$ or sigmoid. The matrices A_k and bias vectors c_k have no sign constraints. In contrast, the weight vectors b_1^+ and v^+ and the weight matrices B_k^+ are elementwise nonnegative. These

constraints preserve the ordering of the quantile-dependent hidden representations throughout the network. Neither the hidden representations nor the final quantile estimates are required to be nonnegative.

The context representation enters each hidden layer through $A_k h_t$. These context-dependent terms and $a(h_t)$ remain unchanged when computing different quantile levels at the same context and therefore do not affect the monotonicity guarantee. To verify this property, fix a context representation h_t and consider any $0 < \tau_1 < \tau_2 < 1$. Since b_1^+ is elementwise nonnegative and ϕ_1 is nondecreasing,

$$r_t^{(1)}(\tau_1) \le r_t^{(1)}(\tau_2)$$

componentwise. At each subsequent layer, multiplication by the elementwise nonnegative matrix B_k^+, addition of the common term $A_k h_t + c_k$, and application of the nondecreasing activation ϕ_k preserve this ordering. Hence,

$$r_t^{(k)}(\tau_1) \le r_t^{(k)}(\tau_2), \qquad k = 1, \dots, K,$$

componentwise. Finally, the nonnegativity of v^+ implies

$$g_{\theta''}(h_t, \tau_2) - g_{\theta''}(h_t, \tau_1) = (v^+)^\top \left[r_t^{(K)}(\tau_2) - r_t^{(K)}(\tau_1) \right] \ge 0.$$

Therefore, for every context z_t,

$$\widehat{q}_T(z_t, \tau_1) \le \widehat{q}_T(z_t, \tau_2), \qquad 0 < \tau_1 < \tau_2 < 1.$$

Thus, the estimated conditional quantile function is nondecreasing in the quantile level, preventing quantile crossing by construction.

If we do not use the partially monotonic prediction head, we use quantile embedding and fuse it with the encoded representation from the sequence encoder following Dabney et al. (2018); Gouttes et al. (2021). Specifically, for each quantile level $\tau \in (0, 1)$, we construct a vector of K cosine features and apply a learned affine transformation followed by a ReLU activation:

$$\begin{aligned} \mathbf{c}(\tau) &= \left[\cos(\pi i \tau)\right]_{i=0}^{K-1}, \\ \phi(\tau) &= \mathrm{ReLU}(\mathbf{W}_\phi \mathbf{c}(\tau) + \mathbf{b}_\phi), \end{aligned}$$

where $\mathbf{W}_\phi \in \mathbb{R}^{d_h \times K}$ and $\mathbf{b}_\phi \in \mathbb{R}^{d_h}$ are learned parameters, and d_h is the dimension of the encoded representation $\mathbf{h}_t$. We then fuse the quantile embedding with $\mathbf{h}_t$ through residual element-wise multiplication:

$$\mathbf{h}_{t,\tau} = \mathbf{h}_t \odot \big(\mathbf{1} + \phi(\tau)\big) = \mathbf{h}_t + \mathbf{h}_t \odot \phi(\tau),$$

where $\odot$ denotes element-wise multiplication and $\mathbf{1}$ is the all-ones vector. This formulation retains an identity pathway while allowing quantile-dependent modulation of the encoded features. The fused representation is fed into the prediction head. The encoder, quantile embedding, and prediction head are trained jointly. Unlike the partially monotonic head, this construction does not guarantee to prevent quantile crossing.

We use the first 75% of the calibration set for training and the remaining 25% for validation. The model is trained for at most 50 epochs with early stopping using a patience of three epochs, and the checkpoint with the lowest validation loss is selected. Training minimizes the quantile loss evaluated at 64 quantile levels sampled for each training observation. Increasing the number of sampled quantile levels did not affect performance in our experiments. For validation, we evaluate the same quantile loss at the target quantile sets $\{0.10, 0.90\}$, $\{0.05, 0.95\}$, and $\{0.025, 0.975\}$, corresponding to target coverage levels of 0.8, 0.9, and 0.95, respectively. We use AdamW (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and a weight decay of 0.01. The initial learning rate is set to 10^{-3} for the RNN encoder and 10^{-4} for the Transformer encoder. The best hyperparameter configuration is selected using the same grid-search procedure as for DSCP-QR, over the search space specified in Table 4. The best hyperparameter configurations selected for each dataset and base predictor are presented in Tables 5–7.

DSCP-Local DSCP-Local constructs prediction intervals using contextual representations from the sequence encoder rather than directly using prediction heads. This design allows flexibility in the choice of the encoder training objective. To encourage the learned representations to capture information about the conditional distribution of prediction residuals, we adopt the same architecture as DSCP-QR and train the model using the quantile loss. The quantile-specific prediction heads support representation learning during training but are not used to construct prediction intervals; instead, DSCP-Local uses only the learned encoder representations.

For similarity computation, we consider dot product, cosine similarity, and squared Euclidean distance, with a temperature parameter controlling the concentration of calibration weights. Given a query representation h_t and a calibration representation h_i, we compute

$$m_i = \begin{cases} T h_t^\top h_i, & \text{dot product,} \\ T \dfrac{h_t^\top h_i}{\|h_t\|_2 \|h_i\|_2}, & \text{cosine similarity,} \\ -\dfrac{\|h_t - h_i\|_2^2}{T}, & \text{squared Euclidean distance,} \end{cases}$$

where $T > 0$. We normalize these scores across the calibration window $\mathcal{C}$ using softmax:

$$w_i = \frac{\exp(m_i)}{\sum_{j \in \mathcal{C}} \exp(m_j)}.$$

The resulting weights define a sampling distribution over calibration residuals, from which we estimate the residual quantiles used to construct prediction intervals. The experiments use softmax weights, while the DSCP-Local theorical analysis uses a radial kernel estimator under Assumptions A.21—A.24. The experimental softmax variants are therefore not directly covered by the stated compact-support and bandwidth conditions. Both formulations use representation-based weighting, and our analysis identifies sufficient conditions under which this principle yields asymptotic conditional coverage.

We split the calibration set chronologically, using the first 60% for training, the next 15% for validation, and the remaining 25% as a held-out calibration subset. This held-out subset is used to define a similarity-weighted sampling distribution. We train the model for at most 50 epochs with early stopping using a patience of three epochs and select the checkpoint with the lowest validation loss. The validation loss is the same quantile loss used for training, evaluated at the target quantile levels. We compare three choices of target quantile levels: the median alone ($\{0.5\}$), three levels ($\{0.1, 0.5, 0.9\}$), and nine levels ($\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$). As shown in Table 18, using multiple quantile levels yields modest performance improvements over median-only training, while increasing the number of levels from three to nine produces no noticeable further improvement. We therefore train the encoder using the three quantile levels $\{0.1, 0.5, 0.9\}$ in all experiments.

We use AdamW (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and a weight decay of 0.01. The initial learning rate is set to 10^{-3} for the RNN encoder and 10^{-4} for the Transformer encoder. To construct prediction intervals with target coverage levels of 0.8, 0.9, and 0.95, we use the target quantile sets $\{0.10, 0.90\}$, $\{0.05, 0.95\}$, and $\{0.025, 0.975\}$, respectively. The best hyperparameter configuration is selected using the same grid-search procedure as for DSCP-QR, over the search space specified in Table 4. The best hyperparameter configurations selected for each dataset and base predictor are presented in Tables 5–7.

B.4 Implementation Details of Baseline Methods

NexCP (Barber et al., 2023) We use the calibration set to calibrate the prediction intervals. For hyperparameter selection, we further split the calibration set into training and validation subsets using an 85% / 15% split. The best hyperparameter configuration is selected using the same grid-search procedure as for the proposed methods, over the search space specified in Table 8. The best hyperparameter configurations selected for each dataset and base predictor are presented in Tables 9–11.

HopCPT (Auer et al., 2024) Because the datasets used in the original HopCPT experiments include the three datasets used in our study (Air, Solar, and Sapflow), we follow the hyperparameter

Table 5: Selected hyperparameters for the proposed methods on the Air dataset across different base predictors.

Method	Hyperparameter	LR	LSTM	Chronos
DSCP-QR-RNN	Model dimension	64	64	32
	Number of layers	1	1	2
	Use of current covariate	True	True	True
	Context window size	200	200	200
	Prediction head hidden layers	1	1	1
	Context features	xry	xr	xry
DSCP-QR-Transformer	Model dimension	64	32	64
	Number of heads	4	2	2
	Number of layers	2	4	2
	Use of current covariate	True	False	False
	Context window size	200	100	200
	Prediction head hidden layers	1	1	1
	Context features	xry	xr	xry
DSCP-QFE-RNN	Model dimension	64	64	64
	Number of layers	1	1	1
	Use of current covariate	True	True	True
	Context window size	100	100	100
	Prediction head hidden layers	1	1	1
	Context features	xry	xry	xry
DSCP-QFE-Transformer	Model dimension	64	64	64
	Number of heads	2	2	2
	Number of layers	2	2	2
	Use of current covariate	False	False	False
	Context window size	100	100	100
	Prediction head hidden layers	1	1	1
	Context features	xry	xry	xry
DSCP-Local-RNN	Model dimension	64	64	32
	Number of layers	2	2	1
	Use of current covariate	False	False	False
	Context window size	200	200	200
	Prediction head hidden layers	1	1	1
	Context features	xr	xr	xry
	Calibration window size	500	500	500
	Similarity function	cos	cos	cos
	Temperature	10.0	10.0	10.0
DSCP-Local-Transformer	Model dimension	32	64	32
	Number of heads	2	4	4
	Number of layers	4	2	4
	Use of current covariate	True	False	False
	Context window size	200	200	200
	Prediction head hidden layers	1	1	1
	Context features	xry	xry	xry
	Calibration window size	500	500	500
	Similarity function	cos	cos	cos
	Temperature	10.0	10.0	10.0

selection procedure described in the original paper. We split the calibration set equally into training and validation subsets.

SPCI (Xu & Xie, 2023b) We calibrate SPCI prediction intervals using the calibration set. For hyperparameter selection, we divide this set into training and validation subsets containing 85% and 15% of the observations, respectively. The best hyperparameter configuration is selected using the same grid-search procedure as for the proposed methods, over the search space specified in Table 8.

Table 6: Selected hyperparameters for the proposed methods on the Solar dataset across different base predictors.

Method	Hyperparameter	LR	LSTM	Chronos
DSCP-QR-RNN	Model dimension	64	32	32
	Number of layers	1	1	1
	Use of current covariate	True	True	True
	Context window size	200	200	100
	Prediction head hidden layers	1	1	1
	Context features	xry	xry	xry
DSCP-QR-Transformer	Model dimension	64	64	64
	Number of heads	4	4	4
	Number of layers	2	4	2
	Use of current covariate	True	False	False
	Context window size	200	100	200
	Prediction head hidden layers	1	1	1
	Context features	xry	xr	xry
DSCP-QFE-RNN	Model dimension	64	64	64
	Number of layers	1	1	1
	Use of current covariate	True	True	True
	Context window size	100	100	100
	Prediction head hidden layers	1	1	1
	Context features	xry	xry	xry
DSCP-QFE-Transformer	Model dimension	64	64	64
	Number of heads	2	2	2
	Number of layers	2	2	2
	Use of current covariate	False	False	False
	Context window size	100	100	100
	Prediction head hidden layers	1	1	1
	Context features	xry	xry	xry
DSCP-Local-RNN	Model dimension	64	64	32
	Number of layers	2	2	1
	Use of current covariate	True	False	False
	Context window size	200	200	200
	Prediction head hidden layers	1	1	1
	Context features	xr	xry	xr
	Calibration window size	500	500	500
	Similarity function	cos	cos	cos
	Temperature	10.0	10.0	10.0
DSCP-Local-Transformer	Model dimension	32	64	32
	Number of heads	2	4	4
	Number of layers	4	2	4
	Use of current covariate	True	False	False
	Context window size	200	200	200
	Prediction head hidden layers	1	1	1
	Context features	xry	xry	xry
	Calibration window size	500	500	500
	Similarity function	cos	cos	cos
	Temperature	10.0	10.0	10.0

The best hyperparameter configurations selected for each dataset and base predictor are presented in Tables 9–11.

In its original formulation, SPCI retrains its random forest regressor at each prediction step, which can be computationally prohibitive for larger datasets. Auer et al. (2024) identify the same computational bottleneck and adopt a modified implementation that trains the regressor only on the calibration set, reporting only a small performance decrease in experiments on their smallest dataset. We therefore adopt the same modification and keep the regressor fixed during prediction. This choice

Table 7: Selected hyperparameters for the proposed methods on the Sapflow dataset across different base predictors.

Method	Hyperparameter	LR	LSTM	Chronos
DSCP-QR-RNN	Model dimension	64	32	32
	Number of layers	1	1	1
	Use of current covariate	False	False	False
	Context window size	100	100	100
	Prediction head hidden layers	1	1	1
	Context features	`xr`	`xr`	`xr`
DSCP-QR-Transformer	Model dimension	64	64	64
	Number of heads	4	4	4
	Number of layers	2	2	2
	Use of current covariate	False	False	False
	Context window size	200	200	200
	Prediction head hidden layers	1	1	1
	Context features	`xr`	`xr`	`xr`
DSCP-QFE-RNN	Model dimension	64	64	64
	Number of layers	1	1	1
	Use of current covariate	False	False	False
	Context window size	100	100	100
	Prediction head hidden layers	1	1	1
	Context features	`xry`	`xry`	`xry`
DSCP-QFE-Transformer	Model dimension	64	64	64
	Number of heads	2	2	4
	Number of layers	2	2	2
	Use of current covariate	False	False	False
	Context window size	100	100	100
	Prediction head hidden layers	1	1	1
	Context features	`xry`	`xry`	`xry`
DSCP-Local-RNN	Model dimension	64	64	32
	Number of layers	2	2	1
	Use of current covariate	False	False	False
	Context window size	200	200	200
	Prediction head hidden layers	1	1	1
	Context features	`xr`	`xry`	`xr`
	Calibration window size	500	500	500
	Similarity function	`cos`	`cos`	`cos`
	Temperature	10.0	10.0	10.0
DSCP-Local-Transformer	Model dimension	32	64	32
	Number of heads	2	4	4
	Number of layers	4	2	4
	Use of current covariate	True	False	False
	Context window size	200	200	200
	Prediction head hidden layers	1	1	1
	Context features	`xry`	`xry`	`xry`
	Calibration window size	500	500	500
	Similarity function	`cos`	`cos`	`cos`
	Temperature	10.0	10.0	10.0

is also consistent with the other methods in our comparison, whose models are not retrained after calibration.

KOWCPI (Lee et al., 2024a) We use calibration set to calibrate the prediction intervals. For hyperparameter selection, we divide this set into training and validation subsets containing 85% and 15% of the observations, respectively. We use Epanechnikov kernel for all experiments. The best hyperparameter configuration is selected using the same grid-search procedure as for the proposed

Table 8: Hyperparameter search spaces used for the baseline methods.

Method	Hyperparameter	Search space
NexCP	Decay factor ρ	$0.9,\ 0.95,\ 0.99,\ 0.999$
	Maximum past-window length	$100,\ 200,\ 500$
SPCI	Number of estimators	$10,\ 20,\ 50$
	Max depth	$2,\ 5,\ 10$
	Window size	$100,\ 200,\ 500$
KOWCPI	Past-window length	$5,\ 10,\ 20,\ 50,\ 100,\ 200$
	Bandwidth	$1,\ 5,\ 10,\ 20,\ 50,\ 100$
ResCP	Spectral radius γ	$0.5,\ 0.9,\ 1.2,\ 1.5$
	Leak rate	$0.5,\ 0.8,\ 1.0$
	Input scaling	$0.1,\ 0.25,\ 0.75,\ 1.5$
	Temperature	$0.03,\ 0.1,\ 0.3$
	Calibration size	$100,\ 500,\ 1500,\ 4000$
DistMatch	KS threshold	$0.01,\ 0.05,\ 0.1$
	Past-window length	100

methods, over the search space specified in Table 8. The best hyperparameter configurations selected for each dataset and base predictor are presented in Tables 9–11.

ResCP (Neglia et al., 2026) We calibrate the prediction intervals using the calibration set. Because the Air and Solar datasets were also evaluated in the original publication, we adopt the hyperparameter configurations reported for these datasets and perform hyperparameter selection only for Sapflow. For this selection, we divide the Sapflow calibration set into training and validation subsets containing 85% and 15% of the observations, respectively. We select the best configuration using the same grid-search procedure as for the proposed methods, over the search space specified in Table 8. The hyperparameter configurations used for each dataset and base predictor are reported in Tables 9–11.

DistMatch (Menadjiev et al., 2026) We tune the KS matching threshold $\gamma \in \{0.01, 0.05, 0.1\}$ while fixing the residual patch length at $w = 100$, following the authors experimental setup. The number of matching trees, bootstrap sampling ratio, and minimum node-size parameter are fixed at 10, 0.9, and 0, respectively, as specified in Appendix of Menadjiev et al. (2026). We select the best configuration using the same grid-search procedure as for the proposed methods, over the search space specified in Table 8. The hyperparameter configurations used for each dataset and base predictor are reported in Tables 9–11.

B.5 Evaluation Metrics

Given a prediction interval at time t,

$$\widehat{C}_t^{\alpha} = \widehat{C}^{\alpha}\left(\hat{q}_t^{\alpha_l}, \hat{q}_t^{\alpha_h}, \hat{y}_t\right) = \left[\hat{y}_t + \hat{q}_t^{\alpha_l}, \hat{y}_t + \hat{q}_t^{\alpha_h}\right],$$

we denote its lower and upper endpoints by

$$L_t = \hat{y}_t + \hat{q}_t^{\alpha_l}, \qquad U_t = \hat{y}_t + \hat{q}_t^{\alpha_h},$$

respectively. Here, α denotes the significance level, such that the target coverage is $1 - \alpha$. Let $\mathcal{D}_{\text{test}}$ denote the set of observations in a test sequence. We evaluate the prediction intervals using the following metrics.

Coverage gap (ΔCov). The coverage gap for a test sequence at significance level α is defined as

$$\Delta\text{Cov} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{t \in \mathcal{D}_{\text{test}}} \mathbb{1}\left(y_t \in \widehat{C}_t^{\alpha}\right) - (1 - \alpha). \tag{55}$$

The coverage gap measures the difference between the empirical coverage and the target coverage $1 - \alpha$. A positive value indicates overcoverage, whereas a negative value indicates undercoverage. For each dataset, we report the coverage gap averaged across all test sequences.

Table 9: Selected hyperparameters for the baseline methods across different base predictors on Air dataset.

Method	Hyperparameter	LR	LSTM	Chronos
NexCP	Decay factor ρ	0.99	0.999	0.99
	Maximum past-window length	500	200	500
SPCI	Number of estimators	50	50	50
	Max depth	10	10	10
	Window size	100	100	200
KOWCPI	Past-window length	100	100	100
ResCP	Spectral radius γ	1.45	1.3	1.0
	Leak rate	0.65	0.95	0.8
	Input scaling	0.75	0.25	0.25
	Temperature	0.15	0.1	0.1
	Calibration size	3800	3200	10500
DistMatch	KS threshold	0.1	0.1	0.1
	Past-window length	100	100	100

Table 10: Selected hyperparameters for the baseline methods across different base predictors on Solar dataset.

Method	Hyperparameter	LR	LSTM	Chronos
NexCP	Decay factor ρ	0.999	0.99	0.99
	Maximum past-window length	100	100	100
SPCI	Number of estimators	50	50	50
	Max depth	10	10	10
	Window size	100	100	100
KOWCPI	Past-window length	100	100	100
ResCP	Spectral radius γ	1.0	0.9	0.9
	Leak rate	0.8	0.75	0.8
	Input scaling	0.2	0.7	0.4
	Temperature	0.1	0.1	0.1
	Calibration size	1500	3900	7600
DistMatch	KS threshold	0.1	0.1	0.1
	Past-window length	100	100	100

Average prediction interval width (PI-W). The average prediction interval width for a test sequence is defined as

$$\text{PI-W} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{t \in \mathcal{D}_{\text{test}}} (U_t - L_t) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{t \in \mathcal{D}_{\text{test}}} (\hat{q}_t^{\alpha_h} - \hat{q}_t^{\alpha_l}) \,. \tag{56}$$

A smaller interval width indicates a more efficient prediction interval, provided that the target coverage is obtained. For each dataset, we report the average prediction interval width across all test sequences.

Winkler score (WS). The Winkler score (Winkler, 1972) jointly evaluates prediction interval width and coverage. It is equal to the interval width when the true outcome falls within the prediction interval and adds a penalty when the outcome falls outside the interval. The penalty is proportional to the distance between the true outcome and the nearest interval endpoint, with a factor of $2/\alpha$. At

Table 11: Selected hyperparameters for the baseline methods across different base predictors on Sapflow dataset.

Method	Hyperparameter	LR	LSTM	Chronos
NexCP	Decay factor ρ	0.99	0.99	0.99
	Maximum past-window length	500	200	200
SPCI	Number of estimators	50	50	50
	Max depth	10	10	10
	Window size	200	200	200
KOWCPI	Past-window length	100	100	100
ResCP	Spectral radius γ	1.2	1.2	1.2
	Leak rate	0.9	0.9	0.9
	Input scaling	0.25	0.25	0.25
	Temperature	0.1	0.1	0.1
	Calibration size	3800	3800	3800
DistMatch	KS threshold	0.1	0.1	0.1
	Past-window length	100	100	100

time t, the Winkler score is defined as

$$\mathrm{WS}_t = \begin{cases} (U_t - L_t) + \dfrac{2}{\alpha}(L_t - y_t), & \text{if } y_t < L_t, \\ (U_t - L_t), & \text{if } L_t \le y_t \le U_t, \\ (U_t - L_t) + \dfrac{2}{\alpha}(y_t - U_t), & \text{if } y_t > U_t. \end{cases} \tag{57}$$

The average Winkler score for a test sequence is then given by

$$\mathrm{WS} = \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{t \in \mathcal{D}_{\mathrm{test}}} \mathrm{WS}_t. \tag{58}$$

Lower Winkler scores indicate smaller prediction intervals with fewer or less severe coverage violations. For each dataset, we report the Winkler score averaged across all test sequences.

Table 12: Coverage gap (ΔCov), prediction interval width (PI-W), and Winkler score (WS) averaged across sequences in each dataset for the proposed methods with different sequence encoder: RNN and Transformer. The target coverage is $1 - \alpha = 0.9$. The best result among two encoder choices for each method except the undercoveraged results based on Winkler score is underlined.

Method	Air								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
QR-RNN	$+0.012_{\pm.016}$	$85.3_{\pm10.5}$	$\underline{121.3}_{\pm12.5}$	$-0.021_{\pm.017}$	$61.1_{\pm9.32}$	$\underline{103.4}_{\pm12.1}$	$+0.014_{\pm.014}$	$87.1_{\pm10.7}$	$\underline{125.5}_{\pm10.6}$
QR-Transformer	$-0.004_{\pm.014}$	$92.3_{\pm14.7}$	$136.3_{\pm18.5}$	$-0.018_{\pm.018}$	$69.4_{\pm10.6}$	$124.2_{\pm16.0}$	$-0.002_{\pm.013}$	$93.5_{\pm9.82}$	$146.3_{\pm13.2}$
QFE-RNN	$-0.064_{\pm.015}$	$61.3_{\pm6.40}$	$\underline{132.7}_{\pm15.6}$	$-0.126_{\pm.026}$	$37.9_{\pm4.73}$	$\underline{124.4}_{\pm16.3}$	$-0.091_{\pm.032}$	$52.5_{\pm7.39}$	$\underline{170.1}_{\pm13.1}$
QFE-Transformer	$-0.084_{\pm.021}$	$63.4_{\pm6.78}$	$152.7_{\pm19.8}$	$-0.112_{\pm.028}$	$45.7_{\pm6.78}$	$143.4_{\pm22.3}$	$-0.096_{\pm.022}$	$58.2_{\pm6.31}$	$196.1_{\pm17.1}$
Local-RNN	$-0.015_{\pm.003}$	$81.0_{\pm10.0}$	$\underline{133.2}_{\pm14.4}$	$-0.013_{\pm.005}$	$67.8_{\pm8.67}$	$\underline{116.0}_{\pm13.3}$	$+0.027_{\pm.006}$	$117.4_{\pm9.14}$	$\underline{160.7}_{\pm11.4}$
Local-Transformer	$-0.007_{\pm.004}$	$83.4_{\pm9.60}$	$140.8_{\pm15.5}$	$-0.007_{\pm.003}$	$71.8_{\pm9.29}$	$127.6_{\pm15.1}$	$+0.015_{\pm.005}$	$130.4_{\pm11.9}$	$179.9_{\pm14.2}$
Method	Solar								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
QR-RNN	$-0.019_{\pm.107}$	$81.9_{\pm14.6}$	$\underline{121.6}_{\pm82.9}$	$-0.006_{\pm.120}$	$55.4_{\pm10.8}$	$\underline{84.0}_{\pm84.0}$	$+0.007_{\pm.020}$	$62.9_{\pm9.97}$	$\underline{82.4}_{\pm11.5}$
QR-Transformer	$-0.031_{\pm.121}$	$87.3_{\pm16.8}$	$150.1_{\pm186.5}$	$-0.005_{\pm.028}$	$62.1_{\pm12.9}$	$90.7_{\pm24.7}$	$-0.001_{\pm.023}$	$75.3_{\pm14.0}$	$102.3_{\pm13.9}$
QFE-RNN	$-0.114_{\pm.019}$	$55.9_{\pm9.50}$	$\underline{159.7}_{\pm41.8}$	$-0.120_{\pm.061}$	$33.4_{\pm8.74}$	$\underline{122.8}_{\pm37.2}$	$-0.103_{\pm.035}$	$42.1_{\pm8.71}$	$\underline{123.3}_{\pm15.6}$
QFE-Transformer	$-0.165_{\pm.098}$	$49.8_{\pm7.16}$	$203.8_{\pm144.5}$	$-0.152_{\pm.105}$	$25.9_{\pm5.84}$	$172.7_{\pm194.0}$	$-0.127_{\pm.023}$	$37.3_{\pm8.23}$	$152.4_{\pm16.5}$
Local-RNN	$-0.001_{\pm.009}$	$91.0_{\pm18.9}$	$124.3_{\pm28.7}$	$+0.016_{\pm.010}$	$64.0_{\pm15.3}$	$\underline{87.0}_{\pm17.6}$	$+0.011_{\pm.007}$	$83.1_{\pm14.3}$	$\underline{109.9}_{\pm13.2}$
Local-Transformer	$-0.001_{\pm.008}$	$88.6_{\pm18.4}$	$\underline{123.8}_{\pm27.6}$	$+0.002_{\pm.009}$	$60.7_{\pm15.7}$	$105.7_{\pm19.0}$	$+0.001_{\pm.011}$	$75.9_{\pm12.4}$	$122.6_{\pm14.3}$
Method	Sapflow								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
QR-RNN	$+0.005_{\pm.064}$	$708.5_{\pm511.9}$	$\underline{994.5}_{\pm687.8}$	$-0.008_{\pm.111}$	$839.3_{\pm917.8}$	$\underline{1142.3}_{\pm1023.3}$	$-0.061_{\pm.065}$	$864.8_{\pm549.1}$	$\underline{1450.2}_{\pm953.0}$
QR-Transformer	$-0.014_{\pm.117}$	$766.1_{\pm450.4}$	$1151.1_{\pm726.4}$	$-0.077_{\pm.162}$	$1339.0_{\pm2828.9}$	$1911.9_{\pm3030.8}$	$-0.092_{\pm.085}$	$1079.5_{\pm720.9}$	$2296.8_{\pm1915.4}$
QFE-RNN	$-0.056_{\pm.089}$	$569.1_{\pm469.1}$	$\underline{1078.2}_{\pm753.2}$	$-0.115_{\pm.128}$	$527.0_{\pm461.3}$	$\underline{1188.3}_{\pm884.6}$	$-0.137_{\pm.101}$	$563.4_{\pm442.8}$	$\underline{1885.1}_{\pm1216.4}$
QFE-Transformer	$-0.046_{\pm.116}$	$609.6_{\pm421.5}$	$1209.8_{\pm829.9}$	$-0.083_{\pm.143}$	$665.3_{\pm670.8}$	$1436.8_{\pm1262.0}$	$-0.141_{\pm.093}$	$838.4_{\pm777.8}$	$2602.0_{\pm2046.1}$
Local-RNN	$+0.001_{\pm.012}$	$558.2_{\pm375.5}$	$\underline{864.2}_{\pm619.6}$	$+0.007_{\pm.013}$	$632.1_{\pm550.7}$	$\underline{911.4}_{\pm743.6}$	$+0.017_{\pm.010}$	$1400.5_{\pm1102.3}$	$\underline{1860.4}_{\pm1382.5}$
Local-Transformer	$+0.009_{\pm.006}$	$586.1_{\pm380.5}$	$912.5_{\pm634.6}$	$+0.007_{\pm.008}$	$898.3_{\pm1357.8}$	$1245.6_{\pm1264.3}$	$+0.007_{\pm.005}$	$1927.9_{\pm1557.9}$	$3044.2_{\pm2340.9}$

Table 13: Coverage gap (ΔCov), prediction interval width (PI-W), and Winkler score (WS) averaged across sequences in each dataset for the proposed methods with different sequence encoder: RNN and Transformer. The target coverage is $1 - \alpha = 0.95$. The best result among two encoder choices for each method except the undercoveraged results based on Winkler score is underlined.

Method	Air								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
QR-RNN	$-0.005_{\pm.009}$	$107.4_{\pm15.0}$	$\underline{155.6}_{\pm17.1}$	$-0.027_{\pm.011}$	$78.1_{\pm7.74}$	$\underline{135.8}_{\pm19.0}$	$+0.001_{\pm.013}$	$109.0_{\pm13.3}$	$\underline{159.8}_{\pm13.8}$
QR-Transformer	$-0.010_{\pm.012}$	$116.0_{\pm18.2}$	$172.7_{\pm21.7}$	$-0.016_{\pm.009}$	$93.7_{\pm14.1}$	$166.2_{\pm21.8}$	$-0.006_{\pm.012}$	$121.1_{\pm14.9}$	$191.3_{\pm16.6}$
QFE-RNN	$-0.097_{\pm.017}$	$64.8_{\pm7.03}$	$\underline{196.8}_{\pm27.0}$	$-0.153_{\pm.032}$	$40.3_{\pm4.47}$	$\underline{203.7}_{\pm32.7}$	$-0.113_{\pm.035}$	$60.0_{\pm8.84}$	$\underline{282.4}_{\pm22.6}$
QFE-Transformer	$-0.118_{\pm.022}$	$66.8_{\pm6.93}$	$234.7_{\pm37.4}$	$-0.138_{\pm.025}$	$49.2_{\pm5.23}$	$230.0_{\pm36.9}$	$-0.128_{\pm0.024}$	$61.8_{\pm7.13}$	$325.3_{\pm27.1}$
Local-RNN	$-0.007_{\pm.003}$	$109.1_{\pm13.5}$	$\underline{172.1}_{\pm16.9}$	$-0.006_{\pm.004}$	$93.9_{\pm12.4}$	$\underline{151.4}_{\pm16.8}$	$+0.009_{\pm.003}$	$150.2_{\pm11.0}$	$\underline{208.5}_{\pm14.8}$
Local-Transformer	$-0.011_{\pm.002}$	$106.0_{\pm12.8}$	$174.5_{\pm18.6}$	$-0.010_{\pm.003}$	$93.7_{\pm11.1}$	$159.1_{\pm18.0}$	$+0.006_{\pm.002}$	$163.9_{\pm14.4}$	$224.0_{\pm16.2}$
Method	Solar								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
QR-RNN	$-0.022_{\pm.013}$	$100.1_{\pm17.8}$	$222.0_{\pm629.5}$	$-0.020_{\pm.134}$	$67.0_{\pm15.1}$	$138.5_{\pm383.9}$	$+0.001_{\pm.020}$	$75.9_{\pm11.3}$	$\underline{96.1}_{\pm12.3}$
QR-Transformer	$-0.011_{\pm.013}$	$114.8_{\pm41.7}$	$\underline{156.3}_{\pm83.1}$	$+0.004_{\pm.018}$	$84.0_{\pm18.0}$	$\underline{112.0}_{\pm29.1}$	$-0.003_{\pm.012}$	$93.0_{\pm13.8}$	$123.5_{\pm14.0}$
QFE-RNN	$-0.160_{\pm0.110}$	$53.8_{\pm8.4}$	$\underline{327.4}_{\pm753.9}$	$-0.166_{\pm0.096}$	$28.7_{\pm6.3}$	$\underline{205.4}_{\pm157.7}$	$-0.118_{\pm0.029}$	$36.7_{\pm7.4}$	$\underline{177.7}_{\pm33.8}$
QFE-Transformer	$-0.190_{\pm.106}$	$52.5_{\pm6.70}$	$364.5_{\pm362.8}$	$-0.140_{\pm.108}$	$41.5_{\pm13.3}$	$311.6_{\pm217.3}$	$-0.155_{\pm.025}$	$37.6_{\pm6.82}$	$274.4_{\pm35.1}$
Local-RNN	$-0.002_{\pm0.007}$	$114.4_{\pm24.8}$	$\underline{149.8}_{\pm39.0}$	$+0.012_{\pm0.007}$	$83.1_{\pm19.6}$	$\underline{107.3}_{\pm20.5}$	$+0.006_{\pm0.006}$	$106.0_{\pm16.4}$	$\underline{135.6}_{\pm14.2}$
Local-Transformer	$+0.004_{\pm0.006}$	$118.1_{\pm25.0}$	$158.5_{\pm36.1}$	$+0.004_{\pm0.006}$	$90.2_{\pm21.8}$	$133.7_{\pm23.6}$	$+0.002_{\pm0.006}$	$107.2_{\pm16.2}$	$155.0_{\pm15.8}$
Method	Sapflow								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
QR-RNN	$-0.018_{\pm0.048}$	$931.1_{\pm675.3}$	$\underline{1340.1}_{\pm955.6}$	$-0.043_{\pm0.108}$	$1141.8_{\pm1113.8}$	$1978.3_{\pm2901.4}$	$-0.070_{\pm0.067}$	$1039.8_{\pm612.6}$	$\underline{2002.1}_{\pm1607.5}$
QR-Transformer	$+0.004_{\pm0.049}$	$1082.3_{\pm722.3}$	$1460.3_{\pm1041.9}$	$-0.002_{\pm0.046}$	$1235.4_{\pm1088.7}$	$\underline{1800.7}_{\pm2008.7}$	$-0.050_{\pm0.056}$	$1651.8_{\pm1395.5}$	$3277.0_{\pm3476.1}$
QFE-RNN	$-0.092_{\pm0.098}$	$632.0_{\pm575.3}$	$\underline{1608.2}_{\pm1330.9}$	$-0.137_{\pm0.143}$	$604.6_{\pm492.2}$	$\underline{1903.0}_{\pm1492.3}$	$-0.160_{\pm0.074}$	$654.6_{\pm477.6}$	$\underline{3161.2}_{\pm2164.8}$
QFE-Transformer	$-0.064_{\pm0.100}$	$808.0_{\pm739.8}$	$1773.9_{\pm1299.4}$	$-0.078_{\pm0.117}$	$942.0_{\pm994.4}$	$2216.9_{\pm2101.4}$	$-0.110_{\pm0.104}$	$1380.6_{\pm1405.6}$	$3756.0_{\pm2527.8}$
Local-RNN	$-0.001_{\pm0.008}$	$743.0_{\pm520.8}$	$\underline{1111.2}_{\pm827.7}$	$+0.002_{\pm0.008}$	$833.4_{\pm716.4}$	$\underline{1162.5}_{\pm954.9}$	$+0.008_{\pm0.008}$	$1783.3_{\pm1301.5}$	$\underline{2269.2}_{\pm1630.1}$
Local-Transformer	$+0.004_{\pm0.004}$	$795.3_{\pm523.8}$	$1184.7_{\pm873.0}$	$+0.002_{\pm0.005}$	$1101.3_{\pm1398.2}$	$1487.1_{\pm1661.7}$	$+0.004_{\pm0.004}$	$2794.6_{\pm2282.1}$	$3930.2_{\pm3185.8}$

Table 14: Coverage gap (ΔCov), prediction interval width (PI-W), and Winkler score (WS) averaged across sequences in each dataset for the proposed methods with different sequence encoder: RNN and Transformer. The target coverage is $1-\alpha = 0.8$. The best result among two encoder choices for each method except the undercoveraged results based on Winkler score is underlined.

Method	Air								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
QR-RNN	$+0.011_{\pm 0.019}$	$60.9_{\pm 7.9}$	$\underline{93.7}_{\pm 10.9}$	$-0.022_{\pm 0.028}$	$43.1_{\pm 6.9}$	$\underline{77.7}_{\pm 9.5}$	$+0.016_{\pm 0.026}$	$62.5_{\pm 8.1}$	$\underline{97.4}_{\pm 7.8}$
QR-Transformer	$+0.016_{\pm 0.023}$	$69.3_{\pm 10.7}$	$105.6_{\pm 14.6}$	$-0.022_{\pm 0.023}$	$47.9_{\pm 4.0}$	$91.9_{\pm 11.7}$	$-0.000_{\pm 0.032}$	$68.2_{\pm 7.6}$	$113.8_{\pm 9.9}$
QFE-RNN	$-0.002_{\pm 0.022}$	$55.0_{\pm 5.8}$	$\underline{96.4}_{\pm 10.3}$	$-0.073_{\pm 0.022}$	$33.1_{\pm 4.1}$	$\underline{81.9}_{\pm 9.7}$	$-0.032_{\pm 0.034}$	$45.4_{\pm 6.6}$	$\underline{109.8}_{\pm 8.6}$
QFE-Transformer	$-0.033_{\pm 0.032}$	$55.2_{\pm 4.4}$	$109.5_{\pm 13.7}$	$-0.055_{\pm 0.030}$	$40.2_{\pm 4.6}$	$95.0_{\pm 12.8}$	$-0.029_{\pm 0.028}$	$51.9_{\pm 6.4}$	$127.4_{\pm 10.7}$
LCP-RNN	$-0.009_{\pm 0.006}$	$58.6_{\pm 7.5}$	$\underline{102.0}_{\pm 11.3}$	$-0.002_{\pm 0.004}$	$47.3_{\pm 6.4}$	$\underline{85.9}_{\pm 10.7}$	$+0.048_{\pm 0.008}$	$86.1_{\pm 6.9}$	$\underline{123.4}_{\pm 9.0}$
LCP-Transformer	$-0.015_{\pm 0.004}$	$57.6_{\pm 7.2}$	$103.2_{\pm 11.8}$	$-0.012_{\pm 0.004}$	$47.0_{\pm 6.0}$	$89.7_{\pm 10.9}$	$+0.016_{\pm 0.008}$	$93.2_{\pm 9.2}$	$141.6_{\pm 11.5}$
Method	Solar								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
QR-RNN	$-0.015_{\pm 0.071}$	$62.3_{\pm 10.1}$	$\underline{95.1}_{\pm 38.9}$	$-0.009_{\pm 0.113}$	$42.0_{\pm 9.1}$	$\underline{72.5}_{\pm 95.5}$	$+0.012_{\pm 0.036}$	$48.0_{\pm 8.5}$	$\underline{68.1}_{\pm 9.8}$
QR-Transformer	$-0.029_{\pm 0.109}$	$65.9_{\pm 12.0}$	$115.7_{\pm 108.8}$	$+0.007_{\pm 0.096}$	$46.9_{\pm 11.4}$	$77.6_{\pm 49.6}$	$+0.007_{\pm 0.036}$	$56.1_{\pm 10.5}$	$83.0_{\pm 11.2}$
QFE-RNN	$-0.078_{\pm 0.101}$	$43.6_{\pm 6.7}$	$\underline{108.4}_{\pm 112.3}$	$-0.060_{\pm 0.108}$	$21.0_{\pm 4.6}$	$\underline{76.2}_{\pm 108.4}$	$-0.047_{\pm 0.046}$	$29.0_{\pm 5.6}$	$\underline{70.4}_{\pm 12.3}$
QFE-Transformer	$-0.100_{\pm 0.081}$	$44.3_{\pm 5.8}$	$127.9_{\pm 50.2}$	$-0.070_{\pm 0.075}$	$22.8_{\pm 5.4}$	$92.4_{\pm 30.3}$	$-0.063_{\pm 0.023}$	$31.3_{\pm 6.4}$	$100.6_{\pm 14.7}$
LCP-RNN	$-0.002_{\pm 0.008}$	$66.3_{\pm 13.1}$	$\underline{100.3}_{\pm 21.4}$	$+0.006_{\pm 0.012}$	$44.5_{\pm 10.9}$	$\underline{68.6}_{\pm 14.3}$	$+0.008_{\pm 0.009}$	$59.7_{\pm 11.1}$	$\underline{86.7}_{\pm 11.6}$
LCP-Transformer	$-0.012_{\pm 0.009}$	$62.0_{\pm 12.5}$	$104.6_{\pm 20.3}$	$-0.011_{\pm 0.011}$	$35.5_{\pm 9.6}$	$79.9_{\pm 15.7}$	$-0.024_{\pm 0.011}$	$49.0_{\pm 9.3}$	$97.2_{\pm 12.1}$
Method	Sapflow								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
QR-RNN	$-0.030_{\pm 0.104}$	$491.2_{\pm 325.1}$	$\underline{766.9}_{\pm 518.2}$	$-0.029_{\pm 0.141}$	$588.5_{\pm 455.5}$	$\underline{934.6}_{\pm 797.3}$	$-0.055_{\pm 0.087}$	$630.1_{\pm 395.9}$	$\underline{1122.2}_{\pm 833.4}$
QR-Transformer	$+0.027_{\pm 0.101}$	$596.5_{\pm 387.9}$	$869.1_{\pm 556.0}$	$-0.006_{\pm 0.150}$	$684.4_{\pm 685.9}$	$1043.9_{\pm 1057.9}$	$-0.108_{\pm 0.093}$	$756.3_{\pm 504.7}$	$1747.9_{\pm 1548.2}$
QFE-RNN	$-0.048_{\pm 0.105}$	$456.0_{\pm 323.5}$	$\underline{864.8}_{\pm 637.2}$	$-0.073_{\pm 0.142}$	$461.4_{\pm 398.4}$	$\underline{862.6}_{\pm 652.1}$	$-0.121_{\pm 0.093}$	$447.0_{\pm 310.1}$	$\underline{1266.9}_{\pm 859.5}$
QFE-Transformer	$-0.027_{\pm 0.128}$	$567.4_{\pm 495.7}$	$940.7_{\pm 630.7}$	$-0.003_{\pm 0.132}$	$496.4_{\pm 425.3}$	$974.4_{\pm 936.9}$	$-0.071_{\pm 0.098}$	$731.6_{\pm 752.0}$	$1632.6_{\pm 1272.6}$
LCP-RNN	$+0.001_{\pm 0.014}$	$391.9_{\pm 253.6}$	$\underline{663.1}_{\pm 458.2}$	$+0.013_{\pm 0.019}$	$446.2_{\pm 398.7}$	$\underline{700.8}_{\pm 570.7}$	$+0.014_{\pm 0.015}$	$953.8_{\pm 731.0}$	$\underline{1445.1}_{\pm 1064.5}$
LCP-Transformer	$+0.006_{\pm 0.006}$	$403.3_{\pm 260.8}$	$691.5_{\pm 464.9}$	$+0.017_{\pm 0.014}$	$579.3_{\pm 785.9}$	$876.3_{\pm 984.4}$	$+0.008_{\pm 0.007}$	$1165.5_{\pm 1008.8}$	$2304.4_{\pm 1886.8}$

Table 15: Coverage gap (ΔCov), prediction interval width (PI-W), and Winkler score (WS), averaged across sequences within each dataset, for all methods. The target coverage is $1-\alpha = 0.95$.

Method	Air								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
Split CP	$+0.014_{\pm 0.008}$	$136.3_{\pm 14.4}$	$198.8_{\pm 17.4}$	$-0.002_{\pm 0.012}$	$104.3_{\pm 11.2}$	$187.2_{\pm 20.0}$	$+0.002_{\pm 0.005}$	$256.0_{\pm 20.4}$	$422.4_{\pm 29.6}$
NexCP	$-0.002_{\pm 0.001}$	$118.2_{\pm 14.4}$	$185.6_{\pm 19.1}$	$-0.005_{\pm 0.001}$	$107.3_{\pm 13.2}$	$179.6_{\pm 19.6}$	$-0.005_{\pm 0.002}$	$253.1_{\pm 18.0}$	$362.8_{\pm 22.4}$
HopCPT	$-0.014_{\pm .015}$	$108.4_{\pm 18.8}$	$196.9_{\pm 19.8}$	$-0.007_{\pm .009}$	$100.6_{\pm 12.2}$	$181.9_{\pm 18.8}$	$-0.010_{\pm .011}$	$232.2_{\pm 28.0}$	$412.9_{\pm 41.2}$
SPCI	$-0.140_{\pm 0.005}$	$80.3_{\pm 9.4}$	$235.1_{\pm 23.7}$	$-0.139_{\pm 0.005}$	$67.9_{\pm 8.4}$	$208.2_{\pm 24.6}$	$-0.118_{\pm 0.008}$	$86.2_{\pm 7.7}$	$230.9_{\pm 19.2}$
KOWCPI	$+0.004_{\pm 0.006}$	$127.0_{\pm 12.9}$	$199.8_{\pm 17.3}$	$-0.008_{\pm 0.010}$	$103.9_{\pm 10.6}$	$187.3_{\pm 18.0}$	$-0.004_{\pm 0.005}$	$251.5_{\pm 18.7}$	$424.1_{\pm 30.0}$
ResCP	$-0.007_{\pm 0.004}$	$108.4_{\pm 14.0}$	$174.0_{\pm 17.4}$	$-0.014_{\pm 0.004}$	$91.8_{\pm 13.6}$	$157.1_{\pm 16.5}$	$+0.020_{\pm 0.001}$	$165.2_{\pm 12.9}$	$215.5_{\pm 15.1}$
DistMatch	$-0.008_{\pm 0.005}$	$107.4_{\pm 11.5}$	$181.3_{\pm 18.5}$	$-0.008_{\pm 0.005}$	$93.7_{\pm 9.5}$	$166.1_{\pm 17.3}$	$+0.016_{\pm 0.003}$	$153.0_{\pm 12.9}$	$216.4_{\pm 13.9}$
DSCP-QR	$-0.005_{\pm .009}$	$107.4_{\pm 15.0}$	$155.6_{\pm 17.1}$	$-0.027_{\pm .011}$	$78.1_{\pm 7.74}$	$135.8_{\pm 19.0}$	$+0.001_{\pm .013}$	$109.0_{\pm 13.3}$	$159.8_{\pm 13.8}$
DSCP-QFE	$-0.097_{\pm .017}$	$64.8_{\pm 7.03}$	$196.8_{\pm 27.0}$	$-0.153_{\pm .032}$	$40.3_{\pm 4.47}$	$203.7_{\pm 32.7}$	$-0.113_{\pm .035}$	$60.0_{\pm 8.84}$	$282.4_{\pm 22.6}$
DSCP-LCP	$-0.007_{\pm .003}$	$109.1_{\pm 13.5}$	$172.1_{\pm 16.9}$	$-0.006_{\pm .004}$	$93.9_{\pm 12.4}$	$151.4_{\pm 16.8}$	$+0.009_{\pm .003}$	$150.2_{\pm 11.0}$	$208.5_{\pm 14.8}$
Method	Solar								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
Split CP	$-0.003_{\pm 0.013}$	$168.2_{\pm 94.2}$	$267.2_{\pm 99.5}$	$-0.004_{\pm 0.012}$	$141.7_{\pm 27.5}$	$237.0_{\pm 35.4}$	$+0.004_{\pm 0.007}$	$251.9_{\pm 26.2}$	$347.3_{\pm 26.9}$
NexCP	$-0.003_{\pm 0.002}$	$159.8_{\pm 30.1}$	$233.5_{\pm 39.2}$	$-0.005_{\pm 0.002}$	$138.9_{\pm 28.7}$	$215.2_{\pm 34.2}$	$-0.004_{\pm 0.001}$	$225.4_{\pm 30.0}$	$311.6_{\pm 34.1}$
HopCPT	$-0.002_{\pm .027}$	$133.0_{\pm 33.6}$	$185.6_{\pm 57.1}$	$+0.003_{\pm .020}$	$104.0_{\pm 28.6}$	$161.9_{\pm 56.7}$	$+0.001_{\pm .025}$	$152.2_{\pm 28.6}$	$223.5_{\pm 66.4}$
SPCI	$-0.130_{\pm 0.025}$	$81.9_{\pm 13.8}$	$222.5_{\pm 34.1}$	$-0.114_{\pm 0.062}$	$59.6_{\pm 13.5}$	$177.6_{\pm 34.8}$	$-0.107_{\pm 0.015}$	$63.2_{\pm 9.4}$	$172.2_{\pm 23.2}$
KOWCPI	$-0.003_{\pm 0.014}$	$167.3_{\pm 54.0}$	$258.1_{\pm 75.8}$	$-0.027_{\pm 0.019}$	$137.7_{\pm 32.2}$	$230.8_{\pm 35.8}$	$-0.012_{\pm 0.014}$	$236.1_{\pm 34.2}$	$340.0_{\pm 31.6}$
ResCP	$-0.019_{\pm 0.005}$	$118.8_{\pm 26.3}$	$202.9_{\pm 42.3}$	$+0.005_{\pm 0.003}$	$93.3_{\pm 19.2}$	$136.7_{\pm 25.5}$	$+0.009_{\pm 0.002}$	$119.5_{\pm 17.3}$	$165.7_{\pm 19.6}$
DistMatch	$-0.003_{\pm 0.007}$	$138.6_{\pm 30.1}$	$220.7_{\pm 41.8}$	$-0.001_{\pm 0.007}$	$119.0_{\pm 22.1}$	$196.7_{\pm 30.5}$	$+0.007_{\pm 0.004}$	$139.4_{\pm 20.2}$	$203.1_{\pm 25.0}$
DSCP-QR	$-0.011_{\pm .013}$	$114.8_{\pm 41.7}$	$156.3_{\pm 83.1}$	$+0.004_{\pm .018}$	$84.0_{\pm 18.0}$	$112.0_{\pm 29.1}$	$+0.001_{\pm .020}$	$75.9_{\pm 11.3}$	$96.1_{\pm 12.3}$
DSCP-QFE	$-0.160_{\pm 0.110}$	$53.8_{\pm 8.4}$	$327.4_{\pm 753.9}$	$-0.166_{\pm 0.096}$	$28.7_{\pm 6.3}$	$205.4_{\pm 157.7}$	$-0.118_{\pm 0.029}$	$36.7_{\pm 7.4}$	$177.7_{\pm 33.8}$
DSCP-Local	$-0.002_{\pm 0.007}$	$114.4_{\pm 24.8}$	$149.8_{\pm 39.0}$	$+0.012_{\pm 0.007}$	$83.1_{\pm 19.6}$	$107.3_{\pm 20.5}$	$+0.006_{\pm 0.006}$	$106.0_{\pm 16.4}$	$135.6_{\pm 14.2}$
Method	Sapflow								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
Split CP	$+0.021_{\pm 0.015}$	$1204.6_{\pm 887.1}$	$1647.6_{\pm 1202.7}$	$+0.014_{\pm 0.015}$	$1954.4_{\pm 3326.8}$	$2636.1_{\pm 4144.9}$	$+0.008_{\pm 0.015}$	$3632.2_{\pm 2827.5}$	$5062.5_{\pm 3690.9}$
NexCP	$-0.001_{\pm 0.003}$	$902.2_{\pm 596.7}$	$1388.5_{\pm 980.2}$	$-0.003_{\pm 0.004}$	$1539.7_{\pm 2505.4}$	$2113.4_{\pm 2951.2}$	$-0.006_{\pm 0.005}$	$3146.6_{\pm 2514.7}$	$4668.5_{\pm 3559.1}$
HopCPT	$+0.005_{\pm .042}$	$1126.5_{\pm 917.4}$	$1517.5_{\pm 1148.6}$	$-0.005_{\pm .056}$	$1201.2_{\pm 1297.8}$	$1635.3_{\pm 1544.6}$	$-0.005_{\pm .047}$	$3170.3_{\pm 2516.1}$	$4017.6_{\pm 2538.3}$
SPCI	$+0.026_{\pm 0.012}$	$1016.6_{\pm 729.3}$	$1329.8_{\pm 938.5}$	$+0.019_{\pm 0.019}$	$908.4_{\pm 658.6}$	$1213.3_{\pm 862.3}$	$-0.026_{\pm 0.033}$	$1028.4_{\pm 751.4}$	$1759.7_{\pm 1231.7}$
KOWCPI	$-0.006_{\pm 0.061}$	$1141.9_{\pm 899.0}$	$1815.3_{\pm 1306.8}$	$-0.007_{\pm 0.063}$	$1325.0_{\pm 977.9}$	$1878.8_{\pm 1283.5}$	$+0.022_{\pm 0.033}$	$4550.0_{\pm 3750.7}$	$5298.7_{\pm 4118.7}$
ResCP	$-0.003_{\pm 0.014}$	$727.8_{\pm 502.4}$	$1185.2_{\pm 855.9}$	$-0.002_{\pm 0.014}$	$806.5_{\pm 709.5}$	$1244.0_{\pm 1010.4}$	$+0.016_{\pm 0.018}$	$1895.1_{\pm 1501.9}$	$2398.2_{\pm 1666.5}$
DistMatch	$+0.010_{\pm 0.011}$	$882.9_{\pm 624.0}$	$1321.7_{\pm 954.2}$	$+0.015_{\pm 0.011}$	$969.0_{\pm 719.1}$	$1393.9_{\pm 1019.9}$	$+0.023_{\pm 0.011}$	$1867.9_{\pm 1288.1}$	$2404.6_{\pm 1570.9}$
DSCP-QR	$-0.018_{\pm 0.048}$	$931.1_{\pm 675.3}$	$1340.1_{\pm 955.6}$	$-0.002_{\pm 0.046}$	$1235.4_{\pm 1088.7}$	$1800.7_{\pm 2008.7}$	$-0.070_{\pm 0.067}$	$1039.8_{\pm 612.6}$	$2002.1_{\pm 1607.5}$
DSCP-QFE	$-0.092_{\pm 0.098}$	$632.0_{\pm 575.3}$	$1608.2_{\pm 1330.9}$	$-0.137_{\pm 0.143}$	$604.6_{\pm 492.2}$	$1903.0_{\pm 1492.3}$	$-0.160_{\pm 0.074}$	$654.6_{\pm 477.6}$	$3161.2_{\pm 2164.8}$
DSCP-Local	$-0.001_{\pm 0.008}$	$743.0_{\pm 520.8}$	$1111.2_{\pm 827.7}$	$+0.002_{\pm 0.008}$	$833.4_{\pm 716.4}$	$1162.5_{\pm 954.9}$	$+0.008_{\pm 0.008}$	$1783.3_{\pm 1301.5}$	$2269.2_{\pm 1630.1}$

Table 16: Coverage gap (ΔCov), prediction interval width (PI-W), and Winkler score (WS), averaged across sequences within each dataset, for all methods. The target coverage is $1 - \alpha = 0.8$.

Method	Air								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
Split CP	$+0.045_{\pm 0.015}$	$68.4_{\pm 7.9}$	$110.1_{\pm 11.7}$	$-0.001_{\pm 0.017}$	$48.9_{\pm 5.2}$	$98.4_{\pm 11.0}$	$+0.030_{\pm 0.016}$	$125.8_{\pm 13.7}$	$227.9_{\pm 17.0}$
NexCP	$+0.001_{\pm 0.002}$	$61.0_{\pm 7.6}$	$106.7_{\pm 12.0}$	$-0.005_{\pm 0.002}$	$51.0_{\pm 5.9}$	$97.3_{\pm 10.9}$	$-0.005_{\pm 0.002}$	$118.0_{\pm 10.8}$	$214.8_{\pm 15.6}$
HopCPT	$-0.023_{\pm .028}$	$55.1_{\pm 8.32}$	$107.7_{\pm 11.8}$	$-0.033_{\pm .033}$	$44.9_{\pm 4.87}$	$96.9_{\pm 10.6}$	$-0.010_{\pm .025}$	$105.7_{\pm 17.8}$	$219.6_{\pm 29.9}$
SPCI	$-0.151_{\pm 0.008}$	$50.2_{\pm 6.1}$	$118.1_{\pm 13.4}$	$-0.146_{\pm 0.008}$	$40.4_{\pm 5.1}$	$99.7_{\pm 12.5}$	$-0.113_{\pm 0.011}$	$53.2_{\pm 4.9}$	$116.7_{\pm 10.3}$
KOWCPI	$+0.019_{\pm 0.012}$	$63.6_{\pm 7.4}$	$110.2_{\pm 11.7}$	$-0.011_{\pm 0.012}$	$48.0_{\pm 5.2}$	$98.5_{\pm 10.8}$	$+0.012_{\pm 0.012}$	$118.6_{\pm 12.2}$	$227.4_{\pm 17.1}$
ResCP	$-0.021_{\pm 0.007}$	$56.8_{\pm 7.3}$	$103.4_{\pm 11.5}$	$-0.025_{\pm 0.006}$	$46.0_{\pm 6.5}$	$89.4_{\pm 10.5}$	$+0.034_{\pm 0.006}$	$89.7_{\pm 7.7}$	$135.9_{\pm 10.6}$
DistMatch	$-0.004_{\pm 0.007}$	$59.9_{\pm 6.7}$	$105.3_{\pm 11.6}$	$-0.004_{\pm 0.009}$	$48.7_{\pm 5.3}$	$92.7_{\pm 10.3}$	$+0.068_{\pm 0.007}$	$86.1_{\pm 7.4}$	$129.3_{\pm 9.9}$
DSCP-QR	$+0.011_{\pm 0.019}$	$60.9_{\pm 7.9}$	$93.7_{\pm 10.9}$	$-0.022_{\pm 0.028}$	$43.1_{\pm 6.9}$	$77.7_{\pm 9.5}$	$+0.016_{\pm 0.026}$	$62.5_{\pm 8.1}$	$97.4_{\pm 7.8}$
DSCP-QFE	$-0.002_{\pm 0.022}$	$55.0_{\pm 5.8}$	$96.4_{\pm 10.3}$	$-0.073_{\pm 0.022}$	$33.1_{\pm 4.1}$	$81.9_{\pm 9.7}$	$-0.032_{\pm 0.034}$	$45.4_{\pm 6.6}$	$109.8_{\pm 8.6}$
DSCP-Local	$-0.009_{\pm 0.006}$	$58.6_{\pm 7.5}$	$102.0_{\pm 11.3}$	$-0.002_{\pm 0.004}$	$47.3_{\pm 6.4}$	$85.9_{\pm 10.7}$	$+0.048_{\pm 0.008}$	$86.1_{\pm 6.9}$	$123.4_{\pm 9.0}$
Method	Solar								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
Split CP	$+0.001_{\pm 0.039}$	$72.0_{\pm 85.2}$	$146.7_{\pm 78.3}$	$+0.004_{\pm 0.032}$	$41.3_{\pm 33.7}$	$119.1_{\pm 31.9}$	$+0.014_{\pm 0.013}$	$79.4_{\pm 16.0}$	$189.1_{\pm 24.8}$
NexCP	$+0.003_{\pm 0.002}$	$68.2_{\pm 13.9}$	$132.2_{\pm 24.6}$	$+0.001_{\pm 0.002}$	$44.8_{\pm 12.4}$	$111.2_{\pm 22.5}$	$+0.002_{\pm 0.002}$	$82.5_{\pm 15.9}$	$180.0_{\pm 24.8}$
HopCPT	$+0.004_{\pm .038}$	$78.6_{\pm 18.9}$	$117.4_{\pm 25.3}$	$-0.007_{\pm .054}$	$52.8_{\pm 19.8}$	$92.3_{\pm 29.1}$	$0.004_{\pm .048}$	$79.0_{\pm 22.9}$	$141.0_{\pm 39.7}$
SPCI	$-0.148_{\pm 0.033}$	$52.4_{\pm 9.5}$	$120.1_{\pm 19.9}$	$-0.118_{\pm 0.092}$	$33.9_{\pm 7.7}$	$87.9_{\pm 17.7}$	$-0.103_{\pm 0.025}$	$40.4_{\pm 6.6}$	$92.4_{\pm 13.9}$
KOWCPI	$-0.003_{\pm 0.015}$	$67.7_{\pm 44.5}$	$141.9_{\pm 51.5}$	$-0.022_{\pm 0.024}$	$39.3_{\pm 18.5}$	$116.1_{\pm 24.7}$	$-0.005_{\pm 0.014}$	$74.6_{\pm 17.9}$	$185.0_{\pm 25.7}$
ResCP	$-0.041_{\pm 0.009}$	$56.5_{\pm 11.9}$	$118.6_{\pm 23.2}$	$+0.006_{\pm 0.006}$	$42.4_{\pm 9.5}$	$80.8_{\pm 16.6}$	$+0.009_{\pm 0.005}$	$56.9_{\pm 9.7}$	$97.9_{\pm 13.8}$
DistMatch	$+0.012_{\pm 0.008}$	$63.3_{\pm 12.4}$	$124.2_{\pm 23.5}$	$+0.035_{\pm 0.011}$	$42.6_{\pm 10.3}$	$99.7_{\pm 19.5}$	$+0.040_{\pm 0.007}$	$57.0_{\pm 9.8}$	$109.9_{\pm 16.0}$
DSCP-QR	$-0.015_{\pm 0.071}$	$62.3_{\pm 10.1}$	$95.1_{\pm 38.9}$	$-0.009_{\pm 0.113}$	$42.0_{\pm 9.1}$	$72.5_{\pm 95.5}$	$+0.012_{\pm 0.036}$	$48.0_{\pm 8.5}$	$68.1_{\pm 9.8}$
DSCP-QFE	$-0.078_{\pm 0.101}$	$43.6_{\pm 6.7}$	$108.4_{\pm 112.3}$	$-0.060_{\pm 0.108}$	$21.0_{\pm 4.6}$	$76.2_{\pm 108.4}$	$-0.047_{\pm 0.046}$	$29.0_{\pm 5.6}$	$70.4_{\pm 12.3}$
DSCP-Local	$-0.002_{\pm 0.008}$	$66.3_{\pm 13.1}$	$100.3_{\pm 21.4}$	$+0.006_{\pm 0.012}$	$44.5_{\pm 10.9}$	$68.6_{\pm 14.3}$	$+0.008_{\pm 0.009}$	$59.7_{\pm 11.1}$	$86.7_{\pm 11.6}$
Method	Sapflow								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
Split CP	$+0.036_{\pm 0.042}$	$480.7_{\pm 339.8}$	$842.9_{\pm 572.8}$	$+0.024_{\pm 0.037}$	$665.4_{\pm 861.6}$	$1461.1_{\pm 2463.1}$	$-0.016_{\pm 0.085}$	$1326.2_{\pm 1165.5}$	$2739.2_{\pm 2111.8}$
NexCP	$+0.004_{\pm 0.005}$	$436.3_{\pm 279.3}$	$794.3_{\pm 532.5}$	$+0.001_{\pm 0.007}$	$750.5_{\pm 1249.2}$	$1320.6_{\pm 2097.6}$	$+0.005_{\pm 0.006}$	$1254.8_{\pm 1067.8}$	$2600.6_{\pm 2072.9}$
HopCPT	$+0.025_{\pm .078}$	$531.4_{\pm 370.7}$	$804.4_{\pm 525.9}$	$-0.013_{\pm .099}$	$587.2_{\pm 529.4}$	$982.3_{\pm 807.2}$	$-0.026_{\pm .090}$	$1460.1_{\pm 1348.7}$	$2474.0_{\pm 1941.6}$
SPCI	$+0.058_{\pm 0.043}$	$444.2_{\pm 292.4}$	$708.8_{\pm 469.4}$	$+0.030_{\pm 0.035}$	$424.0_{\pm 299.6}$	$679.3_{\pm 468.9}$	$-0.042_{\pm 0.100}$	$550.5_{\pm 417.7}$	$999.5_{\pm 684.9}$
KOWCPI	$+0.033_{\pm 0.102}$	$538.2_{\pm 410.8}$	$893.1_{\pm 606.6}$	$+0.017_{\pm 0.082}$	$631.4_{\pm 599.8}$	$1025.9_{\pm 773.1}$	$+0.086_{\pm 0.113}$	$2053.3_{\pm 1864.2}$	$2689.9_{\pm 2130.5}$
ResCP	$-0.013_{\pm 0.035}$	$363.3_{\pm 240.0}$	$681.4_{\pm 463.5}$	$-0.008_{\pm 0.036}$	$401.7_{\pm 363.5}$	$739.3_{\pm 611.7}$	$+0.006_{\pm 0.027}$	$905.4_{\pm 748.5}$	$1432.2_{\pm 1033.1}$
DistMatch	$+0.034_{\pm 0.024}$	$431.2_{\pm 289.4}$	$722.8_{\pm 486.0}$	$+0.047_{\pm 0.023}$	$459.4_{\pm 358.0}$	$758.8_{\pm 555.6}$	$+0.060_{\pm 0.033}$	$949.3_{\pm 723.1}$	$1403.5_{\pm 975.5}$
DSCP-QR	$-0.030_{\pm 0.104}$	$491.2_{\pm 325.1}$	$766.9_{\pm 518.2}$	$-0.029_{\pm 0.141}$	$588.5_{\pm 455.5}$	$934.6_{\pm 797.3}$	$-0.055_{\pm 0.087}$	$630.1_{\pm 395.9}$	$1122.2_{\pm 833.4}$
DSCP-QFE	$-0.030_{\pm 0.104}$	$491.2_{\pm 325.1}$	$766.9_{\pm 518.2}$	$-0.029_{\pm 0.141}$	$588.5_{\pm 455.5}$	$934.6_{\pm 797.3}$	$-0.055_{\pm 0.087}$	$630.1_{\pm 395.9}$	$1122.2_{\pm 833.4}$
DSCP-LCP	$+0.001_{\pm 0.014}$	$391.9_{\pm 253.6}$	$663.1_{\pm 458.2}$	$+0.013_{\pm 0.019}$	$446.2_{\pm 398.7}$	$700.8_{\pm 570.7}$	$+0.014_{\pm 0.015}$	$953.8_{\pm 731.0}$	$1445.1_{\pm 1064.5}$

Table 17: Ablation study of architectural constraints for preventing quantile crossing in DSCP-QR and DSCP-QFE. The suffix "-C" denotes the corresponding method without these constraints. CR denotes the quantile crossing occurrence rate across all sequences.

Method	Air											
	LR				LSTM				Chronos			
	ΔCov	PI-W	WS	CR	ΔCov	PI-W	WS	CR	ΔCov	PI-W	WS	CR
QR-RNN	$+0.012_{\pm .016}$	$85.3_{\pm 10.5}$	$121.3_{\pm 12.5}$	0.000%	$-0.021_{\pm .017}$	$61.1_{\pm 9.32}$	$103.4_{\pm 12.1}$	0.000%	$+0.014_{\pm .014}$	$87.1_{\pm 10.7}$	$125.5_{\pm 10.6}$	0.000%
QR-Transformer	$-0.004_{\pm .014}$	$92.3_{\pm 14.7}$	$136.3_{\pm 18.5}$	0.000%	$-0.018_{\pm .018}$	$69.4_{\pm 10.6}$	$124.2_{\pm 16.0}$	0.000%	$-0.002_{\pm .013}$	$93.5_{\pm 9.82}$	$146.3_{\pm 13.2}$	0.000%
QFE-RNN	$-0.064_{\pm .015}$	$61.3_{\pm 6.40}$	$132.7_{\pm 15.6}$	0.000%	$-0.126_{\pm .026}$	$37.9_{\pm 4.73}$	$124.4_{\pm 16.3}$	0.000%	$-0.091_{\pm .032}$	$52.5_{\pm 7.39}$	$170.1_{\pm 13.1}$	0.000%
QFE-Transformer	$-0.084_{\pm .021}$	$63.4_{\pm 6.78}$	$152.7_{\pm 19.8}$	0.000%	$-0.112_{\pm .028}$	$45.7_{\pm 6.78}$	$143.4_{\pm 22.3}$	0.000%	$-0.096_{\pm .022}$	$58.2_{\pm 6.31}$	$196.1_{\pm 17.1}$	0.000%
QR-RNN-C	$-0.008_{\pm 0.018}$	$80.7_{\pm 10.1}$	$121.6_{\pm 13.2}$	0.000%	$-0.025_{\pm 0.023}$	$60.5_{\pm 10.4}$	$102.7_{\pm 12.6}$	0.000%	$+0.003_{\pm 0.013}$	$84.3_{\pm 9.1}$	$126.1_{\pm 11.1}$	0.013%
QR-Transformer-C	$+0.008_{\pm 0.021}$	$95.5_{\pm 13.6}$	$137.3_{\pm 15.5}$	0.029%	$-0.030_{\pm 0.017}$	$67.6_{\pm 8.6}$	$123.0_{\pm 15.2}$	0.047%	$-0.010_{\pm 0.023}$	$92.6_{\pm 12.3}$	$145.5_{\pm 12.2}$	0.013%
QFE-RNN-C	$-0.020_{\pm 0.021}$	$76.1_{\pm 9.4}$	$124.4_{\pm 14.4}$	0.007%	$-0.047_{\pm 0.026}$	$54.1_{\pm 7.1}$	$109.4_{\pm 11.8}$	0.001%	$-0.006_{\pm 0.019}$	$78.0_{\pm 9.6}$	$126.9_{\pm 10.4}$	0.001%
QFE-Transformer-C	$-0.036_{\pm 0.038}$	$76.6_{\pm 12.4}$	$149.8_{\pm 16.5}$	0.000%	$-0.027_{\pm 0.015}$	$62.3_{\pm 6.9}$	$124.6_{\pm 15.8}$	0.010%	$-0.002_{\pm 0.017}$	$89.0_{\pm 10.3}$	$162.3_{\pm 13.3}$	0.000%
Method	Solar											
	LR				LSTM				Chronos			
	ΔCov	PI-W	WS	CR	ΔCov	PI-W	WS	CR	ΔCov	PI-W	WS	CR
QR-RNN	$-0.019_{\pm .107}$	$81.9_{\pm 14.6}$	$121.6_{\pm 82.9}$	0.000%	$-0.006_{\pm .120}$	$55.4_{\pm 10.8}$	$84.0_{\pm 84.0}$	0.000%	$+0.007_{\pm .020}$	$62.9_{\pm 9.97}$	$82.4_{\pm 11.5}$	0.000%
QR-Transformer	$-0.031_{\pm .121}$	$87.3_{\pm 16.8}$	$150.1_{\pm 186.5}$	0.000%	$-0.005_{\pm .028}$	$62.1_{\pm 12.9}$	$90.7_{\pm 24.7}$	0.000%	$-0.001_{\pm .023}$	$75.3_{\pm 14.0}$	$102.3_{\pm 13.9}$	0.000%
QFE-RNN	$-0.114_{\pm .019}$	$55.9_{\pm 9.50}$	$159.7_{\pm 41.8}$	0.000%	$-0.120_{\pm .061}$	$33.4_{\pm 8.74}$	$122.8_{\pm 37.2}$	0.000%	$-0.103_{\pm .035}$	$42.1_{\pm 8.71}$	$123.3_{\pm 15.6}$	0.000%
QFE-Transformer	$-0.165_{\pm .098}$	$49.8_{\pm 7.16}$	$203.8_{\pm 144.5}$	0.000%	$-0.152_{\pm .105}$	$25.9_{\pm 5.84}$	$172.7_{\pm 194.0}$	0.000%	$-0.127_{\pm .023}$	$37.3_{\pm 8.23}$	$152.4_{\pm 16.5}$	0.000%
QR-RNN-C	$-0.018_{\pm 0.126}$	$85.4_{\pm 16.1}$	$196.4_{\pm 585.9}$	1.704%	$-0.003_{\pm 0.030}$	$57.8_{\pm 10.6}$	$75.3_{\pm 12.9}$	2.483%	$-0.006_{\pm 0.031}$	$64.8_{\pm 11.1}$	$84.5_{\pm 11.2}$	0.194%
QR-Transformer-C	$-0.006_{\pm 0.025}$	$99.2_{\pm 60.8}$	$131.4_{\pm 62.7}$	0.010%	$+0.008_{\pm 0.035}$	$64.5_{\pm 15.3}$	$90.7_{\pm 21.7}$	0.335%	$-0.005_{\pm 0.029}$	$74.9_{\pm 13.4}$	$101.9_{\pm 13.6}$	0.370%
QFE-RNN-C	$-0.033_{\pm 0.111}$	$59.5_{\pm 10.9}$	$99.1_{\pm 84.3}$	0.007%	$-0.027_{\pm 0.037}$	$28.6_{\pm 7.8}$	$44.8_{\pm 9.5}$	2.006%	$-0.028_{\pm 0.038}$	$40.2_{\pm 6.7}$	$63.8_{\pm 7.9}$	0.005%
QFE-Transformer-C	$-0.039_{\pm 0.124}$	$80.7_{\pm 12.4}$	$189.2_{\pm 397.5}$	0.000%	$-0.023_{\pm 0.120}$	$68.7_{\pm 15.2}$	$136.4_{\pm 107.8}$	0.359%	$-0.005_{\pm 0.023}$	$68.6_{\pm 13.1}$	$104.9_{\pm 14.2}$	0.077%
Method	Sapflow											
	LR				LSTM				Chronos			
	ΔCov	PI-W	WS	CR	ΔCov	PI-W	WS	CR	ΔCov	PI-W	WS	CR
QR-RNN	$+0.005_{\pm .064}$	$708.5_{\pm 511.9}$	$994.5_{\pm 687.8}$	0.000%	$-0.008_{\pm .111}$	$839.3_{\pm 917.8}$	$1142.3_{\pm 1023.3}$	0.000%	$-0.061_{\pm .065}$	$864.8_{\pm 549.1}$	$1450.2_{\pm 953.0}$	0.000%
QR-Transformer	$-0.014_{\pm .117}$	$766.1_{\pm 450.4}$	$1151.1_{\pm 726.4}$	0.000%	$-0.077_{\pm .162}$	$1339.0_{\pm 2828.9}$	$1911.9_{\pm 3030.8}$	0.000%	$-0.092_{\pm .085}$	$1079.5_{\pm 720.9}$	$2296.8_{\pm 1915.4}$	0.000%
QFE-RNN	$-0.056_{\pm .089}$	$569.1_{\pm 369.1}$	$1078.2_{\pm 753.2}$	0.000%	$-0.115_{\pm .128}$	$527.0_{\pm 461.3}$	$1188.3_{\pm 884.6}$	0.000%	$-0.137_{\pm .101}$	$563.4_{\pm 442.8}$	$1885.1_{\pm 1216.4}$	0.000%
QFE-Transformer	$-0.046_{\pm .116}$	$609.6_{\pm 421.5}$	$1209.8_{\pm 829.9}$	0.000%	$-0.083_{\pm .143}$	$665.3_{\pm 670.8}$	$1436.8_{\pm 1262.0}$	0.000%	$-0.141_{\pm .093}$	$838.4_{\pm 777.8}$	$2602.0_{\pm 2046.1}$	0.000%
QR-RNN-C	$-0.035_{\pm 0.107}$	$722.9_{\pm 588.9}$	$1059.8_{\pm 730.8}$	0.373%	$-0.112_{\pm 0.154}$	$740.7_{\pm 702.2}$	$1397.7_{\pm 1643.2}$	1.939%	$-0.068_{\pm 0.080}$	$868.0_{\pm 601.0}$	$1428.2_{\pm 997.4}$	0.416%
QR-Transformer-C	$-0.065_{\pm 0.150}$	$754.2_{\pm 495.1}$	$1235.4_{\pm 785.2}$	0.008%	$-0.091_{\pm 0.153}$	$768.1_{\pm 709.2}$	$1531.8_{\pm 1625.4}$	0.569%	$-0.051_{\pm 0.058}$	$1144.4_{\pm 939.8}$	$2016.2_{\pm 1409.2}$	0.338%
QFE-RNN-C	$-0.128_{\pm 0.137}$	$545.1_{\pm 357.8}$	$1151.2_{\pm 965.3}$	0.383%	$-0.136_{\pm 0.134}$	$542.2_{\pm 424.7}$	$1128.8_{\pm 815.8}$	0.223%	$-0.121_{\pm 0.118}$	$801.9_{\pm 623.6}$	$1576.9_{\pm 1392.8}$	0.187%
QFE-Transformer-C	$-0.245_{\pm 0.174}$	$470.3_{\pm 358.5}$	$1581.2_{\pm 1414.0}$	0.259%	$-0.188_{\pm 0.199}$	$481.3_{\pm 319.7}$	$1794.4_{\pm 2675.3}$	0.235%	$-0.156_{\pm 0.109}$	$713.8_{\pm 506.9}$	$2439.2_{\pm 2019.0}$	0.144%

Table 18: Coverage gap (ΔCov), prediction interval width (PI-W), and Winkler score (WS), averaged across sequences within each dataset, for DSCP-Local using RNN and Transformer encoders with three sets of target quantile levels: the median only (0.5), three quantile levels $(0.1, 0.5, 0.9)$, and nine quantile levels $(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$. The target coverage is $1-\alpha = 0.9$.

Method	Air								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
DSCP-RNN-Local-Median	$-0.017_{\pm 0.005}$	$80.8_{\pm 10.4}$	$137.0_{\pm 14.6}$	$-0.019_{\pm 0.006}$	$66.0_{\pm 9.2}$	$119.5_{\pm 14.7}$	$+0.029_{\pm 0.006}$	$120.1_{\pm 10.2}$	$163.9_{\pm 11.8}$
DSCP-RNN-Local-3	$-0.015_{\pm .003}$	$81.0_{\pm 10.0}$	$133.2_{\pm 14.4}$	$-0.013_{\pm .005}$	$67.8_{\pm 8.67}$	$116.0_{\pm 13.3}$	$+0.027_{\pm .006}$	$117.4_{\pm 9.14}$	$160.7_{\pm 11.4}$
DSCP-RNN-Local-9	$-0.006_{\pm 0.004}$	$82.4_{\pm 10.3}$	$134.3_{\pm 14.8}$	$-0.004_{\pm 0.004}$	$67.7_{\pm 9.0}$	$114.0_{\pm 13.6}$	$+0.033_{\pm 0.006}$	$118.3_{\pm 10.9}$	$160.4_{\pm 12.3}$
DSCP-Transformer-Local-Median	$-0.013_{\pm 0.003}$	$81.1_{\pm 10.4}$	$136.1_{\pm 15.5}$	$-0.009_{\pm 0.003}$	$69.4_{\pm 8.2}$	$121.7_{\pm 13.6}$	$+0.011_{\pm 0.005}$	$135.1_{\pm 12.8}$	$188.6_{\pm 14.7}$
DSCP-Transformer-Local-3	$-0.007_{\pm .004}$	$83.4_{\pm 9.60}$	$140.8_{\pm 15.5}$	$-0.007_{\pm .003}$	$71.8_{\pm 9.29}$	$127.6_{\pm 15.1}$	$+0.015_{\pm .005}$	$130.4_{\pm 11.9}$	$179.9_{\pm 14.2}$
DSCP-Transformer-Local-9	$-0.012_{\pm 0.002}$	$80.8_{\pm 10.3}$	$135.8_{\pm 15.2}$	$-0.008_{\pm 0.003}$	$69.5_{\pm 8.4}$	$123.0_{\pm 13.6}$	$+0.017_{\pm 0.003}$	$127.9_{\pm 11.6}$	$178.4_{\pm 13.7}$
Method	Solar								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
DSCP-RNN-Local-Median	$-0.013_{\pm 0.009}$	$89.6_{\pm 19.3}$	$141.7_{\pm 30.2}$	$-0.003_{\pm 0.009}$	$61.0_{\pm 15.7}$	$117.1_{\pm 20.5}$	$+0.007_{\pm 0.007}$	$82.3_{\pm 13.5}$	$116.2_{\pm 13.8}$
DSCP-RNN-Local-3	$-0.001_{\pm .009}$	$91.0_{\pm 18.9}$	$124.3_{\pm 28.7}$	$+0.016_{\pm .010}$	$64.0_{\pm 15.3}$	$87.0_{\pm 17.6}$	$+0.011_{\pm .007}$	$83.1_{\pm 14.3}$	$109.9_{\pm 13.2}$
DSCP-RNN-Local-9	$+0.001_{\pm 0.009}$	$91.5_{\pm 19.1}$	$125.9_{\pm 31.6}$	$+0.011_{\pm 0.010}$	$62.3_{\pm 15.9}$	$87.1_{\pm 17.8}$	$+0.010_{\pm 0.007}$	$83.3_{\pm 14.2}$	$110.4_{\pm 13.1}$
DSCP-Transformer-Local-Median	$+0.009_{\pm 0.007}$	$93.6_{\pm 18.9}$	$138.1_{\pm 27.6}$	$+0.005_{\pm 0.007}$	$65.0_{\pm 15.9}$	$113.2_{\pm 19.8}$	$-0.000_{\pm 0.009}$	$86.3_{\pm 15.6}$	$145.0_{\pm 17.6}$
DSCP-Transformer-Local-3	$-0.001_{\pm .008}$	$88.6_{\pm 18.4}$	$123.8_{\pm 27.6}$	$+0.002_{\pm .009}$	$60.7_{\pm 15.7}$	$105.7_{\pm 19.0}$	$+0.001_{\pm .011}$	$75.9_{\pm 12.4}$	$122.6_{\pm 14.3}$
DSCP-Transformer-Local-9	$+0.003_{\pm 0.008}$	$88.5_{\pm 18.4}$	$130.9_{\pm 30.7}$	$+0.002_{\pm 0.007}$	$60.2_{\pm 16.2}$	$110.4_{\pm 20.0}$	$+0.007_{\pm 0.009}$	$79.2_{\pm 13.1}$	$128.7_{\pm 13.6}$
Method	Sapflow								
	LR			LSTM			Chronos		
	ΔCov	PI-W	WS	ΔCov	PI-W	WS	ΔCov	PI-W	WS
DSCP-RNN-Local-Median	$-0.004_{\pm 0.008}$	$542.9_{\pm 348.2}$	$878.3_{\pm 605.7}$	$+0.007_{\pm 0.010}$	$625.7_{\pm 527.0}$	$941.8_{\pm 746.5}$	$+0.016_{\pm 0.012}$	$1512.6_{\pm 1199.0}$	$2040.5_{\pm 1520.1}$
DSCP-RNN-Local-3	$+0.001_{\pm .012}$	$558.2_{\pm 375.5}$	$864.2_{\pm 619.6}$	$+0.007_{\pm .013}$	$632.1_{\pm 550.7}$	$911.4_{\pm 743.6}$	$+0.017_{\pm .010}$	$1400.5_{\pm 1102.3}$	$1860.4_{\pm 1382.5}$
DSCP-RNN-Local-9	$-0.001_{\pm 0.014}$	$546.3_{\pm 356.4}$	$856.3_{\pm 609.4}$	$+0.008_{\pm 0.014}$	$616.9_{\pm 526.0}$	$905.2_{\pm 737.6}$	$+0.025_{\pm 0.013}$	$1370.2_{\pm 1075.7}$	$1771.3_{\pm 1321.5}$
DSCP-Transformer-Local-Median	$+0.009_{\pm 0.007}$	$589.2_{\pm 384.4}$	$916.0_{\pm 639.3}$	$+0.007_{\pm 0.007}$	$868.0_{\pm 1172.2}$	$1207.7_{\pm 1387.7}$	$+0.010_{\pm 0.010}$	$2031.6_{\pm 1703.9}$	$3242.6_{\pm 2676.1}$
DSCP-Transformer-Local-3	$+0.009_{\pm .006}$	$586.1_{\pm 380.5}$	$912.5_{\pm 634.6}$	$+0.007_{\pm .008}$	$898.3_{\pm 1357.8}$	$1245.6_{\pm 1264.3}$	$+0.007_{\pm .005}$	$1927.9_{\pm 1557.9}$	$3044.2_{\pm 2340.9}$
DSCP-Transformer-Local-9	$+0.009_{\pm 0.007}$	$585.5_{\pm 380.2}$	$913.0_{\pm 635.2}$	$+0.006_{\pm 0.007}$	$870.7_{\pm 1204.3}$	$1220.5_{\pm 1439.9}$	$+0.007_{\pm 0.007}$	$1856.9_{\pm 1535.2}$	$2953.7_{\pm 2358.1}$