RARECP: REGIME-AWARE RETRIEVAL FOR EFFICIENT CONFORMAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction has become increasingly popular for uncertainty quantification, providing distribution-free coverage guarantees and extensions that remain applicable to sequential and time-series data. Yet, broad common protocol comparisons between conformal methods and the native uncertainty estimates of recent time series foundation models remain limited. We show that, for one-step prediction intervals, existing conformal approaches can be substantially less efficient than native intervals. To close this gap, we introduce *RareCP*, a weighted conformal prediction method with a retrieval architecture tailored to learning recurring residual regimes and adapting to the current forecasting context. We characterize conditions under which calibration consistently estimates the fitted model's population quantiles, isolating its remaining learning and approximation error. We additionally use adaptive conformal inference (ACI) for long-run empirical coverage control under distribution drift.

1 INTRODUCTION

Point forecasts are often insufficient when decisions depend on the range and asymmetry of plausible outcomes. Although time series foundation models (TSFMs) provide probabilistic forecasts, their model-specific estimates generally lack distribution-free coverage guarantees. Conformal prediction (CP) instead calibrates any fixed predictor on held-out data and provides finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Lei et al., 2018). Prior work has conformalized or calibrated TSFM forecasts (Achour et al., 2025; Adler et al., 2026; Cho & Matteson, 2026), but broad matched comparisons remain scarce. Moreover, learned conformal methods are typically fitted separately per dataset and cannot share information.

Temporal dependence and distribution shift typically invalidate the exchangeability assumption used by classical split CP. Online methods respond to observed coverage errors (Gibbs & Candes, 2021; Gibbs & Candès, 2024), while weighted methods emphasize recent or contextually similar residuals (Barber et al., 2023; Lee et al., 2025; Auer et al., 2023; Neglia et al., 2025). However, residual

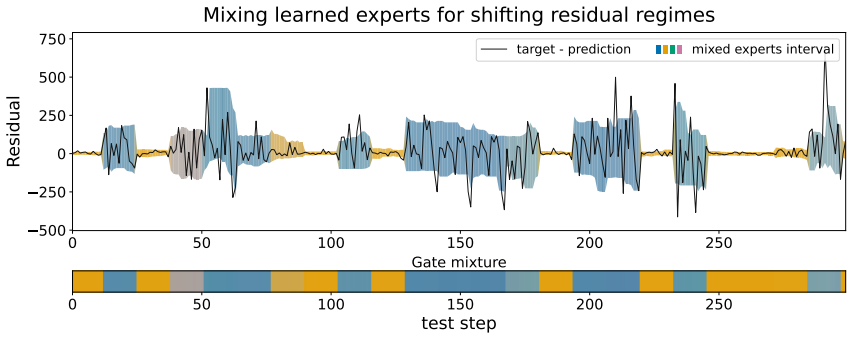


Figure 1: RareCP's nominal 80% prediction interval on `m4_daily`. The gate combines expert residual distributions with weights that vary across histories. The resulting interval narrows in calm periods and widens in volatile periods.

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.

regimes can recur: recency weighting may miss distant recurrences, whereas dense similarity can mix incompatible regimes. This motivates learning sparse, dataset-adaptive residual neighborhoods, while the retrieved examples and expert weights make each interval's construction more inspectible.

We introduce *RareCP*, a residual calibration method trained jointly across datasets. Dataset-conditioned affine maps encode queries and calibration histories, while sparse retrieval experts identify relevant signed residuals. A history-dependent gate mixes their distributions before extracting lower and upper quantiles. The resulting asymmetric intervals adapt to the current history and can be traced to retrieved residuals and expert contributions (Figure 1).

Our main contributions are:

- **A matched uncertainty benchmark.** We compare 13 CP methods and nine native TSFM intervals across 22 GIFT-Eval datasets on four forecasting backbones under a common one-step-ahead protocol.
- **Joint learning of residual regimes.** RareCP combines dataset-conditioned retrieval, sparse residual neighborhoods, and history-dependent expert mixing in a model trained jointly across datasets.
- **Theoretical analysis of residual retrieval.** We establish conditions under which calibration error vanishes as the retrieval pool grows, isolating the fitted model's learning and approximation error, and illustrate the resulting locality–sample-size trade-off empirically.

2 Related work

Online conformal prediction. Conformal prediction provides distribution-free, finite sample coverage under exchangeability (Vovk et al., 2005; Lei et al., 2018), which temporal dependence and drift violate. Online variants update the miscoverage level or threshold: ACI uses a fixed step size (Gibbs & Candes, 2021); DtACI aggregates several rates (Gibbs & Candès, 2024); SF-OGD and SAOCP use scale-free and adaptive online learning (Bhatnagar et al., 2023); and decaying step sizes recover the true quantile for i.i.d. scores (Angelopoulos et al., 2024). Conformal PID, ECI, and COP respectively add integral and scorecasting terms (Angelopoulos et al., 2023), smooth error quantification (Wu et al., 2025), and an optimistic update based on the estimated score distribution (Hu et al., 2026). These methods primarily adapt through temporal feedback, whereas RareCP learns a retrieval geometry over historical forecasting contexts.

Local and learned calibration weights. Localized CP emphasizes calibration scores near the test point (Guan, 2023), since exact distribution-free conditional coverage is unattainable without further assumptions (Barber et al., 2020). For time series, NexCP weights by recency (Barber et al., 2023), while KOWCPI uses kernel-based weights to estimate residual score quantiles and provides guarantees under strong mixing (Lee et al., 2025). Learned schemes retrieve errors using Hopfield attention (Auer et al., 2023), reservoir embeddings (Neglia et al., 2025), or similarities between mixture-of-experts gating vectors (Kong et al., 2026). MoE-CP reuses the point predictor's gate as a fixed representation for a single dense calibration weighting rule. Its experts belong to the point predictor rather than to the calibration mechanism. RareCP instead wraps any fixed point forecaster with multiple sparse retrieval experts and a separate gate that mixes their residual distributions. EnbPI and SPCI instead model the residual distribution through bootstrap ensembles and sequential conditional estimation (Xu & Xie, 2021; 2023). Existing schemes generally use one globally shared similarity function, whereas RareCP generates sample-dependent maps. Its top-K_n signed residuals yield asymmetric intervals without the quantile models of conformalized quantile regression (Romano et al., 2019).

Time series foundation models. TSFMs often provide native quantiles in zero-shot forecasting. Representative models include TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024; 2025), Moirai (Woo et al., 2024; Liu et al., 2026), Toto (Cohen et al., 2024; Khwaja et al., 2026), PatchTST-FM (Wen et al., 2026), TimeGPT (Garza et al., 2024), TiRex (Auer et al., 2025; Podest et al., 2026), TTM (Ekambaram et al., 2024), and TabPFN-TS-3 (Hoo et al., 2025; Grinsztajn et al., 2026). Their native intervals lack distribution-free finite sample coverage guarantees, while point-only forecasters provide no intervals. RareCP handles both cases by wrapping any fixed point forecaster, whether a TSFM or a classical method such as ARIMA (Box & Jenkins, 1970).

Hypernetworks and gated mixtures. Hypernetworks generate target network parameters from a conditioning input (Ha et al., 2016), typically a task or client descriptor (Shamsian et al., 2021). RareCP likewise conditions on a dataset descriptor, but also on the history being encoded. The resulting DCA adapts retrieval to each dataset while generating a sample-specific affine map for every query and calibration history. A trained gate then combines the experts' residual distributions, following the mixture-of-experts principle (Jacobs et al., 1991).

3 METHOD

RareCP constructs asymmetric prediction intervals around a fixed forecasting backbone by retrieving residuals from relevant forecasting contexts. Dataset-conditioned affine maps allow joint training across datasets, while a gate combines experts with different residual neighborhoods.

Weighted residual calibration. Let μ be a fixed point predictor and $\{(H_i, Y_i)\}_{i=1}^n$ the calibration examples drawn from the observed past across all series of the query's dataset. Our jointly trained model supports dataset-wide calibration, making relevant residuals from other series available through a shared learned retrieval mechanism. Each H_i contains the history available for predicting Y_i, with signed residual $E_i = Y_i - \mu(H_i)$. For a query history h, nonnegative weights $w_{i,n}(h)$ summing to one define the weighted residual cumulative distribution function (CDF) and its quantiles:

$$\widehat{F}_n(r \mid h) = \sum_{i=1}^{n} w_{i,n}(h)\mathbf{1}\{E_i \le r\}, \tag{1}$$

$$\widehat{Q}_{p,n}(h) = \inf\{r : \widehat{F}_n(r \mid h) \ge p\}. \tag{2}$$

For miscoverage level $\alpha \in (0,1)$, write $p_- = \alpha/2$ and $p_+ = 1 - \alpha/2$. The prediction interval is

$$\widehat{C}_{n,\alpha}(h) = \big[\mu(h) + \widehat{Q}_{p_-,n}(h),\ \mu(h) + \widehat{Q}_{p_+,n}(h)\big]. \tag{3}$$

Signed residuals allow the lower and upper tails to adapt separately.

3.1 ARCHITECTURE AND TRAINING

RareCP combines M retrieval experts, each learning a similarity rule for selecting and weighting calibration residuals. A gate determines their contributions to the final residual distribution (Figure 2).

Dataset-adapted retrieval geometry. To enable joint training across datasets, each expert uses a dataset-conditioned affine map (DCA) whose parameters are generated by a hypernetwork. For expert $e \in \{1, \dots, M\}$ and a sample history x from dataset d,

$$(W_e(x), b_e(x)) = \mathrm{DCA}_e(x, \eta_d). \tag{4}$$

Each sample is encoded using its own generated map, producing the normalized representation

$$u_e(x) = \frac{W_e(x)x + b_e(x)}{\|W_e(x)x + b_e(x)\|_2}. \tag{5}$$

Thus $u_e(h)$ represents the query, and its inner product with $u_e(H_i)$ measures their cosine similarity. In practice, each history is concatenated with its point forecast; this augmentation and the dataset index on the expert functions are suppressed in the notation.

The dataset descriptor is

$$\eta_d = (d, \mu_d, \sigma_d, \log N_d), \tag{6}$$

where d identifies the dataset, μ_d and σ_d are the componentwise mean and standard deviation of its fitting inputs, and N_d is their number. Section 4.3 shows that the generated affine maps vary little across histories within a dataset but differ substantially across datasets, supporting their role in adapting retrieval to dataset characteristics.

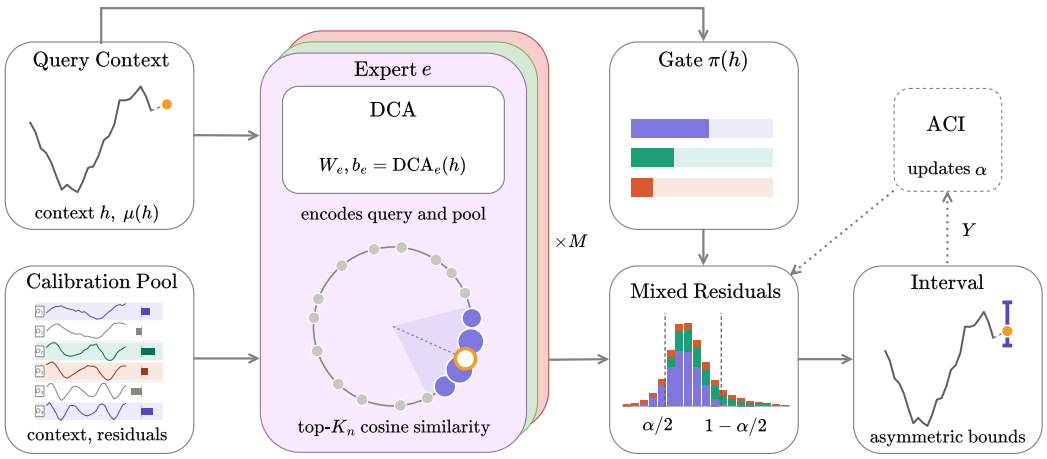


Figure 2: RareCP overview at inference. For a query context h with point forecast $\mu(h)$, each of M experts uses a dataset-conditioned affine map (DCA) to encode the query and calibration histories. Calibration representations are cached. Each expert retrieves the K_n most cosine-similar samples and weights them. A gate $\pi(h)$ mixes the experts' signed residuals, and the $\alpha/2$ and $1-\alpha/2$ quantiles give an asymmetric interval around $\mu(h)$. ACI updates the miscoverage level α online from observed targets Y. Training is described in Section 3.1.

Retrieval and mixture weights. Let $s_{i,e}(h) = \langle u_e(h), u_e(H_i)\rangle$ and $A_{e,n}(h)$ contain the indices of the K_n largest similarities for expert e, breaking ties by index. With retrieval temperature $\tau_e > 0$, the weights of expert e are

$$w_{i,e,n}(h) = \frac{\exp\{s_{i,e}(h)/\tau_e\}\mathbf{1}\{i \in A_{e,n}(h)\}}{\sum_{j\in A_{e,n}(h)} \exp\{s_{j,e}(h)/\tau_e\}}. \tag{7}$$

Each expert selects its own support.

A gate produces $\pi(h) = \mathrm{softmax}(\mathrm{Gate}(h, \eta_d))$. The final calibration weights are

$$w_{i,n}(h) = \sum_{e=1}^{M} \pi_e(h) w_{i,e,n}(h). \tag{8}$$

Weights from experts retrieving the same example are added. We then construct the prediction interval from the quantiles of the combined residual distribution.

Training. Training proceeds in two stages. We first fit DCA experts across datasets with an interval-score objective. We then freeze the experts and train the gate on the mixed residual distribution, with an entropy regularizer that discourages collapse onto a single expert.

We train across datasets using leave-one-out minibatch episodes: each example serves as a query and is excluded from its retrieval candidates Auer et al. (2023). For interval endpoints $\ell_- \le \ell_+$, the Winkler score is

$$\mathrm{Wink}_\alpha(\ell_-, \ell_+; Y) = (\ell_+ - \ell_-) + \frac{2}{\alpha}\big[(\ell_- - Y)_+ + (Y - \ell_+)_+\big], \tag{9}$$

where $x_+ = \max\{x, 0\}$. It balances interval width against penalties for missed targets.

For optimization, we replace hard weighted quantiles with a sigmoid-based approximation in cumulative-probability space and replace the positive-part penalties with smooth softplus functions. The CDF-smoothing temperature τ_{cdf} follows a cyclic schedule: higher temperatures distribute gradients across a broader range of residual weights, while lower temperatures concentrate them near the target quantiles. As both smoothing temperatures vanish, the surrogate approaches the hard Winkler score away from weighted-CDF boundaries. We average the surrogate over a small grid of miscoverage levels around the target. Appendix A defines the surrogate, its two smoothing temperatures, and the expert and gate objectives.

Inference. RareCP caches calibration representations and residuals. For each query, it computes the query representations and gate weights, retrieves the expert neighborhoods, and extracts exact weighted quantiles from at most MK_n distinct residuals. During chronological evaluation, new examples enter the calibration pool only after their targets are observed. We maintain a fixed-capacity pool by removing the oldest entries as new ones arrive, following the online memory updates of HopCPT Auer et al. (2023).

We use ACI Gibbs & Candes (2021) to adjust the miscoverage level from observed coverage errors, targeting long-run empirical coverage under drift. The analysis below separately studies quantile accuracy at a fixed nominal level.

3.2 Theoretical analysis

Our analysis asks what increasing the calibration size can achieve once RareCP has been trained. There are two sources of error: estimating the fitted retrieval distribution from finitely many calibration examples, and a possible mismatch between that fitted distribution and the target residual law. The theorem gives sufficient conditions on K_n under which calibration recovers the optimal history-conditional quantiles up to population learning and approximation error.

Setup and population target. Fix a query history h and a miscoverage level α. All learned functions, dataset descriptors, and hyperparameters are fitted using data independent of the calibration pool and then held fixed. Write

$$U_i = (u_e(H_i))_{e=1}^M, \qquad u = (u_e(h))_{e=1}^M.$$

The sample encoders u_e, including their DCA experts, are fixed, as are the gate weights $\pi_e(h)$ at the query.

As expert e retrieves increasingly local neighborhoods, suppose its weighted population residual distribution approaches a CDF $G_e(\cdot \mid u)$. The corresponding population mixture is

$$G_\pi(r \mid u) = \sum_{e=1}^{M} \pi_e(h) G_e(r \mid u), \qquad Q_{p,\pi}(u) = \inf\{r : G_\pi(r \mid u) \geq p\}. \tag{10}$$

This is the distribution that calibration aims to estimate. It need not equal the target residual CDF $F_H(\cdot \mid h)$.

Let $Q_p(h)$ denote the target residual quantile. For $p_- = \alpha/2$ and $p_+ = 1 - \alpha/2$, define the remaining population learning and approximation error as

$$\varepsilon_{\mathrm{q}}(h) = \max_{p \in \{p_-, p_+\}} |Q_{p,\pi}(u) - Q_p(h)|. \tag{11}$$

Only the quantiles of the gated mixture need to match the target; individual experts may represent different residual distributions.

Sources of error. Following the sampling-and-localization decomposition used in ResCP Neglia et al. (2025), and retaining the fitted mixture's discrepancy from the target law, we obtain

$$\Delta_{n,\mathcal{R}}(h) := \max_{r \in \mathcal{R}} |\widehat{F}_n(r \mid h) - F_H(r \mid h)| \leq R_{n,\mathcal{R}}(u) + B_{n,\mathcal{R}}(u) + D_{\mathcal{R}}(h), \tag{12}$$

where $\mathcal{R}$ is a finite deterministic set of residual thresholds. Here R is sampling error from finite, temporally dependent calibration data, B is localization bias from retrieving a neighborhood, and

$$D_{\mathcal{R}}(h) = \max_{r \in \mathcal{R}} |G_\pi(r \mid u) - F_H(r \mid h)|$$

is the population mismatch of the frozen experts and gate. Under the assumptions and retrieval schedule below, increasing calibration size makes R and B vanish, leaving D as the fitted model's population discrepancy. The resulting quantile error converges to $\varepsilon_{\mathrm{q}}(h)$, establishing asymptotic optimality of calibration up to the population learning and approximation error.

Assumptions. Appendix B states the precise conditions. Their roles are:

- **(A1) Decaying temporal dependence.** The calibration examples are geometrically beta mixing: dependence weakens exponentially with temporal separation.
- **(A2) Controlled retrieval weights.** Within each selected neighborhood, the unnormalized weighting factors are bounded above and away from zero. This prevents a few examples from retaining a disproportionate share of the weight as the neighborhood grows.
- **(A3) Stable local residual laws.** Each expert's weighted population residual distribution approaches $G_e(\cdot \mid u)$ as its retrieval neighborhood shrinks. This condition controls localization bias.
- **(A4) Available local examples.** Every fixed, sufficiently small query neighborhood retains positive average probability as the calibration pool grows, and the population distance distributions are continuous. These conditions allow increasingly local retrieval as more calibration data become available.
- **(A5) Identifiable mixture quantiles.** The target probability levels identify finite, isolated quantiles of G_π. This allows CDF convergence to imply quantile convergence.
- **(A6) Regular target quantiles and finite risk.** The target law satisfies the quantile regularity and integrability conditions needed to connect population Winkler risk to target-quantile error. This assumption is used for the training connection below, rather than for consistency with respect to the fitted mixture.

For RareCP, **(A2)** holds automatically at fixed positive retrieval temperatures because cosine similarities lie in $[-1, 1]$:

$$\exp(-1/\tau_e) \le \exp\{s_{i,e}(h)/\tau_e\} \le \exp(1/\tau_e).$$

These bounds apply to the factors within the selected support; examples outside that support receive zero weight.

Consistency as calibration grows. The retrieval size must balance two goals: using enough examples to reduce sampling error and retaining a sufficiently small neighborhood to reduce localization bias.

Theorem 1 (Consistency for the fitted mixture)**.** *Under the independently fitted setup and **(A1)–(A5)**, let the number of experts M and the positive retrieval temperatures be fixed. Choose a deterministic retrieval schedule satisfying*

$$\frac{K_n}{n} \longrightarrow 0, \qquad \frac{\log(\mathrm{e}n/K_n)}{K_n} \longrightarrow 0, \tag{13}$$

where e *is Euler's number. Then*

$$\max_{p \in \{p_-, p_+\}} |\widehat{Q}_{p,n}(h) - Q_{p,\pi}(u)| \xrightarrow{\mathbb{P}} 0.$$

If the target quantiles are finite, it follows that

$$\max_{p \in \{p_-, p_+\}} |\widehat{Q}_{p,n}(h) - Q_p(h)| \xrightarrow{\mathbb{P}} \varepsilon_{\mathrm{q}}(h).$$

The first schedule condition makes the retrieved fraction vanish, allowing neighborhoods to become local. The second makes the neighborhood grow sufficiently quickly to control sampling error under dependence. For example, $K_n \asymp n^a$ with $0 < a < 1$ satisfies both conditions. The proof appears in Appendix B, Theorem 2.

Localization can be necessary when residual laws vary across neighborhoods: with bounded scores and temperatures bounded away from zero, dense retrieval retains nonvanishing weight on nonlocal examples whenever these form a nonvanishing fraction of the pool, and can therefore remain biased. Top-K_n retrieval with $K_n/n \to 0$ avoids this obstacle while allowing fixed or lower bounded positive temperatures; see Appendix B, Remark 1.

Table 1: Results at 80% target coverage. Cells report nWink and nW / Cov, with mean Winkler score and mean interval width normalized by target std. The first two panels compare CP competitors and the bottom panel reports native TSFM uncertainty estimates alongside RareCP. Values are unweighted means over 22 datasets. All non-deterministic methods were run with 10 seeds, and we report the mean. MoE-CP uses no forecasting backbone. **Bold** marks the lowest nWink among the CP competitors and RareCP for each backbone. Underlining marks the lowest nWink overall for each backbone after including the corresponding TSFM result.

Main Results

Conformal Prediction Competitors (1/2)

	ACI		COP		Decay-OGD		DtACI		ECI		PID	
Backbone	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑
TiRex	0.88	0.47 / 0.79	0.87	0.44 / 0.78	0.90	0.39 / 0.79	0.89	0.41 / 0.80	0.86	0.47 / 0.80	0.87	0.46 / 0.79
TimesFM2.5	0.88	0.44 / 0.80	0.87	0.47 / 0.79	0.90	0.40 / 0.79	0.89	0.42 / 0.80	0.86	0.46 / 0.79	0.88	0.47 / 0.79
Toto2	0.83	0.42 / 0.80	0.83	0.43 / 0.79	0.85	0.39 / 0.79	0.84	0.39 / 0.80	0.82	0.43 / 0.79	0.85	0.45 / 0.80
ARIMA	1.05	0.56 / 0.80	1.04	0.57 / 0.79	1.09	0.52 / 0.79	1.07	0.55 / 0.80	1.04	0.57 / 0.80	1.04	0.57 / 0.79

Conformal Prediction Competitors (2/2)

	SAOCP		SF-OGD		NexCP		KOWCPI		HopCPT		ResCP	
Backbone	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑
TiRex	0.89	0.54 / 0.85	0.90	0.41 / 0.79	0.87	0.43 / 0.79	0.81	0.29 / 0.68	0.79	0.42 / 0.77	0.75	0.36 / 0.75
TimesFM2.5	0.89	0.53 / 0.85	0.90	0.40 / 0.79	0.86	0.43 / 0.79	0.85	0.28 / 0.68	0.84	0.43 / 0.78	0.75	0.35 / 0.75
Toto2	0.87	0.52 / 0.85	0.85	0.39 / 0.80	0.82	0.40 / 0.79	0.79	0.28 / 0.68	0.81	0.41 / 0.78	0.75	0.34 / 0.74
ARIMA	1.08	0.67 / 0.85	1.08	0.53 / 0.79	1.04	0.55 / 0.79	0.95	0.35 / 0.68	1.07	0.62 / 0.79	0.91	0.40 / 0.74
MoE-CP *(no backbone)*	1.02	0.47 / 0.74										

Native TSFM UQ

Model	nWink ↓	nW ↓ / Cov ↑	**Model**	nWink ↓	nW ↓ / Cov ↑	**Model**	nWink ↓	nW ↓ / Cov ↑
TiRex	0.71	0.38 / 0.79	TiRex2	0.70	0.37 / 0.83	Chronos2	0.68	0.40 / 0.78
Moirai2	0.73	0.38 / 0.77	PatchTSTFM	0.72	0.39 / 0.76	TabPFN-TS3	0.86	0.47 / 0.82
TimesFM2.5	0.75	0.40 / 0.84	Toto2	0.64	0.37 / 0.81	TTM	0.88	0.61 / 0.90

RareCP (Ours)

Backbone	nWink ↓	nW ↓ / Cov ↑
TiRex	**0.61**	0.38 / 0.79
TimesFM2.5	**0.71**	0.37 / 0.78
Toto2	**0.70**	0.40 / 0.79
ARIMA	**0.76**	0.46 / 0.80

Connection to training. Under **(A6)**, vanishing excess conditional population Winkler risk implies $\varepsilon_{\mathrm{q}}(h) \to 0$ (Appendix B.4), motivating interval-score training of the complete mixture. Theorem 1 then connects calibration consistency to the Winkler-optimal history-conditional endpoints. Finite-data optimization of the smooth surrogate aims to reduce this endpoint error, but approximation, surrogate, generalization, and optimization errors may remain.

4 Experiments

We evaluate RareCP on one-step forecasting across 22 GIFT-Eval datasets (Aksu et al., 2024).

4.1 Setup

Benchmark and protocol. We construct a fixed chronological evaluation set for each of the 22 GIFT-Eval datasets, using up to 50,000 test forecasts or all available forecasts for shorter datasets. Validation and calibration splits always precede evaluation. All targets are one-step-ahead, and RareCP represents each forecasting context using the most recent 64 observations. Every CP method receives the same calibration examples, test order, and cached point forecasts within each backbone. Dataset statistics and split sizes are reported in Appendix C.2.

Methods. We evaluate CP methods on four point-forecasting backbones: TiRex, TimesFM2.5, Toto2, and AutoARIMA as implemented in StatsForecast (Garca et al., 2022). Table 1 includes eight online and four weighted calibration competitors, described in Section 2. We separately compare RareCP with the native uncertainty estimates of the evaluated TSFMs. This comparison includes matched TiRex, TimesFM2.5, and Toto2 backbones, while ARIMA tests whether RareCP transfers to a forecaster without a native uncertainty head.

Model selection and evaluation. All model selection uses chronologically earlier validation data. For each dataset, we chronologically split the remaining pre-test data into 10% for fitting RareCP and 90% for calibration. RareCP and its dataset descriptors are fitted on the former and then frozen; the latter initializes its calibration pool. Competitor hyperparameters are tuned per dataset over the search spaces in Appendix C.4. Methods process test targets chronologically and update their calibration state according to their algorithms. We average non-deterministic methods over ten seeds.

4.2 RESULTS

Evaluation criteria. Table 1 reports unweighted means over the 22 datasets at 80% target coverage. We use normalized Winkler score (*nWink*) as the primary metric because it penalizes both wide intervals and coverage failures. Normalized width (*nW*) and empirical coverage (*Cov*) show how each method achieves its score. Both metrics are normalized by the target standard deviation computed on the calibration split, with the denominator clipped below at a small constant to ensure stability for nearly constant targets. Full dataset level results are provided in Appendix C.3.

Conformal prediction. RareCP attains the lowest nWink among CP methods on all four backbones. Relative to ResCP, the strongest competitor, it reduces nWink by 19% on TiRex, 5% on TimesFM2.5, 7% on Toto2, and 16% on ARIMA. ResCP produces slightly narrower intervals but undercovers (0.74 to 0.75), while RareCP stays close to the 0.80 target (0.78 to 0.80). KOWCPI shows the same effect more strongly. Its narrow intervals reach only 0.68 coverage, so width alone does not measure efficiency.

Native TSFM uncertainty. On the TiRex backbone, RareCP reaches an nWink of 0.61, the lowest in Table 1. This is 5% below Toto2 (0.64), the strongest native model. On matched backbones, RareCP improves on native TiRex by 14% and on native TimesFM2.5 by 5%, but trails native Toto2 (0.70 vs. 0.64). With ARIMA, a forecaster without a native uncertainty head, RareCP reaches 0.76 and still outperforms the native uncertainty of TabPFN-TS3 and TTM.

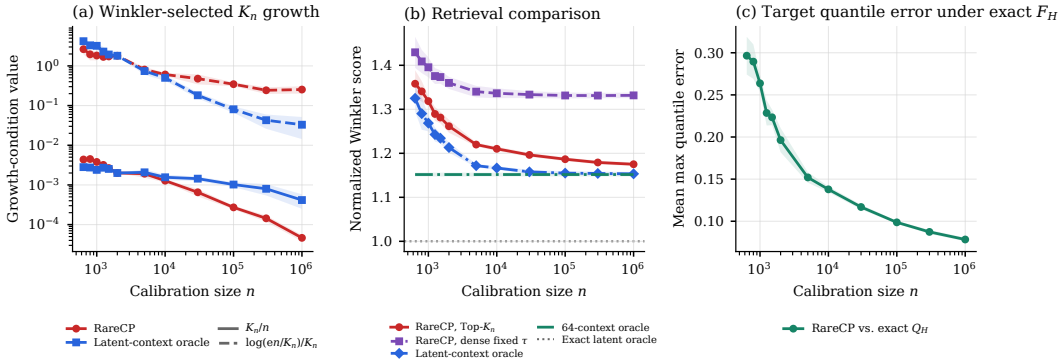


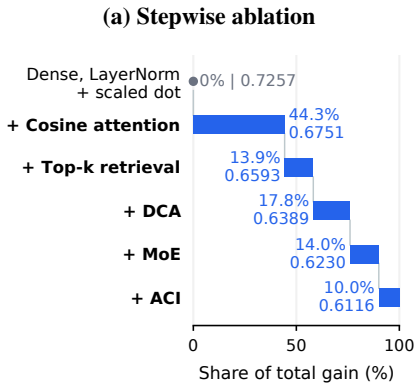
Figure 3: Controlled retrieval scaling on a stochastic seasonal process. **(a)** Growth conditions K_n/n (solid) and $\log(en/K_n)/K_n$ (dashed) for validation-selected K_n. **(b)** Test Winkler score relative to the exact latent oracle (dotted line). **(c)** RareCP's mean maximum endpoint error relative to the exact history-conditional quantiles. Means and 95% confidence intervals over ten seeds.

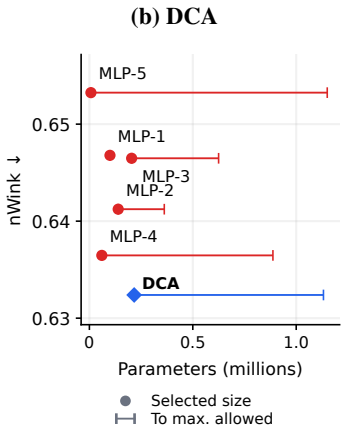
4.3 MECHANISM AND ABLATION STUDIES

Illustration of retrieval scaling. We select a stochastic seasonal process with phase-dependent volatility to illustrate the retrieval behavior described by our theory. With RareCP trained independently and frozen, the validation-selected K_n grows with calibration size while K_n/n and $\log(en/K_n)/K_n$ generally decrease (Figure 3(a)). These trends agree with the sufficient retrieval conditions. Localization is not universally necessary. For i.i.d. Gaussian noise with a zero predictor, uniform retrieval with $K_n = n$ is optimal and uses all available residuals without localization bias.

As calibration grows, sparse retrieval improves interval quality while dense fixed-temperature retrieval plateaus (panel (b)). The target-quantile error also decreases (panel (c)), consistent with convergence toward the remaining learning and approximation error $\varepsilon_{\mathrm{q}}(h)$. The experiment uses no ACI or rolling calibration updates. Details appear in Appendix C.1.

Component ablations. Starting from a learned dense dot-product similarity, we add components cumulatively using the TiRex backbone (Figure 4(a)). Cosine normalization yields the largest gain. The dataset-conditioned affine map (DCA), sparse top-K retrieval, and residual experts add comparable gains, while ACI corrects the remaining coverage drift. The DCA outperforms MLP encoders





(c) MoE

Experts	nWink ↓	h
1	0.6336 ± 0.0059	0.19
3	0.6165 ± 0.0002	0.61
10	0.6116 ± 0.0010	1.75
30	**0.6105** ± 0.0013	5.72

Mean ± sample SD.
3 seeds; 10 seeds for 10 experts.

Figure 4: Ablations on the TiRex backbone. **(a)** Cumulative component gains over a dense scaled dot-product baseline. **(b)** Encoder size versus nWink for DCA (blue) and MLPs (red; labels indicate layer depth). Whiskers extend to the maximum parameter budget. **(c)** Nested expert sweep before the final ACI correction, with mean wall-clock runtime per seed in hours.

with equal or larger parameter budgets (Figure 4(b)). Its low-capacity affine maps adapt retrieval at the dataset level without enlarging a shared embedding. Each expert's generated maps have an average cosine similarity of 0.99 to their dataset mean, versus only 0.38 between the means of different datasets. The expert mixture also improves performance, while Figure 1 illustrates how the gate varies expert contributions across contexts. Ten experts recover 95% of the performance gain achieved by 30 experts at under a third of the runtime (Figure 4(c)).

5 Conclusion

We show that conformal prediction can be competitive with native TSFM uncertainty in one-step-ahead forecasting. RareCP uses a single calibration model trained jointly across datasets, combining dataset-conditioned affine maps with sparse residual retrieval and expert mixing. Across 22 GIFT-Eval datasets and four forecasting backbones, it achieves the best aggregate interval efficiency among the evaluated CP methods and competitive results against native TSFM intervals. The retrieved histories and their weighted residuals make interval construction inspectable, allowing users to relate changes in uncertainty to the historical examples supporting a prediction. For example, a wider electricity-demand interval can be investigated by checking whether the retrieved histories include periods affected by outages or unusual demand fluctuations. Our theory establishes sufficient retrieval conditions under which calibration estimation error vanishes, leaving learning and approximation error. ACI separately targets long-run empirical coverage under distributional drift.

5.1 Limitations

RareCP requires offline training and retrieval from a calibration pool, making it more computationally demanding than simple online CP methods. Our evaluation is limited to one-step-ahead intervals and does not establish performance for longer horizons or multivariate targets. Finally, common targets, metrics, and matched backbones do not fully control the comparison with native TSFM uncertainty. CP methods use calibration residuals and online feedback, while native uncertainty heads differ in pretraining, capacity, and training objectives.

5.2 Future Work

Future work could extend RareCP to multistep and multivariate forecasting and reduce computational cost through approximate nearest-neighbor search and simpler dataset-conditioned maps. Adaptive retrieval sizes and explicit drift detection could improve robustness to rare residual regimes and abrupt changes in the residual distribution.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, propose wording for parts of the manuscript, support literature retrieval and discovery, and assist in developing and checking mathematical claims. We did not use these tools for research ideation or execution, to generate synthetic datasets, or for any other purpose.

We reviewed and revised all AI-assisted writing and checked suggested sources against the original publications and verified that they support the associated claims. We treated mathematical outputs only as assistance, reviewed the arguments, and verified the resulting claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The RareCP architecture, training procedure, and inference algorithm are specified in Section 3, with the training objectives detailed in Appendix A. The assumptions underlying our theoretical results are stated in Section 3.2, and complete proofs are provided in Appendix B. The experimental protocol, chronological data splits, model selection procedure, and evaluation criteria are described in Section 4. Additional dataset statistics, experimental details, full results, hyperparameter settings, and computational costs are reported in Appendix C. The anonymous supplementary material includes the code and execution instructions needed to reproduce the RareCP and baseline results, together with a virtual environment specification. The code will be released through a public repository upon acceptance.

REFERENCES

Sami Achour, Yassine Bouher, Duong Nguyen, and Nicolas Chesneau. Foundation models for time series forecasting: Application in conformal prediction, 2025. URL https://arxiv.org/abs/2507.08858.

Coen Adler, Yuxin Chang, Felix Draxler, Samar Abdi, and Padhraic Smyth. Beyond accuracy: Are time series foundation models well-calibrated?, 2026. URL https://arxiv.org/abs/2510.16060.

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation, 2024. URL https://arxiv.org/abs/2410.10393.

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal pid control for time series prediction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 23047–23074. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/47f2fad8c1111d07f83c91be7870f8db-Paper-Conference.pdf.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Online conformal prediction with decaying step sizes, 2024. URL https://arxiv.org/abs/2402.01139.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024. URL https://arxiv.org/abs/2403.07815.

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

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks, 2023. URL https://arxiv.org/abs/2303.12783.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning, 2025. URL https://arxiv.org/abs/2505.23719.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference, 2020. URL https://arxiv.org/abs/1903.04684.

Rina Foygel Barber, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability, 2023. URL https://arxiv.org/abs/2202.13415.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning, 2023. URL https://arxiv.org/abs/2302.07869.

George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, CA, 1970.

Jason B. Cho and David S. Matteson. Improving conditional coverage in time-series foundation models via direct volatility modeling. In *1st ICLR Workshop on Time Series in the Age of Large Models*, 2026. URL https://openreview.net/forum?id=dZx0R4Cm8d.

Ben Cohen, Emaad Khwaja, Kan Wang, Charles Masson, Elise Ramé, Youssef Doubli, and Othmane Abou-Amal. Toto: Time series optimized transformer for observability, 2024. URL https://arxiv.org/abs/2407.07874.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 74147–74181. Curran Associates, Inc., 2024. doi: 10.52202/079017-2359. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/874a4d89f2d04b4bcf9a2c19545cf040-Paper-Conference.pdf.

Azul Garca, Max Mergenthaler Canseco, Challú Cristian, and Kin G. Olivares. StatsForecast: Lightning fast forecasting with statistical and econometric models. PyCon Salt Lake City, Utah, US 2022, 2022. URL https://github.com/Nixtla/statsforecast.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1, 2024. URL https://arxiv.org/abs/2310.03589.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf.

Isaac Gibbs and Emmanuel Candès. Conformal inference for online prediction with arbitrary distribution shifts. *J. Mach. Learn. Res.*, 25(1), January 2024. ISSN 1532-4435.

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

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3: Technical report, 2026. URL https://arxiv.org/abs/2605.13986.

Leying Guan. Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 03 2023. ISSN 1464-3510. doi: 10.1093/biomet/asac040. URL https://doi.org/10.1093/biomet/asac040.

David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. *CoRR*, abs/1609.09106, 2016. URL http://arxiv.org/abs/1609.09106.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From Tables to Time: Extending TabPFN-v2 to Time Series Forecasting. *arXiv e-prints*, art. arXiv:2501.02945, January 2025. doi: 10.48550/arXiv.2501.02945.

Dongjian Hu, Junxi Wu, Shu-Tao Xia, and Changliang Zou. Distribution-informed online conformal prediction. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 58230–58255, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5f8241131e6fd428aa49914da76b8ad0-Paper-Conference.pdf.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991. URL https://api.semanticscholar.org/CorpusID:572361.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, Chenghao Liu, Ameet Talwalkar, and David Asker. Toto 2.0: Time series forecasting enters the scaling era, 2026. URL https://arxiv.org/abs/2605.20119.

Jingsen Kong, Wenlu Tang, Dezheng Kong, Linglong Kong, Guangren Yang, and Bei Jiang. Adaptive conformal prediction via mixture-of-experts gating similarity. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vCmnu4q8C3.

Jonghyeok Lee, Chen Xu, and Yao Xie. Kernel-based optimally weighted conformal time-series prediction, 2025. URL https://arxiv.org/abs/2405.16828.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://doi.org/10.1080/01621459.2017.1307116.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting, 2026. URL https://arxiv.org/abs/2511.11698.

Roberto Neglia, Andrea Cini, Michael M. Bronstein, and Filippo Maria Bianchi. ResCP: Reservoir Conformal Prediction for Time Series Forecasting. *arXiv e-prints*, art. arXiv:2510.05060, October 2025. doi: 10.48550/arXiv.2510.05060.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex-2: Generalizing tirex to multivariate data and streaming, 2026. URL https://arxiv.org/abs/2607.01204.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks, 2021. URL https://arxiv.org/abs/2103.04628.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. 01 2005. doi: 10.1007/b106715.

Yunshi Wen, Wesley M. Gifford, Chandra Reddy, Lam M. Nguyen, Jayant Kalagnanam, and Anak Agung Julius. Revisiting the generic transformer: Deconstructing a strong baseline for time series foundation models, 2026. URL https://arxiv.org/abs/2602.06909.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

Junxi Wu, Dongjian Hu, Yajie Bao, Shu-Tao Xia, and Changliang Zou. Error-quantified conformal inference for time series. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 81566–81593, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/cab5ae2704d3e01f06a92512a5376b87-Paper-Conference.pdf.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11559–11569. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/xu21h.html.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series, 2023. URL https://arxiv.org/abs/2212.03463.

A SMOOTH WINKLER SURROGATE AND TRAINING LOSSES

RareCP learns retrieval weights by scoring intervals on held-out training queries. Exact weighted quantiles change discontinuously when cumulative weights cross a target probability. We therefore smooth quantile selection during training and also smooth the Winkler score's positive-part penalties. Evaluation uses the exact weighted quantiles in equation 2 and the exact score in equation 9.

A.1 A SMOOTH WEIGHTED QUANTILE

Consider one training query with a weighted residual distribution

$$\mathsf{P} = \sum_{i=1}^{s} w_i \delta_{E_i}, \qquad w_i > 0, \qquad \sum_{i=1}^{s} w_i = 1,$$

where δ_{E_i} is a point mass at residual E_i. Zero-weight entries are omitted. This distribution may come from one retrieval expert or from the gated mixture.

Sort the residuals as $E_{(1)} \le \cdots \le E_{(s)}$, carrying their weights along, and define

$$C_0 = 0, \qquad C_i = \sum_{k=1}^{i} w_{(k)}, \qquad i = 1, \ldots, s. \tag{14}$$

Residual $E_{(i)}$ occupies the probability interval $(C_{i-1}, C_i]$. The exact p-quantile selects the first residual whose cumulative weight reaches p:

$$Q_p(\mathsf{P}) = E_{(k)}, \qquad k = \min\{i : C_i \ge p\}.$$

Smoothing selection in probability space. Let $\sigma(x) = (1 + \exp(-x))^{-1}$. For a CDF-smoothing temperature $\tau_{\mathrm{cdf}} > 0$, define

$$b_i(p; \tau_{\mathrm{cdf}}) = \sigma\left(\frac{p - C_{i-1}}{\tau_{\mathrm{cdf}}}\right) - \sigma\left(\frac{p - C_i}{\tau_{\mathrm{cdf}}}\right), \tag{15}$$

$$\gamma_i(p; \tau_{\mathrm{cdf}}) = \frac{b_i(p; \tau_{\mathrm{cdf}})}{\sum_{k=1}^{s} b_k(p; \tau_{\mathrm{cdf}})}. \tag{16}$$

The sigmoid difference is a smooth indicator that p falls inside the probability interval assigned to residual $E_{(i)}$. Since $C_{i-1} \le C_i$, these differences are nonnegative, and their sum is

$$\sum_{i=1}^{s} b_i(p; \tau_{\mathrm{cdf}}) = \sigma\left(\frac{p}{\tau_{\mathrm{cdf}}}\right) - \sigma\left(\frac{p-1}{\tau_{\mathrm{cdf}}}\right) > 0.$$

The normalized coefficients therefore define the smooth quantile

$$\widetilde{Q}_p(\mathsf{P}; \tau_{\mathrm{cdf}}) = \sum_{i=1}^{s} \gamma_i(p; \tau_{\mathrm{cdf}}) E_{(i)}. \tag{17}$$

At higher temperatures, several residuals contribute around the target probability, distributing gradients across their weights. At lower temperatures, the contribution concentrates near the selected quantile. The smoothing acts on cumulative probabilities, not on distances between residual values. It smooths quantile extraction on the selected support; it does not smooth the discrete top-K support selection.

Zero-temperature behavior. For a fixed weighted support, if $C_{k-1} < p < C_k$, then

$$\widetilde{Q}_p(\mathsf{P}; \tau_{\mathrm{cdf}}) \longrightarrow E_{(k)} = Q_p(\mathsf{P}) \qquad \text{as } \tau_{\mathrm{cdf}} \downarrow 0.$$

At an exact internal boundary $p = C_k$, with $k < s$, the symmetric sigmoid construction instead satisfies

$$\widetilde{Q}_p(\mathsf{P}; \tau_{\mathrm{cdf}}) \longrightarrow \frac{E_{(k)} + E_{(k+1)}}{2}.$$

Thus convergence to the chosen hard-quantile convention holds away from exact cumulative-weight boundaries. This fixed-support limit alone does not establish a uniform approximation when the learned weights or calibration size change.

A.2 THE SMOOTH INTERVAL SCORE

Let $E_j = Y_j - \mu(H_j)$ be the residual of the held-out training query. For a miscoverage level α, put

$$p_- = \alpha/2, \qquad p_+ = 1 - \alpha/2, \qquad \widetilde{q}_- = \widetilde{Q}_{p_-}(\mathsf{P}; \tau_{\mathrm{cdf}}), \qquad \widetilde{q}_+ = \widetilde{Q}_{p_+}(\mathsf{P}; \tau_{\mathrm{cdf}}).$$

We smooth the positive-part function with a separate temperature $\tau_{\mathrm{pen}} > 0$:

$$\mathrm{sp}_{\tau_{\mathrm{pen}}}(x) = \tau_{\mathrm{pen}} \log(1 + \exp(x/\tau_{\mathrm{pen}})). \tag{18}$$

It satisfies

$$0 \le \mathrm{sp}_{\tau_{\mathrm{pen}}}(x) - x_+ \le \tau_{\mathrm{pen}} \log 2,$$

so it approaches the positive-part penalty uniformly as $\tau_{\mathrm{pen}} \downarrow 0$.

The smooth residual-space Winkler score is

$$\begin{aligned} \widetilde{\mathrm{Wink}}_\alpha(\mathsf{P}; E_j, \tau_{\mathrm{cdf}}, \tau_{\mathrm{pen}}) = (\widetilde{q}_+ - \widetilde{q}_-) \\ + \frac{2}{\alpha}\left[\mathrm{sp}_{\tau_{\mathrm{pen}}}(\widetilde{q}_- - E_j) + \mathrm{sp}_{\tau_{\mathrm{pen}}}(E_j - \widetilde{q}_+)\right]. \end{aligned} \tag{19}$$

The first term penalizes width. The two remaining terms penalize a target below the lower endpoint or above the upper endpoint.

Because $E_j = Y_j - \mu(H_j)$, this is equivalent to scoring the prediction interval

$$\left[\mu(H_j) + \widetilde{q}_-,\ \mu(H_j) + \widetilde{q}_+\right]$$

against Y_j: the point forecast cancels from the penalty arguments and from the interval width.

The temperatures have different roles. τ_{cdf} controls how sharply the residual quantiles are selected, whereas τ_{pen} controls the smoothness of the missed-target penalties. Both are distinct from the retrieval temperatures τ_e in equation 7. For a fixed support whose target probabilities are not cumulative boundaries, taking both smoothing temperatures to zero recovers the exact weighted-quantile Winkler score.

Training across interval levels. Let $\mathcal{G}_\alpha \subset (0, 1)$ be a finite grid centered around the target miscoverage level α_{tar}. We train with

$$\mathcal{W}(\mathsf{P}; E_j) = \frac{1}{|\mathcal{G}_\alpha|} \sum_{\alpha \in \mathcal{G}_\alpha} \widetilde{\mathrm{Wink}}_\alpha(\mathsf{P}; E_j, \tau_{\mathrm{cdf}}, \tau_{\mathrm{pen}}). \tag{20}$$

Temperature arguments are suppressed in $\mathcal{W}$ for readability. In the experiments, τ_{cdf} follows a cyclic schedule with alternating higher- and lower-temperature phases. This schedule is an optimization choice; it is not an assumption that the smoothing temperature converges to zero.

A.3 EXPERT AND GATE OBJECTIVES

Leave-one-out retrieval episodes. Let $\mathfrak{D}$ be the collection of fitting datasets and $\mathcal{J}_d$ the training-query indices for dataset d. For query $j \in \mathcal{J}_d$, its own example is excluded from the retrieval candidates before top-K selection.

Expert e generates

$$(W_{e,d,j}, b_{e,d,j}) = \mathrm{DCA}_e(H_j, \eta_d). \tag{21}$$

As in the main text, the map is applied to both the query and all candidate inputs, with point-forecast augmentation suppressed.

Write $A^{(-j)}_{e,d,j}$ for its selected support and $w^{(-j)}_{i,e,d,j}$ for the normalized weights obtained from top-K_n selection and equation 7. The expert's training distribution is

$$\mathsf{P}^{(-j)}_{e,d,j} = \sum_{i \in A^{(-j)}_{e,d,j}} w^{(-j)}_{i,e,d,j} \delta_{E_i}.$$

The superscript $(-j)$ records that the query is excluded from its own retrieval support. It does not imply that the fitting data are independent of the final calibration pool.

For compactness, define the dataset-balanced average

$$\langle f_{d,j}\rangle_{\mathrm{fit}} = \frac{1}{|\mathfrak{D}|}\sum_{d\in\mathfrak{D}}\frac{1}{|\mathcal{J}_d|}\sum_{j\in\mathcal{J}_d} f_{d,j}.$$

Thus every dataset receives equal total weight.

Teacher anchoring. Let the prefit teacher map corresponding to expert e, dataset d, and query j be

$$T_{e,d,j}(x) = W^{\mathrm{T}}_{e,d,j}x + b^{\mathrm{T}}_{e,d,j}. \tag{22}$$

A teacher shared across a dataset is represented by dropping its j dependence. The parameter-distance penalty is

$$\mathcal{A}_{e,d,j} = \|W_{e,d,j} - W^{\mathrm{T}}_{e,d,j}\|_F^2 + \|b_{e,d,j} - b^{\mathrm{T}}_{e,d,j}\|_2^2.$$

Writing θ_e for the expert hypernetwork parameters, its objective is

$$\mathcal{L}^{(e)}_{\mathrm{enc}}(\theta_e) = \left\langle \mathcal{W}(\mathsf{P}^{(-j)}_{e,d,j}; E_j) + \lambda_{\mathrm{anch}}\mathcal{A}_{e,d,j}\right\rangle_{\mathrm{fit}}. \tag{23}$$

The score trains intervals on held-out queries, while the anchor regularizes the generated retrieval maps toward their teachers.

Training the mixture gate. After freezing the experts, we train a gate with parameters ψ:

$$\pi_{d,j} = \mathrm{softmax}\big(\mathrm{Gate}_\psi(H_j, \eta_d)\big).$$

Extend each expert's weights by zero outside its support. The mixture weights and residual distribution are

$$w^{(-j)}_{i,d,j} = \sum_{e=1}^{M}\pi_{e,d,j}w^{(-j)}_{i,e,d,j}, \qquad \mathsf{P}^{(-j)}_{\pi,d,j} = \sum_{i\in\cup_e A^{(-j)}_{e,d,j}} w^{(-j)}_{i,d,j}\delta_{E_i}.$$

If several experts retrieve the same example, their contributions are added before quantile extraction.

Define the gate entropy

$$\mathcal{H}(\pi_{d,j}) = -\sum_{e=1}^{M}\pi_{e,d,j}\log\pi_{e,d,j}.$$

The gate objective is

$$\mathcal{L}_{\mathrm{gate}}(\psi) = \left\langle \mathcal{W}(\mathsf{P}^{(-j)}_{\pi,d,j}; E_j) - \lambda_{\mathrm{ent}}\mathcal{H}(\pi_{d,j})\right\rangle_{\mathrm{fit}}. \tag{24}$$

The interval-score term evaluates the complete mixed residual distribution. The entropy term encourages the gate to retain multiple experts.

Connection to the population analysis. Appendix B.4 relates the exact population Winkler risk at the fitted mixture quantiles to $\varepsilon_{\mathrm{q}}(h)$. The training objective above uses finite leave-one-out supports, smoothed quantiles, and regularization. Transferring optimization of this objective to the population result requires additional control of surrogate approximation, generalization, and optimization error. Theorem 2 therefore retains $\varepsilon_{\mathrm{q}}(h)$ explicitly rather than assuming that training eliminates it.

B THEORETICAL DETAILS

B.1 SETUP AND ERROR DECOMPOSITION

Let $\mathcal{D}_n = \{(H_i, Y_i)\}_{i=1}^n$ be the calibration data, where $H_i \in \mathbb{R}^{d_H}$ contains the information available for predicting the associated target Y_i. The index i labels a forecast example; the calibration pool may contain examples from several time series or datasets. Write

$$E_i = Y_i - \mu(H_i), \qquad U_i = \phi(H_i) \in \mathbb{R}^{d_U},$$

for the residual and representation. The predictor μ, encoder ϕ, expert functions, gate, and their hyperparameters are fixed using training or validation data independent of the calibration data. We condition on these independent training and validation data throughout, suppressing this conditioning from the notation.

Fix a query history h and let $u = \phi(h)$. For a target-population history–residual pair (H, E), define

$$F_H(r \mid h) = \mathbb{P}(E \le r \mid H = h).$$

For RareCP, $\phi(H) = (u_e(H))_{e=1}^M$ collects the expert representations. Each history is encoded using the affine maps generated from that history and its own fixed dataset descriptor. Throughout this analysis, the query history h is fixed. We index the empirical residual CDF and its quantiles by $u = \phi(h)$, writing $\widehat{F}_n(r \mid u)$ and $\widehat{Q}_{p,n}(u)$ for the quantities denoted by $\widehat{F}_n(r \mid h)$ and $\widehat{Q}_{p,n}(h)$ in the main text. Likewise, $\pi_e(u)$ denotes the gate weight $\pi_e(h)$ at this fixed query; this notation does not require the gate to be a function of the retrieval representation alone.

Let $\mathcal{S}$ be a fixed closed set containing u and the supports of all calibration and target representations. For $\alpha \in (0, 1)$, put $p_- = \alpha/2$, $p_+ = 1 - \alpha/2$, and

$$Q_p(h) = \inf\{r : F_H(r \mid h) \ge p\}.$$

Fix a compact interval $\mathcal{I} \subset \mathbb{R}$, required below to contain the relevant population quantiles in its interior.

There are a fixed number M of experts. Expert e uses a measurable nonnegative retrieval distance d_e and a positive measurable weighting function a_e. For a deterministic sequence $K_n \in \{1, \dots, n\}$, let $A_{e,n}(u)$ contain the indices of the K_n smallest distances $d_e(u, U_i)$, breaking ties by index. Define

$$w_{i,e,n}(u) = \frac{a_e(u, U_i)\mathbf{1}\{i \in A_{e,n}(u)\}}{\sum_{j \in A_{e,n}(u)} a_e(u, U_j)}, \qquad w_{i,n}(u) = \sum_{e=1}^M \pi_e(u) w_{i,e,n}(u),$$

where $\pi_e(u) \ge 0$ and $\sum_e \pi_e(u) = 1$. The consolidated weights account for indices retrieved by several experts. Although the retrieval functions are frozen, the weights remain random through the calibration representations.

The expert CDFs and their mixture are

$$\widehat{F}_{e,n}(r \mid u) = \sum_{i=1}^n w_{i,e,n}(u)\mathbf{1}\{E_i \le r\}, \qquad \widehat{F}_n(r \mid u) = \sum_{e=1}^M \pi_e(u)\widehat{F}_{e,n}(r \mid u).$$

Their mixture quantiles and prediction interval are

$$\widehat{Q}_{p,n}(u) = \inf\{r : \widehat{F}_n(r \mid u) \ge p\},$$
$$\widehat{C}_{n,\alpha}(h) = \big[\mu(h) + \widehat{Q}_{p_-,n}(u),\ \mu(h) + \widehat{Q}_{p_+,n}(u)\big].$$

To define the population distribution estimated by each expert, put

$$B_e(u, b) = \{v : d_e(u, v) \le b\}, \qquad p_{e,n}(b) = \frac{1}{n}\sum_{i=1}^n \mathbb{P}\{U_i \in B_e(u, b)\}.$$

Whenever $p_{e,n}(b) > 0$, define the weighted population neighborhood CDF

$$\overline{F}_{e,n,b}(r \mid u) = \frac{\sum_{i=1}^n \mathbb{E}[a_e(u, U_i)\mathbf{1}\{U_i \in B_e(u, b)\}\mathbf{1}\{E_i \le r\}]}{\sum_{i=1}^n \mathbb{E}[a_e(u, U_i)\mathbf{1}\{U_i \in B_e(u, b)\}]}. \tag{25}$$

This is a ratio of population expectations. Assumption **(A3)** requires these local distributions to approach an expert-specific CDF $G_e(\cdot \mid u)$ as the neighborhood shrinks. The population mixture and its quantiles are

$$G_\pi(r \mid u) = \sum_{e=1}^{M} \pi_e(u) G_e(r \mid u), \qquad Q_{p,\pi}(u) = \inf\{r : G_\pi(r \mid u) \geq p\}.$$

Individual G_e may differ from the target law $F_H(\cdot \mid h)$. The population learning and approximation error at the selected quantiles is

$$\varepsilon_{\mathrm{q}}(h) = \max_{p \in \{p_-, p_+\}} |Q_{p,\pi}(u) - Q_p(h)|. \tag{26}$$

It depends on the fitted experts and gate and is retained explicitly.

Let $b_0 > 0$ be the radius in the assumptions below. Whenever $K_n/n \leq \min_e p_{e,n}(b_0)$, choose the deterministic radii

$$b_{e,n} = \inf\{b \in [0, b_0] : p_{e,n}(b) \geq K_n/n\}.$$

Under **(A4)**, $p_{e,n}(b_{e,n}) = K_n/n$. These radii exist for all sufficiently large n under the retrieval schedule used below. Abbreviate

$$\overline{F}_{e,n}(r \mid u) = \overline{F}_{e,n,b_{e,n}}(r \mid u), \qquad \overline{F}_n(r \mid u) = \sum_{e=1}^{M} \pi_e(u) \overline{F}_{e,n}(r \mid u).$$

For a nonempty finite deterministic set $\mathcal{R} \subset \mathcal{I}$, define

$$\begin{aligned}
\Delta_{n,\mathcal{R}}(h) &= \max_{r \in \mathcal{R}} |\widehat{F}_n(r \mid u) - F_H(r \mid h)|, \\
R_{n,\mathcal{R}}(u) &= \max_{r \in \mathcal{R}} |\widehat{F}_n(r \mid u) - \overline{F}_n(r \mid u)|, \\
B_{n,\mathcal{R}}(u) &= \max_{r \in \mathcal{R}} |\overline{F}_n(r \mid u) - G_\pi(r \mid u)|, \\
D_{\mathcal{R}}(h) &= \max_{r \in \mathcal{R}} |G_\pi(r \mid u) - F_H(r \mid h)|.
\end{aligned}$$

Adding and subtracting $\overline{F}_n(r \mid u)$ and $G_\pi(r \mid u)$ gives

$$\boxed{\Delta_{n,\mathcal{R}}(h) \leq R_{n,\mathcal{R}}(u) + B_{n,\mathcal{R}}(u) + D_{\mathcal{R}}(h).} \tag{27}$$

Here R is the sampling fluctuation, B is the localization bias relative to the population mixture, and D is its discrepancy from the target CDF. Temporal dependence, including context overlap, enters R. Expert biases are combined within G_π before measuring D. Both $D_{\mathcal{R}}(h)$ and $\varepsilon_{\mathrm{q}}(h)$ are independent of the calibration size for the fixed fitted model.

B.2 Assumptions

The calibration pool is placed in a fixed order for defining dependence, for example by source and then by time. Its joint distribution may change with n; an additional n subscript on individual observations is suppressed. Stationarity of the pooled data is not required.

(A1) Geometric beta mixing of the calibration data. Write $V_i = (H_i, E_i)$. For sigma-fields $\mathcal{A}, \mathcal{B}$, define

$$\beta(\mathcal{A}, \mathcal{B}) = \frac{1}{2} \sup \sum_{j,\ell} |\mathbb{P}(A_j \cap B_\ell) - \mathbb{P}(A_j)\mathbb{P}(B_\ell)|,$$

where the supremum is over finite measurable partitions (A_j) and (B_ℓ) belonging to $\mathcal{A}$ and $\mathcal{B}$, respectively. For $1 \leq k < n$, set

$$\beta_n(k) = \max_{1 \leq j \leq n-k} \beta\big(\sigma(V_1, \ldots, V_j), \sigma(V_{j+k}, \ldots, V_n)\big),$$

and set $\beta_n(k) = 0$ for $k \geq n$. Assume that, uniformly in n,

$$\beta_n(k) \leq C_\beta e^{-c_\beta k}, \qquad k \geq 1,$$

for finite constants $C_\beta, c_\beta > 0$. This condition is evaluated after conditioning on the independent training and validation data. It is satisfied, for example, by concatenating independent source samples whose within-source processes obey a common geometric beta-mixing bound. Their marginal distributions may differ.

(A2) Bounded weighting factors. All frozen functions are measurable. There are constants $0 < a_- \le a_+ < \infty$ such that

$$a_- \le a_e(u, v) \le a_+, \qquad e = 1, \dots, M, \quad v \in \mathcal{S}.$$

Write $\kappa = a_+/a_-$. Bounded similarity scores s_e and fixed temperatures $\tau_e > 0$ satisfy this condition through $a_e(u, v) = \exp\{s_e(u, v)/\tau_e\}$. These retrieval temperatures are distinct from any temperature used to smooth the training loss.

(A3) Stable local expert distributions. There are CDFs $G_e(\cdot \mid u)$ and a nondecreasing function $\omega : [0, b_0] \to [0, 1]$, with $\omega(b) \to 0$ as $b \downarrow 0$, such that, for every expert, every n, and every $0 < b \le b_0$ with $p_{e,n}(b) > 0$,

$$\sup_{r \in \mathcal{I}} |\overline{F}_{e,n,b}(r \mid u) - G_e(r \mid u)| \le \omega(b).$$

Thus each expert's local pooled residual distribution stabilizes as its neighborhood shrinks. This condition allows different expert limits and differing source distributions; it imposes no equality between an individual G_e and $F_H(\cdot \mid h)$.

(A4) Local recurrence and continuous distances. Each $p_{e,n}$ is continuous on $[0, b_0]$ and $p_{e,n}(0) = 0$. For every expert and every fixed $0 < b \le b_0$, assume

$$\liminf_{n \to \infty} p_{e,n}(b) > 0.$$

(A5) Identification of the mixture quantiles. For $p \in \{p_-, p_+\}$, $Q_{p,\pi}(u)$ is finite and lies in the interior of $\mathcal{I}$, and

$$G_\pi(Q_{p,\pi}(u) - a \mid u) < p < G_\pi(Q_{p,\pi}(u) + a \mid u)$$

for every sufficiently small $a > 0$. This ensures that small CDF estimation errors cannot move the estimated quantiles across a plateau at either selected probability level.

(A6) Target quantile regularity and integrability. The target residual has $\mathbb{E}[|E| \mid H = h] < \infty$. For $p \in \{p_-, p_+\}$, $Q_p(h)$ is finite and lies in the interior of $\mathcal{I}$. The CDF $F_H(\cdot \mid h)$ is continuous at $Q_p(h)$, and

$$F_H(Q_p(h) - a \mid h) < p < F_H(Q_p(h) + a \mid h)$$

for every sufficiently small $a > 0$.

Assumptions **(A1)**–**(A5)** concern estimation of the fitted population mixture. Assumption **(A6)** identifies the target endpoints and makes their expected Winkler scores finite. No assumption requires the population learning and approximation error to vanish.

B.3 Quantile consistency

Theorem 2 (Consistency for the fitted mixture at a fixed query)**.** *Assume **(A1)–(A5)**. Choose a deterministic retrieval schedule satisfying*

$$\frac{K_n}{n} \longrightarrow 0, \qquad \frac{\log(en/K_n)}{K_n} \longrightarrow 0. \tag{28}$$

Then, for every nonempty finite $\mathcal{R} \subset \mathcal{I}$,

$$\max_{r \in \mathcal{R}} |\widehat{F}_n(r \mid u) - G_\pi(r \mid u)| \xrightarrow{\mathbb{P}} 0, \qquad \Delta_{n,\mathcal{R}}(h) \xrightarrow{\mathbb{P}} D_{\mathcal{R}}(h).$$

Moreover,

$$\max_{p \in \{p_-, p_+\}} |\widehat{Q}_{p,n}(u) - Q_{p,\pi}(u)| \xrightarrow{\mathbb{P}} 0.$$

If the target quantiles are finite, their estimation errors satisfy

$$\max_{p \in \{p_-, p_+\}} |\widehat{Q}_{p,n}(u) - Q_p(h)| \le \underbrace{\max_{p \in \{p_-, p_+\}} |\widehat{Q}_{p,n}(u) - Q_{p,\pi}(u)|}_{\textit{calibration estimation error}} + \underbrace{\varepsilon_{\mathrm{q}}(h)}_{\textit{population learning and approximation error}}, \tag{29}$$

and the left-hand side converges in probability to $\varepsilon_{\mathrm{q}}(h)$. In particular, if $\varepsilon_{\mathrm{q}}(h) = 0$, the prediction interval endpoints converge to $\mu(h) + Q_{p_-}(h)$ and $\mu(h) + Q_{p_+}(h)$.

The first condition in equation 28 makes the population neighborhoods shrink; the second controls the sampling fluctuation under geometric beta mixing. Every choice $K_n \asymp n^a$ with $0 < a < 1$ satisfies both conditions. We first establish a variance bound, then control the terms in equation 27.

Lemma 1 (Variance for localized sums)**.** *Under **(A1)**, define*

$$C_{\mathrm{mix}} = 3 + \frac{2[1 + \log(1 + 4C_\beta)]}{c_\beta} + \frac{2}{e^{c_\beta} - 1}.$$

For a deterministic measurable set B, put $p = n^{-1}\sum_i \mathbb{P}(U_i \in B) \in (0, 1]$. If g_i is a measurable function of V_i satisfying $|g_i| \le A\mathbf{1}\{U_i \in B\}$ for some $A > 0$, then

$$\mathrm{Var}\left(\sum_{i=1}^{n} g_i\right) \le C_{\mathrm{mix}} A^2 np \log(\mathrm{e}/p). \tag{30}$$

Proof. Let $p_i = \mathbb{P}(U_i \in B)$. Then $\mathrm{Var}(g_i) \le A^2 p_i$. Cauchy–Schwarz gives

$$\sum_{i=1}^{n-k} |\operatorname{Cov}(g_i, g_{i+k})| \le A^2 \sum_{i=1}^{n-k} \sqrt{p_i p_{i+k}} \le A^2 np.$$

The beta-mixing covariance inequality gives the alternative bound $4A^2 nC_\beta e^{-c_\beta k}$. Hence,

$$\mathrm{Var}\left(\sum_{i=1}^{n} g_i\right) \le A^2 n \left[p + 2\sum_{k\ge 1} \min\{p, 4C_\beta e^{-c_\beta k}\}\right].$$

Splitting the lag sum at $\lceil c_\beta^{-1} \log((1 + 4C_\beta)/p)\rceil$ and summing its geometric tail bounds the bracket by $C_{\mathrm{mix}} p \log(\mathrm{e}/p)$. □

Lemma 2 (Sampling fluctuation R)**.** *Assume **(A1)**, **(A2)**, and **(A4)**, and let $p_{*,n} = \min_e p_{e,n}(b_0)$. If $K_n/n \le p_{*,n}$, then, for $J = |\mathcal{R}|$ and $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$R_{n,\mathcal{R}}(u) \le 2\kappa \sqrt{\frac{C_{\mathrm{mix}} M(J + 1) \log(\mathrm{e}n/K_n)}{\delta K_n}}. \tag{31}$$

Consequently, equation 28 implies $R_{n,\mathcal{R}}(u) \xrightarrow{\mathbb{P}} 0$.

Proof. For each expert, use the deterministic radius $b_{e,n}$ with $p_{e,n}(b_{e,n}) = K_n/n$. Define

$$I_{e,n} = \{i : U_i \in B_e(u, b_{e,n})\}, \qquad N_{e,n} = |I_{e,n}|,$$

and

$$S_{e,n}(r) = \sum_{i \in I_{e,n}} a_e(u, U_i)\big[\mathbf{1}\{E_i \le r\} - \overline{F}_{e,n}(r \mid u)\big].$$

The definition equation 25 gives $\mathbb{E}S_{e,n}(r) = 0$. Individual summands need not have zero expectation. Lemma 1 gives

$$\mathrm{Var}(N_{e,n}) \le C_{\mathrm{mix}} K_n \log(\mathrm{e}n/K_n), \qquad \mathrm{Var}(S_{e,n}(r)) \le C_{\mathrm{mix}} a_+^2 K_n \log(\mathrm{e}n/K_n).$$

Since $\mathbb{E}N_{e,n} = K_n$, Chebyshev's inequality and a union bound over the M counts and MJ residual sums show that, with probability at least $1 - \delta$,

$$|N_{e,n} - K_n| \le t, \qquad |S_{e,n}(r)| \le a_+ t, \qquad t = \sqrt{\frac{C_{\mathrm{mix}} M(J + 1) K_n \log(\mathrm{e}n/K_n)}{\delta}},$$

simultaneously for all experts and thresholds.

The KNN set $A_{e,n}(u)$ and the ball set $I_{e,n}$ are nested, hence

$$|A_{e,n}(u) \triangle I_{e,n}| = |K_n - N_{e,n}|.$$

Since $|\mathbf{1}\{E_i \le r\} - \overline{F}_{e,n}(r \mid u)| \le 1$,

$$\left|\sum_{i \in A_{e,n}(u)} a_e(u, U_i)\big[\mathbf{1}\{E_i \le r\} - \overline{F}_{e,n}(r \mid u)\big]\right| \le |S_{e,n}(r)| + a_+|K_n - N_{e,n}| \le 2a_+ t.$$

The normalization is at least $a_- K_n$. Dividing by it bounds $|\widehat{F}_{e,n}(r \mid u) - \overline{F}_{e,n}(r \mid u)|$. Taking the gate-weighted average proves equation 31. □

Lemma 3 (Localization bias B)**.** *Under **(A3)**, **(A4)**, and $K_n/n \to 0$, the deterministic radius $b_n = \max_e b_{e,n}$ converges to zero, and, for all sufficiently large n,*

$$B_{n,\mathcal{R}}(u) \le \omega(b_n) \longrightarrow 0. \tag{32}$$

*Under **(A1)**, **(A4)**, and $K_n/n \to 0$, the maximum empirical retrieval radius*

$$\rho_n(u) = \max_{1\le e\le M} \max_{i\in A_{e,n}(u)} d_e(u, U_i)$$

also converges to zero in probability.

Proof. Fix $b \in (0, b_0]$. By **(A4)** and $K_n/n \to 0$, $p_{e,n}(b) > K_n/n$ eventually for every expert. The definition of $b_{e,n}$ therefore gives $b_{e,n} \le b$ eventually. Since there are finitely many experts and b can be arbitrarily small, $b_n \to 0$. Assumption **(A3)** and normalization of the gate imply

$$B_{n,\mathcal{R}}(u) \le \sum_{e=1}^{M} \pi_e(u) \max_{r\in\mathcal{R}} |\overline{F}_{e,n}(r \mid u) - G_e(r \mid u)| \le \omega(b_n).$$

For the empirical radius, let $N_{e,n}(b) = \sum_i \mathbf{1}\{U_i \in B_e(u, b)\}$. Since $p\log(\mathrm{e}/p) \le 1$ for $0 < p \le 1$, Lemma 1 gives $\mathrm{Var}(N_{e,n}(b)) \le C_{\mathrm{mix}} n$. Chebyshev's inequality yields, for all sufficiently large n,

$$\mathbb{P}\{N_{e,n}(b) < K_n\} \le \frac{C_{\mathrm{mix}}}{n[p_{e,n}(b) - K_n/n]^2} \longrightarrow 0.$$

A union bound over the finitely many experts proves $\mathbb{P}\{\rho_n(u) > b\} \to 0$ for every fixed $b > 0$. □

Lemma 4 (Population discrepancy D)**.** *For every nonempty finite $\mathcal{R} \subset \mathcal{I}$,*

$$|\Delta_{n,\mathcal{R}}(h) - D_{\mathcal{R}}(h)| \le R_{n,\mathcal{R}}(u) + B_{n,\mathcal{R}}(u).$$

Thus vanishing sampling and localization errors leave exactly the population mixture discrepancy on these thresholds.

Proof. The reverse triangle inequality for the maximum norm gives

$$|\Delta_{n,\mathcal{R}}(h) - D_{\mathcal{R}}(h)| \le \max_{r\in\mathcal{R}} |\widehat{F}_n(r \mid u) - G_\pi(r \mid u)|.$$

Adding and subtracting $\overline{F}_n(r \mid u)$ bounds the right-hand side by $R_{n,\mathcal{R}}(u) + B_{n,\mathcal{R}}(u)$. □

Proof of Theorem 2. Lemmas 2 and 3 give

$$\max_{r\in\mathcal{R}} |\widehat{F}_n(r \mid u) - G_\pi(r \mid u)| \le R_{n,\mathcal{R}}(u) + B_{n,\mathcal{R}}(u) \xrightarrow{\mathbb{P}} 0.$$

Lemma 4 then gives $\Delta_{n,\mathcal{R}}(h) \xrightarrow{\mathbb{P}} D_{\mathcal{R}}(h)$.

Fix a sufficiently small $a > 0$ so that

$$\mathcal{R}_a = \{Q_{p,\pi}(u) - a, Q_{p,\pi}(u) + a : p \in \{p_-, p_+\}\} \subset \mathcal{I}.$$

By **(A5)**,

$$m(a) = \min_{p\in\{p_-,p_+\}} \{p - G_\pi(Q_{p,\pi}(u) - a \mid u),\ G_\pi(Q_{p,\pi}(u) + a \mid u) - p\} > 0.$$

On the event $\max_{r\in\mathcal{R}_a} |\widehat{F}_n(r \mid u) - G_\pi(r \mid u)| < m(a)$,

$$\widehat{F}_n(Q_{p,\pi}(u) - a \mid u) < p < \widehat{F}_n(Q_{p,\pi}(u) + a \mid u)$$

for both probabilities. Monotonicity of the empirical CDF gives $|\widehat{Q}_{p,n}(u) - Q_{p,\pi}(u)| \le a$. This event has probability tending to one, and a can be arbitrarily small, proving consistency for the mixture quantiles.

The triangle inequality gives equation 29. Its reverse form gives

$$\left| \max_{p\in\{p_-,p_+\}} |\widehat{Q}_{p,n}(u) - Q_p(h)| - \varepsilon_{\mathrm{q}}(h) \right| \le \max_{p\in\{p_-,p_+\}} |\widehat{Q}_{p,n}(u) - Q_{p,\pi}(u)| \xrightarrow{\mathbb{P}} 0.$$

The endpoint conclusion follows. □

Remark 1 (Dense retrieval and temperature)**.** Consider dense softmax weights

$$\widetilde{w}_{i,n}(u) = \frac{\exp\{s(u, U_i)/\tau_n\}}{\sum_{j=1}^{n} \exp\{s(u, U_j)/\tau_n\}},$$

with a fixed score satisfying $s_- \le s(u, v) \le s_+$ for $v \in \mathcal{S}$. If $\tau_n \ge \tau_0 > 0$ along a subsequence, then

$$\widetilde{w}_{i,n}(u) \ge \frac{\exp\{-(s_+ - s_-)/\tau_0\}}{n}$$

along that subsequence. Suppose a fixed neighborhood B of u satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(U_i \notin B) \ge p_{\text{out}} > 0$$

for all sufficiently large n. Under **(A1)**, Lemma 1 applied to the complement of B shows that its empirical fraction is at least $p_{\text{out}}/2$ with probability tending to one. Consequently, along the same subsequence, with probability tending to one,

$$\sum_{i:U_i \notin B} \widetilde{w}_{i,n}(u) \ge \frac{p_{\text{out}}}{2} e^{-(s_+ - s_-)/\tau_0} > 0.$$

Thus dense retrieval with bounded fixed scores cannot guarantee localization unless $\tau_n \to 0$, when temperature is its only changing localization parameter. Nonvanishing mass outside the neighborhood can prevent convergence to the local population distribution when residual laws vary across neighborhoods. Vanishing temperature alone is not sufficient: the score must favor the query neighborhood and the weighted sampling fluctuation must vanish. Top-K_n retrieval provides a localization parameter through $K_n/n \to 0$, while permitting fixed positive temperatures within the selected set. Equation equation 28 also controls its sampling fluctuation. Localization is a sufficient route to these expert limits; it is not necessary for correct target quantiles when a nonlocal distribution or the expert mixture already has those quantiles.

B.4 Population quantile error and the Winkler objective

The remaining error $\varepsilon_{\text{q}}(h)$ concerns the endpoints of the complete population mixture. Its connection to the learning objective follows from the standard quantile interpretation of the interval score; see (Gneiting & Raftery, 2007, Section 6). For residual endpoints $\ell \le v$, define

$$\text{Wink}_\alpha(\ell, v; E) = (v - \ell) + \frac{2}{\alpha}(\ell - E)_+ + \frac{2}{\alpha}(E - v)_+, \qquad (x)_+ = \max\{x, 0\},$$

and let

$$L_h(\ell, v) = \mathbb{E}[\text{Wink}_\alpha(\ell, v; E) \mid H = h]\,.$$

The oracle conditional risk is

$$L_h^\star = L_h\big(Q_{p_-}(h), Q_{p_+}(h)\big)\,.$$

Remark (Winkler risk and population quantile error). The interval score is strictly consistent for the corresponding lower and upper quantiles. In particular, for the population mixture quantiles,

$$L_h\big(Q_{p_-,\pi}(u), Q_{p_+,\pi}(u)\big) - L_h^\star = \frac{2}{\alpha} \sum_{p \in \{p_-, p_+\}} \int_{Q_p(h)}^{Q_{p,\pi}(u)} \big(F_H(t \mid h) - p\big)\, dt.$$

Under the strict-identification condition in (A6), every term on the right-hand side is nonnegative, and the excess risk is zero if and only if

$$Q_{p,\pi}(u) = Q_p(h) \qquad \text{for } p \in \{p_-, p_+\},$$

or equivalently $\varepsilon_{\text{q}}(h) = 0$. Moreover, along any sequence of fitted population mixtures, vanishing excess Winkler risk implies $\varepsilon_{\text{q}}(h) \to 0$.

Thus the population Winkler objective directly targets the remaining endpoint error in Theorem 2. Importantly, this does not require each expert distribution to equal the target conditional law: individual experts may be biased as long as their gated mixture has the correct two target quantiles.

If the mixture family can represent the oracle endpoints and learning attains the oracle conditional Winkler risk, then $\varepsilon_{\mathrm{q}}(h) = 0$, and Theorem 2 yields target quantile consistency. For an independent test residual drawn from $F_H(\cdot \mid h)$, continuity at the target quantiles then also gives limiting coverage $1 - \alpha$.

This argument concerns the exact population Winkler score evaluated at the hard mixture quantiles. RareCP is trained with the smooth surrogate described in Appendix A. Relating that surrogate to the population score additionally requires control of surrogate approximation, generalization, and optimization error. We therefore do not assume that training drives $\varepsilon_{\mathrm{q}}(h)$ to zero; the calibration result retains $\varepsilon_{\mathrm{q}}(h)$ as the population approximation and learning error of the fitted model.

C EXPERIMENTS AND EVALUATION

This section provides additional experimental details and evaluation results that complement the main paper.

C.1 CONTROLLED RETRIEVAL-SCALING EXPERIMENT

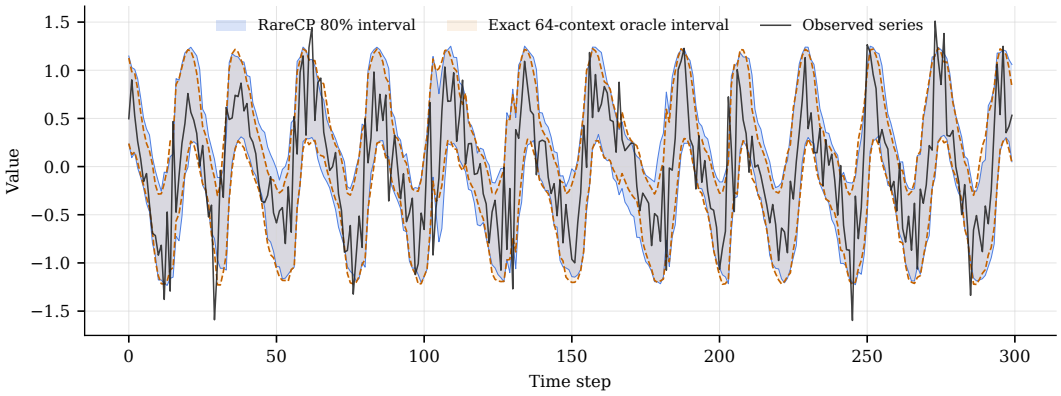


Figure 5: **Prediction intervals on the controlled stochastic-cycle process.** We show 300 consecutive observations from one held-out test series. RareCP uses a calibration set of size $n = 30{,}000$ and the per-expert neighborhood size $K_n = 24$ selected by validation Winkler score. Blue shading denotes RareCP's 80% prediction interval, and the dashed orange boundaries denote the exact history-conditional 80% oracle interval obtained from $F_H(\cdot \mid H_t)$. The oracle conditions on exactly the same 64 observations available to RareCP and integrates over $p(Z_t \mid H_t)$; it does not observe the realized latent state. The black line shows the generated observations.

This subsection describes the controlled study in Section **??**. The generator provides an exact latent conditional distribution, while a numerical filter provides a reference conditional on the same observation history available to RareCP.

Data-generating process. We generate independent realizations of

$$Z_t = \mathrm{wrap}\left(Z_{t-1} + \frac{2\pi}{24} + \eta_t\right), \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.16^2), \tag{33}$$

$$Y_t = m(Z_t) + s(Z_t)\epsilon_t, \qquad \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{34}$$

$$m(z) = 0.70\sin z + 0.22\sin(2z), \qquad s(z) = 0.30\exp(0.45\cos z). \tag{35}$$

Here wrap maps angles to $[-\pi, \pi)$, Z_0 is uniform on the circle, and the two innovation sequences and initial state are mutually independent. The model combines a harmonic seasonal mean, random timing fluctuations, and seasonally varying volatility.

The uniform phase distribution is stationary. The wrapped-Gaussian transition density is continuous and strictly positive on the compact circle, giving geometric mixing of the latent process. Independent observation noise and fixed-length observation windows preserve geometric mixing, although overlapping contexts introduce dependence between adjacent forecast examples.

RareCP receives

$$H_t = (Y_{t-64}, \ldots, Y_{t-1}).$$

The fixed backbone predicts the unconditional mean, $\mu(H_t) = 0$, so the signed residual is $E_t = Y_t$. Writing Φ for the standard normal CDF, the exact latent conditional residual law and quantiles are

$$F_Z(r \mid z) = \Phi\left(\frac{r - m(z)}{s(z)}\right), \qquad Q_p^Z(z) = m(z) + s(z)\Phi^{-1}(p). \tag{36}$$

The latent state is not provided as input to RareCP.

Training and evaluation data. We train the full teacher-bank, hypernetwork, distillation, and ten-expert MoE pipeline on a separate, independently generated panel containing 32,768 training examples. Dataset descriptors are computed from the fitting data. The trained functions, descriptors, and retrieval temperature are then frozen. We use neither ACI nor rolling calibration updates.

Evaluation uses ten seeds, each containing 32 independent series, with total calibration sizes ranging from 640 to 10^6. Each seed provides 8,192 validation queries and 8,192 test queries from disjoint groups of series. The same test observations are retained across calibration sizes. Variation across evaluation seeds therefore reflects calibration and evaluation randomness, rather than representation retraining.

Each expert retrieves from a temporally thinned candidate pool containing approximately $n/65$ examples, where n denotes the total calibration size. Some of the smallest nominal sizes produce identical candidate pools, explaining the initial plateau and irregularity in the main-text curves. If $n_e \asymp n/65$ denotes an expert's candidate-pool size, then

$$\frac{K_n}{n_e} \asymp \frac{65K_n}{n}, \qquad \frac{\log(\mathrm{e}n_e/K_n)}{K_n} = \frac{\log(\mathrm{e}n/K_n)}{K_n} + O(K_n^{-1}).$$

Thus the fixed thinning factor does not change the two asymptotic growth conditions, although it matters when interpreting the finite-sample diagnostics.

Selection of the retrieval size. For each retrieval method, evaluation seed, and calibration size, we select K_n on a separate selection set by minimizing normalized Winkler score:

$$K_n^\star \in \arg\min_K \frac{1}{|\mathcal{T}_{\mathrm{sel}}|\sigma_{\mathrm{sel}}} \sum_{t\in\mathcal{T}_{\mathrm{sel}}} \mathrm{Wink}_{0.2}\left(\widehat{Q}^{(K)}_{0.1,n}(H_t), \widehat{Q}^{(K)}_{0.9,n}(H_t); E_t\right).$$

Here $\sigma_{\mathrm{sel}} > 0$ is a fixed target-scale normalizer shared by all candidates. The selected K_n is then evaluated on the disjoint test set. Selection uses observed targets, not latent states or oracle quantiles.

Latent references. Latent-oracle retrieval uses the true conditional mean and log-scale state to retrieve calibration residuals, assigning uniform weights within each expert neighborhood. It still incurs finite-calibration error.

The exact latent oracle instead uses $Q_p^Z(Z_t)$ from equation 36 directly. It requires no calibration set and has no retrieval-size parameter.

Reference using the same observation history. The 64-context oracle knows the generator but receives only H_t. Its predictive residual CDF is

$$F_H(r \mid h) = \int_{-\pi}^{\pi} F_Z(r \mid z)\, p(z \mid H_t = h)\, dz. \tag{37}$$

This integrates uncertainty about the current phase instead of conditioning on its realized value.

We approximate the predictive phase distribution by circular-grid filtering. For each context, the filter starts from the stationary uniform prior, processes the 64 observations using the transition law and Gaussian observation likelihood, and applies one prediction step after the final observation. It uses neither the realized latent state nor observations outside the context. The resulting mixture CDF is inverted to obtain its 0.1 and 0.9 quantiles.

The reported filter uses 512 grid states. Doubling the resolution to 1,024 states changes the checked quantiles by less than 4.3×10^{-11}. This oracle has no calibration or retrieval error and provides a reference with the same input information as RareCP.

Reported metrics. We report the mean maximum endpoint error

$$\widehat{\mathcal{E}}_H = \frac{1}{|\mathcal{T}_{\mathrm{test}}|} \sum_{t\in\mathcal{T}_{\mathrm{test}}} \max_{p\in\{0.1,0.9\}} \left|\widehat{Q}_{p,n}(H_t) - Q_p(H_t)\right|,$$

where $Q_p(H_t)$ is computed using the history-conditional numerical filter described above, together with the oracle-relative Winkler score defined below.

C.2 Benchmark statistics

Table 2: Per-dataset benchmark statistics for the 22 GIFT-Eval datasets. Each series is split chronologically into val_calib, val_test, calib, and test, and all splits use the same fixed set of series (#series). Hyperparameter search trains on val_calib and is scored on val_test. Test starts at the official GIFT-Eval test start and is capped at 50,000 targets. σ_{norm} is the std of all pre-test values of a dataset. Targets are divided by it, so nWink and nW are comparable across datasets. Mean, std, min, max, and interquartile range (IQR) are computed on the test targets in original units. Missing values are set to the dataset mean.

		Split sizes					**Test targets (original units)**				
Dataset	#series	val_calib	val_test	calib	test	σ_{norm}	mean	std	min	max	IQR
electricity_15T	370	4,500	7,500	30,000	50,000	3,584.3	509.3	2,281.4	0.00	57,400	233.8
electricity_H	370	4,500	7,500	30,000	50,000	14,383	2,156.0	10,132	0.00	255,600	949.1
electricity_D	370	4,500	7,500	30,000	50,000	300,791	62,015	331,366	0.00	8,794,099	26,315
solar_10T	137	4,500	7,500	30,000	50,000	10.19	3.739	7.684	0.00	70.40	3.650
solar_H	137	4,500	7,500	30,000	50,000	61.06	31.18	54.81	0.00	427.5	45.70
LOOP_SEATTLE_5T	323	4,500	7,500	30,000	50,000	12.76	60.86	8.437	2.078	77.96	6.055
LOOP_SEATTLE_H	323	4,500	7,500	30,000	50,000	11.94	59.11	10.32	3.256	73.95	6.210
SZ_TAXI_15T	156	4,500	7,500	30,000	50,000	12.18	10.64	9.799	0.00	61.73	12.99
SZ_TAXI_H	156	4,500	7,500	30,000	14,976	11.58	10.73	9.326	0.00	57.01	12.44
M_DENSE_H	30	4,500	7,500	30,000	28,800	803.6	563.8	796.6	0.00	4,980.0	521.3
M_DENSE_D	30	1,830	3,060	12,390	2,700	639.4	577.7	642.0	0.00	2,993.5	431.6
m4_hourly	414	4,500	7,500	30,000	19,872	35,495	7,324.8	42,010	10.00	691,371	970.0
m4_daily	3,572	4,500	7,500	50,000	50,008	4,238.8	7,006.2	5,572.4	38.00	47,330	9,892.7
m4_weekly	359	4,500	7,500	30,000	4,667	3,511.1	5,488.9	5,244.1	110.5	36,404	5,136.0
m4_monthly	3,125	4,500	7,500	50,000	50,000	3,222.4	6,698.0	5,437.8	141.7	51,621	7,048.7
m4_quarterly	6,250	4,500	7,500	50,000	50,000	3,469.5	7,075.4	5,193.2	83.40	68,878	7,400.2
temperature_rain_with_missing	3,125	4,500	7,500	50,000	50,000	14.82	17.16	24.03	0.00	206.4	26.00
saugeenday_D	1	2,472	4,120	16,485	600	39.13	30.87	40.25	7.600	425.0	19.22
saugeenday_W	1	339	565	2,263	160	35.36	33.13	38.50	6.871	291.3	22.23
saugeenday_M	1	67	112	453	84	28.50	33.31	29.86	6.573	150.4	24.23
us_births_D	1	711	1,185	4,745	600	1,069.8	10,668	1,245.4	7,835.0	12,851	2,215.8
us_births_W	1	92	154	621	112	5,142.8	73,665	3,354.3	66,467	82,962	3,942.5
Total	19,252	77,511	129,196	596,957	672,579						

C.3 FULL EXPERIMENT RESULTS

C.3.1 FULL CONFORMAL PREDICTION AND UNCERTAINTY RESULTS

Table 3: Full results for conformal prediction competitors (1/2) at 80% target coverage across four backbones using the resampled runs. Cells report nWink and nW / Cov normalized by the target standard deviation. **bold** identifies the lowest nWink among the CP competitors and RareCP, while underlining identifies the lowest nWink overall across all three tables. Rankings use unrounded values and also apply to the unweighted *Average* rows.

	Conformal Prediction Competitors (1/2)											
	ACI		COP		Decay-OGD		DtACI		ECI		PID	
Dataset	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑
TiRex backbone												
electricity_15T	0.034	0.022 / 0.796	0.033	0.020 / 0.778	0.045	0.015 / 0.799	0.037	0.017 / 0.800	0.034	0.021 / 0.790	0.034	0.020 / 0.789
electricity_D	0.043	0.029 / 0.794	0.043	0.028 / 0.793	0.063	0.015 / 0.799	0.047	0.023 / 0.800	0.045	0.029 / 0.797	0.045	0.030 / 0.798
electricity_H	0.031	0.018 / 0.793	0.034	0.016 / 0.775	0.040	0.012 / 0.799	0.033	0.015 / 0.800	0.031	0.017 / 0.795	0.031	0.019 / 0.788
LOOP_SEATTLE_5T	1.029	0.623 / 0.800	1.017	0.599 / 0.777	1.052	0.605 / 0.800	1.056	0.602 / 0.800	1.010	0.599 / 0.781	1.014	0.601 / 0.782
LOOP_SEATTLE_H	1.194	0.487 / 0.800	1.191	0.473 / 0.792	1.195	0.461 / 0.800	1.197	0.472 / 0.800	1.185	0.480 / 0.792	1.188	0.479 / 0.797
m4_daily	0.085	0.049 / 0.797	0.085	0.047 / 0.786	0.088	0.043 / 0.800	0.090	0.048 / 0.800	0.085	0.049 / 0.790	0.086	0.050 / 0.787
m4_hourly	0.008	0.005 / 0.794	0.008	0.004 / 0.798	0.009	0.004 / 0.800	0.009	0.004 / 0.802	0.008	0.004 / 0.794	0.009	0.004 / 0.795
m4_monthly	0.361	0.216 / 0.792	0.363	0.198 / 0.776	0.424	0.170 / 0.800	0.386	0.188 / 0.800	0.358	0.202 / 0.785	0.360	0.202 / 0.783
m4_quarterly	0.539	0.280 / 0.797	0.533	0.279 / 0.793	0.559	0.229 / 0.800	0.567	0.228 / 0.800	0.527	0.279 / 0.787	0.527	0.277 / 0.785
m4_weekly	0.303	0.183 / 0.796	0.306	0.179 / 0.796	0.314	0.161 / 0.799	0.304	0.163 / 0.798	0.302	0.167 / 0.779	0.303	0.165 / 0.782
M_DENSE_D	0.422	0.216 / 0.802	0.414	0.183 / 0.790	0.419	0.190 / 0.799	0.415	0.208 / 0.800	0.406	0.211 / 0.784	0.413	0.182 / 0.791
M_DENSE_H	0.389	0.236 / 0.799	0.388	0.231 / 0.798	0.391	0.225 / 0.800	0.386	0.230 / 0.800	0.385	0.233 / 0.793	0.387	0.231 / 0.790
saugeenday_D	0.601	0.303 / 0.799	0.637	0.285 / 0.772	0.695	0.267 / 0.802	0.652	0.212 / 0.802	0.745	0.394 / 0.825	0.701	0.358 / 0.813
saugeenday_M	3.010	1.689 / 0.798	2.889	1.263 / 0.786	2.853	1.190 / 0.774	2.876	1.335 / 0.810	2.919	1.573 / 0.798	2.919	1.324 / 0.786
saugeenday_W	2.529	0.972 / 0.794	2.481	0.910 / 0.800	2.466	0.775 / 0.806	2.483	0.798 / 0.806	2.346	1.155 / 0.800	2.442	1.040 / 0.794
solar_10T	0.210	0.099 / 0.797	0.215	0.097 / 0.692	0.256	0.083 / 0.800	0.220	0.065 / 0.801	0.221	0.123 / 0.859	0.223	0.123 / 0.859
solar_H	0.415	0.151 / 0.800	0.394	0.199 / 0.799	0.413	0.149 / 0.800	0.412	0.131 / 0.800	0.386	0.197 / 0.795	0.397	0.205 / 0.795
SZ_TAXI_15T	0.975	0.638 / 0.798	0.959	0.636 / 0.798	0.960	0.616 / 0.800	0.982	0.615 / 0.802	0.958	0.637 / 0.798	0.963	0.634 / 0.795
SZ_TAXI_H	0.687	0.447 / 0.799	0.678	0.446 / 0.798	0.676	0.442 / 0.800	0.688	0.434 / 0.800	0.676	0.446 / 0.799	0.678	0.448 / 0.799
temperature_rain	4.221	2.286 / 0.798	4.123	2.407 / 0.798	4.464	2.116 / 0.800	4.426	2.121 / 0.800	4.098	2.392 / 0.799	4.131	2.408 / 0.799
us_births_D	1.346	0.766 / 0.793	1.299	0.694 / 0.792	1.371	0.480 / 0.673	1.297	0.672 / 0.790	1.318	0.629 / 0.778	1.290	0.728 / 0.792
us_births_W	0.983	0.536 / 0.759	0.973	0.557 / 0.768	1.031	0.410 / 0.643	0.983	0.548 / 0.777	0.972	0.563 / 0.786	0.973	0.570 / 0.786
Average	0.883	0.466 / 0.795	0.867	0.443 / 0.784	0.899	0.394 / 0.786	0.888	0.415 / 0.799	0.864	0.473 / 0.796	0.869	0.459 / 0.795
TimesFM2.5 backbone												
electricity_15T	0.034	0.023 / 0.796	0.033	0.021 / 0.780	0.047	0.015 / 0.799	0.037	0.017 / 0.800	0.033	0.021 / 0.787	0.034	0.022 / 0.790
electricity_D	0.044	0.029 / 0.795	0.044	0.027 / 0.788	0.063	0.017 / 0.799	0.048	0.024 / 0.800	0.043	0.027 / 0.795	0.044	0.027 / 0.795
electricity_H	0.033	0.020 / 0.794	0.032	0.018 / 0.782	0.041	0.015 / 0.799	0.035	0.016 / 0.800	0.032	0.019 / 0.791	0.032	0.019 / 0.796
LOOP_SEATTLE_5T	1.002	0.639 / 0.799	0.990	0.607 / 0.774	1.030	0.619 / 0.800	1.016	0.609 / 0.800	0.980	0.608 / 0.782	1.001	0.601 / 0.779
LOOP_SEATTLE_H	1.154	0.484 / 0.800	1.146	0.506 / 0.799	1.152	0.477 / 0.800	1.157	0.478 / 0.800	1.146	0.504 / 0.799	1.146	0.498 / 0.797
m4_daily	0.085	0.050 / 0.796	0.086	0.052 / 0.788	0.088	0.043 / 0.800	0.090	0.048 / 0.800	0.084	0.049 / 0.790	0.088	0.052 / 0.795
m4_hourly	0.008	0.005 / 0.793	0.009	0.005 / 0.794	0.010	0.004 / 0.800	0.009	0.004 / 0.804	0.009	0.005 / 0.797	0.009	0.005 / 0.796
m4_monthly	0.361	0.218 / 0.792	0.355	0.205 / 0.771	0.421	0.177 / 0.800	0.393	0.198 / 0.800	0.355	0.207 / 0.782	0.356	0.209 / 0.782
m4_quarterly	0.543	0.296 / 0.795	0.543	0.297 / 0.797	0.568	0.234 / 0.800	0.570	0.232 / 0.800	0.537	0.289 / 0.786	0.536	0.279 / 0.788
m4_weekly	0.312	0.191 / 0.796	0.317	0.187 / 0.796	0.326	0.170 / 0.799	0.316	0.170 / 0.798	0.310	0.183 / 0.785	0.317	0.183 / 0.792
M_DENSE_D	0.379	0.201 / 0.800	0.361	0.174 / 0.779	0.373	0.169 / 0.801	0.374	0.172 / 0.798	0.386	0.199 / 0.799	0.380	0.188 / 0.793
M_DENSE_H	0.386	0.229 / 0.799	0.386	0.229 / 0.793	0.389	0.218 / 0.800	0.387	0.227 / 0.800	0.384	0.225 / 0.797	0.388	0.230 / 0.790
saugeenday_D	0.648	0.320 / 0.787	0.673	0.305 / 0.752	0.709	0.265 / 0.803	0.687	0.217 / 0.802	0.668	0.290 / 0.782	0.668	0.250 / 0.773
saugeenday_M	2.794	1.424 / 0.810	3.012	1.514 / 0.774	2.794	1.223 / 0.750	2.849	1.465 / 0.798	2.962	1.512 / 0.786	3.277	1.824 / 0.786
saugeenday_W	2.586	0.807 / 0.800	2.512	1.074 / 0.800	2.594	0.873 / 0.794	2.567	0.845 / 0.812	2.453	1.142 / 0.794	2.376	0.996 / 0.794
solar_10T	0.218	0.106 / 0.797	0.223	0.094 / 0.785	0.265	0.079 / 0.800	0.221	0.069 / 0.801	0.226	0.092 / 0.772	0.223	0.116 / 0.834
solar_H	0.586	0.278 / 0.800	0.550	0.343 / 0.799	0.579	0.292 / 0.800	0.584	0.266 / 0.800	0.527	0.333 / 0.789	0.555	0.347 / 0.799
SZ_TAXI_15T	0.971	0.642 / 0.798	0.961	0.628 / 0.776	0.950	0.618 / 0.800	0.975	0.613 / 0.802	0.949	0.634 / 0.798	0.964	0.645 / 0.795
SZ_TAXI_H	0.683	0.450 / 0.799	0.675	0.447 / 0.798	0.675	0.449 / 0.800	0.685	0.439 / 0.800	0.674	0.448 / 0.798	0.675	0.443 / 0.798
temperature_rain	4.202	2.205 / 0.799	4.125	2.420 / 0.799	4.430	2.050 / 0.800	4.301	2.071 / 0.800	4.070	2.343 / 0.800	4.118	2.389 / 0.799
us_births_D	1.169	0.598 / 0.793	1.111	0.555 / 0.787	1.117	0.467 / 0.783	1.090	0.488 / 0.790	1.112	0.499 / 0.788	1.111	0.501 / 0.787
us_births_W	1.096	0.567 / 0.777	1.084	0.575 / 0.768	1.130	0.426 / 0.670	1.097	0.580 / 0.786	1.083	0.570 / 0.777	1.086	0.569 / 0.777
Average	0.877	0.445 / 0.796	0.874	0.467 / 0.785	0.898	0.405 / 0.791	0.886	0.420 / 0.799	0.865	0.464 / 0.790	0.881	0.472 / 0.792
Toto2 backbone												
electricity_15T	0.029	0.019 / 0.795	0.029	0.017 / 0.781	0.037	0.013 / 0.799	0.031	0.015 / 0.800	0.030	0.019 / 0.790	0.029	0.018 / 0.791
electricity_D	0.044	0.028 / 0.795	0.044	0.028 / 0.793	0.062	0.016 / 0.799	0.048	0.021 / 0.800	0.045	0.029 / 0.798	0.044	0.029 / 0.797
electricity_H	0.031	0.018 / 0.793	0.030	0.018 / 0.787	0.037	0.012 / 0.799	0.032	0.014 / 0.800	0.029	0.017 / 0.791	0.030	0.017 / 0.793
LOOP_SEATTLE_5T	1.011	0.603 / 0.800	0.999	0.580 / 0.782	1.023	0.591 / 0.800	1.024	0.588 / 0.800	0.988	0.586 / 0.782	1.001	0.581 / 0.782
LOOP_SEATTLE_H	1.095	0.422 / 0.800	1.093	0.414 / 0.786	1.097	0.408 / 0.800	1.098	0.415 / 0.800	1.091	0.419 / 0.795	1.090	0.423 / 0.798
m4_daily	0.086	0.050 / 0.795	0.085	0.048 / 0.791	0.088	0.042 / 0.800	0.089	0.047 / 0.800	0.092	0.056 / 0.791	0.087	0.050 / 0.795
m4_hourly	0.008	0.005 / 0.795	0.008	0.004 / 0.795	0.009	0.004 / 0.800	0.008	0.004 / 0.801	0.008	0.004 / 0.801	0.007	0.004 / 0.793
m4_monthly	0.356	0.212 / 0.795	0.347	0.199 / 0.782	0.411	0.163 / 0.800	0.376	0.183 / 0.800	0.347	0.197 / 0.786	0.356	0.197 / 0.790
m4_quarterly	0.528	0.271 / 0.797	0.518	0.278 / 0.779	0.547	0.224 / 0.800	0.531	0.225 / 0.800	0.516	0.270 / 0.785	0.517	0.266 / 0.783
m4_weekly	0.308	0.171 / 0.797	0.301	0.165 / 0.778	0.311	0.161 / 0.798	0.303	0.157 / 0.797	0.296	0.162 / 0.780	0.305	0.163 / 0.794
M_DENSE_D	0.399	0.173 / 0.801	0.398	0.167 / 0.796	0.404	0.167 / 0.800	0.399	0.180 / 0.800	0.393	0.167 / 0.787	0.398	0.164 / 0.794
M_DENSE_H	0.366	0.215 / 0.800	0.366	0.215 / 0.797	0.369	0.208 / 0.800	0.366	0.212 / 0.800	0.365	0.211 / 0.798	0.365	0.211 / 0.794
saugeenday_D	0.608	0.300 / 0.787	0.651	0.300 / 0.788	0.652	0.291 / 0.800	0.617	0.202 / 0.798	0.653	0.313 / 0.803	0.629	0.290 / 0.790
saugeenday_M	2.584	1.320 / 0.810	2.646	1.255 / 0.798	2.492	1.183 / 0.798	2.567	1.348 / 0.798	2.800	1.531 / 0.774	3.219	1.980 / 0.786
saugeenday_W	2.540	0.944 / 0.794	2.640	1.327 / 0.787	2.592	0.834 / 0.794	2.610	0.824 / 0.806	2.437	0.984 / 0.794	2.457	1.017 / 0.794
solar_10T	0.199	0.110 / 0.782	0.199	0.092 / 0.754	0.242	0.085 / 0.800	0.211	0.059 / 0.801	0.199	0.091 / 0.795	0.210	0.112 / 0.850
solar_H	0.464	0.156 / 0.800	0.465	0.139 / 0.790	0.454	0.184 / 0.800	0.453	0.154 / 0.800	0.433	0.221 / 0.799	0.464	0.151 / 0.800
SZ_TAXI_15T	0.981	0.639 / 0.797	0.956	0.632 / 0.797	0.954	0.613 / 0.800	0.979	0.604 / 0.802	0.951	0.621 / 0.797	0.967	0.637 / 0.794
SZ_TAXI_H	0.683	0.442 / 0.800	0.675	0.439 / 0.798	0.675	0.440 / 0.800	0.685	0.430 / 0.800	0.675	0.440 / 0.798	0.675	0.437 / 0.798
temperature_rain	4.079	2.173 / 0.799	3.986	2.297 / 0.798	4.309	2.042 / 0.800	4.217	2.014 / 0.800	3.942	2.248 / 0.800	3.993	2.307 / 0.798
us_births_D	1.109	0.476 / 0.793	1.105	0.460 / 0.785	1.110	0.421 / 0.752	1.123	0.495 / 0.793	1.107	0.469 / 0.790	1.106	0.462 / 0.783
us_births_W	0.766	0.513 / 0.795	0.747	0.480 / 0.786	0.771	0.389 / 0.696	0.765	0.491 / 0.786	0.747	0.478 / 0.777	0.747	0.487 / 0.795
Average	0.831	0.421 / 0.796	0.831	0.435 / 0.788	0.847	0.386 / 0.793	0.842	0.395 / 0.799	0.825	0.433 / 0.791	0.850	0.455 / 0.795
ARIMA backbone												
electricity_15T	0.052	0.031 / 0.794	0.053	0.028 / 0.775	0.065	0.020 / 0.799	0.054	0.023 / 0.800	0.057	0.036 / 0.796	0.052	0.031 / 0.789
electricity_D	0.050	0.033 / 0.794	0.049	0.032 / 0.791	0.071	0.020 / 0.799	0.055	0.027 / 0.800	0.054	0.037 / 0.798	0.054	0.036 / 0.797
electricity_H	0.075	0.042 / 0.794	0.077	0.039 / 0.776	0.090	0.028 / 0.799	0.075	0.032 / 0.800	0.072	0.040 / 0.796	0.074	0.043 / 0.796
LOOP_SEATTLE_5T	1.071	0.666 / 0.800	1.046	0.647 / 0.779	1.088	0.659 / 0.800	1.095	0.645 / 0.800	1.043	0.653 / 0.789	1.051	0.671 / 0.797
LOOP_SEATTLE_H	1.498	0.720 / 0.800	1.486	0.726 / 0.798	1.499	0.699 / 0.800	1.482	0.713 / 0.800	1.484	0.725 / 0.796	1.488	0.720 / 0.796
m4_daily	0.087	0.051 / 0.795	0.085	0.049 / 0.788	0.089	0.043 / 0.801	0.089	0.047 / 0.800	0.091	0.052 / 0.798	0.088	0.051 / 0.789
m4_hourly	0.016	0.009 / 0.800	0.016	0.008 / 0.796	0.018	0.007 / 0.800	0.016	0.007 / 0.802	0.016	0.008 / 0.795	0.016	0.008 / 0.795
m4_monthly	0.419	0.249 / 0.793	0.414	0.241 / 0.783	0.488	0.200 / 0.800	0.457	0.221 / 0.800	0.414	0.236 / 0.785	0.434	0.231 / 0.794
m4_quarterly	0.602	0.327 / 0.794	0.611	0.335 / 0.779	0.646	0.259 / 0.800	0.650	0.258 / 0.800	0.603	0.324 / 0.794	0.599	0.323 / 0.785
m4_weekly	0.348	0.210 / 0.795	0.349	0.212 / 0.773	0.369	0.191 / 0.799	0.357	0.187 / 0.798	0.360	0.202 / 0.796	0.347	0.203 / 0.781
M_DENSE_D	0.555	0.387 / 0.795	0.543	0.365 / 0.795	0.563	0.372 / 0.801	0.566	0.393 / 0.802	0.545	0.377 / 0.811	0.544	0.360 / 0.794
M_DENSE_H	0.686	0.428 / 0.800	0.678	0.411 / 0.797	0.680	0.406 / 0.800	0.679	0.416 / 0.800	0.677	0.410 / 0.798	0.679	0.412 / 0.797
saugeenday_D	0.638	0.336 / 0.787	0.634	0.324 / 0.775	0.703	0.304 / 0.798	0.690	0.254 / 0.807	0.647	0.291 / 0.798	0.639	0.329 / 0.783
saugeenday_M	3.413	1.394 / 0.821	3.468	1.243 / 0.780	3.415	1.400 / 0.821	3.503	1.325 / 0.810	3.421	1.267 / 0.798	3.462	1.203 / 0.786
saugeenday_W	2.841	0.936 / 0.794	2.710	0.980 / 0.794	2.786	0.903 / 0.794	2.823	0.892 / 0.806	2.753	1.087 / 0.794	2.748	1.071 / 0.794
solar_10T	0.242	0.128 / 0.798	0.236	0.124 / 0.821	0.290	0.115 / 0.800	0.257	0.087 / 0.801	0.244	0.126 / 0.787	0.232	0.128 / 0.829
solar_H	0.883	0.449 / 0.800	0.883	0.446 / 0.798	0.879	0.439 / 0.800	0.877	0.442 / 0.800	0.880	0.446 / 0.799	0.882	0.446 / 0.799
SZ_TAXI_15T	0.987	0.683 / 0.795	0.973	0.678 / 0.799	0.970	0.653 / 0.800	0.995	0.646 / 0.802	0.970	0.670 / 0.798	0.970	0.668 / 0.798
SZ_TAXI_H	0.764	0.506 / 0.799	0.755	0.503 / 0.798	0.757	0.504 / 0.800	0.768	0.493 / 0.800	0.753	0.505 / 0.799	0.758	0.499 / 0.795
temperature_rain	3.782	2.006 / 0.798	3.706	2.168 / 0.799	4.002	1.967 / 0.800	3.998	2.003 / 0.799	3.668	2.152 / 0.799	3.730	2.225 / 0.799
us_births_D	2.826	2.079 / 0.772	2.830	2.146 / 0.790	3.099	1.497 / 0.627	2.837	2.190 / 0.787	2.830	2.146 / 0.790	2.830	2.139 / 0.792
us_births_W	1.328	0.741 / 0.786	1.304	0.757 / 0.795	1.316	0.678 / 0.759	1.314	0.717 / 0.768	1.310	0.743 / 0.786	1.299	0.733 / 0.786
Average	1.053	0.564 / 0.796	1.041	0.566 / 0.791	1.086	0.516 / 0.791	1.074	0.546 / 0.799	1.041	0.570 / 0.796	1.044	0.570 / 0.794

Table 4: Full results for conformal prediction competitors (2/2) at 80% target coverage across four backbones using the resampled runs. Cells report nWink and nW / Cov normalized by the target standard deviation. **bold** identifies the lowest nWink among the CP competitors and RareCP, while underlining identifies the lowest nWink overall across all three tables. Rankings use unrounded values and also apply to the unweighted *Average* rows.

	Conformal Prediction Competitors (2/2)											
	SAOCP		SF-OGD		NexCP		KOWCPI		HopCPT		ResCP	
Dataset	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑
						TiRex backbone						
electricity_15T	0.057	0.047 / 0.912	0.044	0.016 / 0.799	0.032	0.019 / 0.786	0.054	0.002 / 0.480	0.032	0.022 / 0.801	0.041	0.014 / 0.702
electricity_D	0.087	0.073 / 0.927	0.061	0.017 / 0.799	0.041	0.025 / 0.790	0.070	0.008 / 0.736	0.041	0.028 / 0.821	0.048	0.021 / 0.712
electricity_H	0.046	0.036 / 0.911	0.039	0.013 / 0.799	0.029	0.017 / 0.790	0.045	0.006 / 0.714	0.028	0.017 / 0.799	0.032	0.014 / 0.728
LOOP_SEATTLE_5T	1.015	0.715 / 0.832	1.049	0.609 / 0.800	1.000	0.615 / 0.791	1.080	0.466 / 0.715	1.066	0.564 / 0.755	1.049	0.563 / 0.766
LOOP_SEATTLE_H	1.185	0.702 / 0.838	1.193	0.464 / 0.800	1.187	0.481 / 0.795	1.102	0.328 / 0.717	1.185	0.590 / 0.840	0.899	0.432 / 0.759
m4_daily	0.115	0.086 / 0.849	0.088	0.043 / 0.800	0.082	0.046 / 0.785	0.097	0.023 / 0.563	0.077	0.045 / 0.751	0.080	0.035 / 0.760
m4_hourly	0.009	0.006 / 0.917	0.009	0.004 / 0.800	0.008	0.004 / 0.783	0.008	0.002 / 0.594	0.007	0.004 / 0.782	0.009	0.003 / 0.727
m4_monthly	0.378	0.240 / 0.822	0.419	0.174 / 0.800	0.352	0.195 / 0.783	0.460	0.107 / 0.694	0.308	0.190 / 0.745	0.397	0.166 / 0.721
m4_quarterly	0.557	0.330 / 0.834	0.560	0.228 / 0.800	0.520	0.262 / 0.790	0.598	0.168 / 0.736	0.446	0.262 / 0.722	0.549	0.226 / 0.775
m4_weekly	0.309	0.205 / 0.835	0.311	0.168 / 0.799	0.297	0.166 / 0.785	0.302	0.074 / 0.595	0.261	0.142 / 0.757	0.316	0.136 / 0.727
M_DENSE_D	0.411	0.237 / 0.815	0.415	0.194 / 0.800	0.401	0.201 / 0.779	0.395	0.084 / 0.568	0.394	0.163 / 0.748	0.384	0.134 / 0.665
M_DENSE_H	0.394	0.264 / 0.823	0.390	0.226 / 0.800	0.387	0.230 / 0.792	0.485	0.114 / 0.699	0.376	0.199 / 0.737	0.398	0.205 / 0.755
saugeenday_D	0.694	0.446 / 0.872	0.700	0.283 / 0.800	0.637	0.276 / 0.802	0.581	0.155 / 0.695	0.594	0.230 / 0.721	0.646	0.199 / 0.693
saugeenday_M	3.058	1.870 / 0.810	2.892	1.222 / 0.774	2.954	1.146 / 0.726	2.715	1.333 / 0.798	2.487	2.155 / 0.917	2.652	1.325 / 0.837
saugeenday_W	2.453	1.239 / 0.856	2.487	0.825 / 0.806	2.478	0.855 / 0.781	2.220	0.585 / 0.756	**1.947**	0.763 / 0.703	2.099	0.873 / 0.844
solar_10T	0.224	0.147 / 0.900	0.252	0.084 / 0.800	0.195	0.098 / 0.784	0.217	0.037 / 0.659	0.267	0.126 / 0.761	0.245	0.046 / 0.754
solar_H	0.407	0.228 / 0.849	0.413	0.150 / 0.800	0.414	0.151 / 0.795	0.338	0.077 / 0.759	0.335	0.134 / 0.820	0.360	0.144 / 0.825
SZ_TAXI_15T	0.960	0.663 / 0.831	0.965	0.614 / 0.800	0.950	0.611 / 0.789	1.084	0.519 / 0.733	0.961	0.643 / 0.811	1.030	0.610 / 0.811
SZ_TAXI_H	0.682	0.473 / 0.826	0.678	0.439 / 0.800	0.675	0.433 / 0.790	0.739	0.406 / 0.770	0.666	0.439 / 0.802	0.722	0.423 / 0.783
temperature_rain	4.235	2.410 / 0.817	4.471	2.115 / 0.800	4.129	2.290 / 0.801	3.103	1.044 / 0.673	3.620	1.364 / 0.624	2.337	1.262 / 0.733
us_births_D	1.311	0.765 / 0.810	1.335	0.555 / 0.732	1.290	0.681 / 0.777	1.032	0.423 / 0.648	1.258	0.592 / 0.754	1.256	0.528 / 0.714
us_births_W	1.021	0.665 / 0.804	0.979	0.508 / 0.750	1.002	0.610 / 0.795	1.013	0.482 / 0.679	0.982	0.508 / 0.731	0.955	0.494 / 0.702
Average	0.891	0.539 / 0.850	0.898	0.407 / 0.794	0.866	0.428 / 0.786	0.806	0.293 / 0.681	0.788	0.417 / 0.768	0.750	0.357 / 0.750
						TimesFM2.5 backbone						
electricity_15T	0.058	0.049 / 0.925	0.046	0.016 / 0.799	0.033	0.020 / 0.784	0.043	0.007 / 0.579	0.033	0.022 / 0.796	0.036	0.016 / 0.696
electricity_D	0.087	0.073 / 0.923	0.061	0.019 / 0.799	0.042	0.026 / 0.787	0.056	0.011 / 0.599	0.041	0.028 / 0.817	0.048	0.021 / 0.706
electricity_H	0.048	0.037 / 0.908	0.041	0.015 / 0.799	0.031	0.018 / 0.788	0.048	0.007 / 0.748	0.030	0.019 / 0.796	0.033	0.015 / 0.731
LOOP_SEATTLE_5T	0.997	0.716 / 0.834	1.023	0.623 / 0.800	0.989	0.613 / 0.789	1.056	0.526 / 0.762	1.055	0.546 / 0.744	1.063	0.559 / 0.751
LOOP_SEATTLE_H	1.146	0.716 / 0.842	1.153	0.476 / 0.800	1.147	0.497 / 0.798	1.108	0.339 / 0.711	1.186	0.591 / 0.819	0.919	0.448 / 0.771
m4_daily	0.115	0.086 / 0.854	0.088	0.043 / 0.800	0.082	0.046 / 0.785	0.095	0.023 / 0.570	0.077	0.045 / 0.752	0.080	0.032 / 0.703
m4_hourly	0.009	0.007 / 0.904	0.009	0.004 / 0.800	0.008	0.004 / 0.782	0.011	0.001 / 0.606	0.007	0.004 / 0.791	0.010	0.003 / 0.732
m4_monthly	0.382	0.245 / 0.824	0.418	0.179 / 0.800	0.355	0.197 / 0.783	0.461	0.103 / 0.681	0.307	0.189 / 0.746	0.399	0.169 / 0.723
m4_quarterly	0.563	0.333 / 0.831	0.567	0.234 / 0.800	0.527	0.273 / 0.794	0.532	0.162 / 0.651	0.446	0.260 / 0.714	0.558	0.214 / 0.730
m4_weekly	0.319	0.213 / 0.830	0.323	0.172 / 0.799	0.305	0.173 / 0.783	0.329	0.084 / 0.594	0.269	0.144 / 0.752	0.320	0.139 / 0.714
M_DENSE_D	0.382	0.218 / 0.824	0.372	0.168 / 0.800	0.359	0.171 / 0.780	0.355	0.089 / 0.585	0.356	0.144 / 0.737	0.357	0.129 / 0.668
M_DENSE_H	0.393	0.258 / 0.825	0.388	0.220 / 0.800	0.383	0.220 / 0.793	0.477	0.114 / 0.715	0.388	0.217 / 0.781	0.394	0.198 / 0.756
saugeenday_D	0.697	0.431 / 0.857	0.710	0.276 / 0.805	0.655	0.275 / 0.793	0.631	0.149 / 0.710	0.689	0.275 / 0.848	0.621	0.182 / 0.771
saugeenday_M	2.948	1.758 / 0.798	2.824	1.193 / 0.750	2.813	1.170 / 0.750	2.461	1.374 / 0.798	2.632	1.502 / 0.839	2.546	1.239 / 0.794
saugeenday_W	2.572	1.264 / 0.844	2.552	0.781 / 0.806	2.536	0.898 / 0.794	2.429	0.720 / 0.762	2.722	1.759 / 0.913	2.334	0.861 / 0.808
solar_10T	0.230	0.151 / 0.882	0.256	0.086 / 0.800	0.206	0.103 / 0.801	0.218	0.044 / 0.686	0.273	0.076 / 0.788	0.259	0.072 / 0.726
solar_H	0.579	0.384 / 0.851	0.583	0.278 / 0.800	0.575	0.289 / 0.782	0.364	0.149 / 0.792	0.589	0.300 / 0.822	0.353	0.210 / 0.843
SZ_TAXI_15T	0.954	0.666 / 0.840	0.960	0.616 / 0.800	0.944	0.607 / 0.786	1.077	0.520 / 0.736	0.943	0.581 / 0.791	0.992	0.562 / 0.761
SZ_TAXI_H	0.678	0.477 / 0.827	0.675	0.443 / 0.800	0.673	0.435 / 0.787	0.741	0.387 / 0.758	0.658	0.413 / 0.777	0.744	0.420 / 0.783
temperature_rain	4.204	2.352 / 0.819	4.442	2.049 / 0.800	4.099	2.219 / 0.797	3.907	0.555 / 0.480	3.661	1.356 / 0.608	2.275	1.207 / 0.724
us_births_D	1.127	0.633 / 0.827	1.118	0.462 / 0.783	1.104	0.492 / 0.785	1.061	0.411 / 0.755	1.114	0.483 / 0.789	1.112	0.452 / 0.775
us_births_W	1.127	0.630 / 0.786	1.098	0.509 / 0.732	1.109	0.609 / 0.812	1.130	0.482 / 0.714	1.107	0.482 / 0.728	1.096	0.527 / 0.768
Average	0.892	0.532 / 0.848	0.896	0.403 / 0.794	0.862	0.425 / 0.788	0.845	0.284 / 0.681	0.845	0.429 / 0.780	0.752	0.349 / 0.747
						Toto2 backbone						
electricity_15T	0.045	0.036 / 0.923	0.037	0.014 / 0.799	**0.028**	0.016 / 0.785	0.042	0.003 / 0.548	0.028	0.018 / 0.796	0.030	0.014 / 0.722
electricity_D	0.089	0.076 / 0.939	0.061	0.017 / 0.799	0.042	0.025 / 0.788	0.057	0.010 / 0.594	**0.041**	0.027 / 0.818	0.047	0.022 / 0.723
electricity_H	0.038	0.028 / 0.886	0.036	0.013 / 0.799	0.028	0.016 / 0.797	0.042	0.006 / 0.754	**0.027**	0.017 / 0.799	0.035	0.012 / 0.718
LOOP_SEATTLE_5T	0.991	0.689 / 0.828	1.020	0.593 / 0.800	0.979	0.599 / 0.790	1.042	0.534 / 0.775	1.038	0.536 / 0.747	1.040	0.556 / 0.773
LOOP_SEATTLE_H	1.091	0.603 / 0.829	1.095	0.410 / 0.800	1.092	0.423 / 0.798	1.091	0.331 / 0.747	1.108	0.479 / 0.820	1.124	0.391 / 0.713
m4_daily	0.114	0.084 / 0.848	0.087	0.043 / 0.800	0.081	0.045 / 0.786	0.089	0.027 / 0.622	0.077	0.045 / 0.751	0.080	0.032 / 0.725
m4_hourly	0.009	0.006 / 0.903	0.009	0.004 / 0.800	0.007	0.004 / 0.782	0.008	0.002 / 0.639	0.007	0.004 / 0.789	0.009	0.003 / 0.691
m4_monthly	0.369	0.236 / 0.830	0.403	0.169 / 0.800	0.343	0.189 / 0.783	0.430	0.086 / 0.635	0.298	0.182 / 0.743	0.388	0.161 / 0.720
m4_quarterly	0.550	0.334 / 0.838	0.549	0.223 / 0.800	0.510	0.262 / 0.793	0.587	0.134 / 0.661	0.439	0.255 / 0.720	0.547	0.202 / 0.730
m4_weekly	0.306	0.196 / 0.830	0.308	0.163 / 0.799	0.296	0.162 / 0.785	0.318	0.066 / 0.576	0.263	0.133 / 0.752	0.314	0.134 / 0.732
M_DENSE_D	0.401	0.227 / 0.831	0.402	0.171 / 0.800	0.389	0.181 / 0.784	0.416	0.074 / 0.569	0.385	0.133 / 0.735	0.360	0.122 / 0.701
M_DENSE_H	0.373	0.243 / 0.821	0.368	0.208 / 0.800	**0.364**	0.210 / 0.793	0.401	0.124 / 0.659	0.367	0.215 / 0.787	0.375	0.189 / 0.755
saugeenday_D	0.658	0.386 / 0.845	0.659	0.286 / 0.800	0.602	0.256 / 0.790	0.672	0.114 / 0.643	0.644	0.240 / 0.835	0.590	0.294 / 0.859
saugeenday_M	3.084	2.023 / 0.833	2.521	1.211 / 0.786	2.526	1.151 / 0.774	2.380	1.248 / 0.810	2.473	1.437 / 0.825	2.474	1.092 / 0.770
saugeenday_W	2.579	1.238 / 0.850	2.590	0.799 / 0.794	2.560	0.909 / 0.787	2.393	0.583 / 0.731	2.773	1.861 / 0.922	2.280	0.806 / 0.823
solar_10T	0.219	0.142 / 0.908	0.238	0.088 / 0.800	0.190	0.094 / 0.776	0.261	0.013 / 0.597	0.266	0.074 / 0.789	0.227	0.078 / 0.688
solar_H	0.454	0.251 / 0.851	0.458	0.173 / 0.800	0.464	0.149 / 0.800	0.374	0.083 / 0.763	0.465	0.151 / 0.805	0.401	0.117 / 0.698
SZ_TAXI_15T	0.957	0.657 / 0.834	0.961	0.608 / 0.800	0.948	0.600 / 0.785	1.069	0.498 / 0.732	0.944	0.573 / 0.791	1.147	0.656 / 0.811
SZ_TAXI_H	0.679	0.469 / 0.826	0.676	0.439 / 0.800	0.673	0.431 / 0.791	0.743	0.346 / 0.726	**0.657**	0.407 / 0.778	0.719	0.423 / 0.783
temperature_rain	4.083	2.315 / 0.817	4.319	2.050 / 0.800	3.985	2.192 / 0.802	3.155	1.034 / 0.660	3.685	1.337 / 0.617	2.541	1.304 / 0.707
us_births_D	1.151	0.590 / 0.810	1.106	0.447 / 0.775	1.108	0.471 / 0.788	1.116	0.409 / 0.742	1.105	0.463 / 0.780	1.107	0.441 / 0.766
us_births_W	0.802	0.523 / 0.804	0.751	0.452 / 0.750	0.786	0.503 / 0.795	0.759	0.422 / 0.723	0.752	0.438 / 0.741	0.757	0.450 / 0.759
Average	0.866	0.516 / 0.849	0.848	0.390 / 0.796	0.818	0.404 / 0.789	0.793	0.279 / 0.677	0.811	0.410 / 0.779	0.754	0.341 / 0.744
						ARIMA backbone						
electricity_15T	0.083	0.066 / 0.912	0.064	0.021 / 0.799	0.048	0.026 / 0.790	0.063	0.009 / 0.570	0.047	0.027 / 0.792	0.054	0.021 / 0.714
electricity_D	0.094	0.079 / 0.914	0.069	0.022 / 0.799	0.047	0.029 / 0.789	0.072	0.006 / 0.578	0.046	0.032 / 0.823	0.052	0.023 / 0.706
electricity_H	0.113	0.087 / 0.900	0.087	0.031 / 0.800	0.072	0.038 / 0.789	0.082	0.012 / 0.614	0.066	0.037 / 0.796	0.068	0.027 / 0.725
LOOP_SEATTLE_5T	1.053	0.777 / 0.840	1.094	0.656 / 0.800	1.048	0.651 / 0.790	1.162	0.392 / 0.624	1.120	0.566 / 0.735	1.154	0.560 / 0.708
LOOP_SEATTLE_H	1.483	0.999 / 0.852	1.498	0.700 / 0.800	1.485	0.731 / 0.803	1.249	0.458 / 0.719	1.518	0.858 / 0.850	1.449	0.489 / 0.657
m4_daily	0.113	0.084 / 0.846	0.089	0.044 / 0.801	0.082	0.046 / 0.785	0.084	0.026 / 0.650	0.077	0.045 / 0.752	0.082	0.035 / 0.763
m4_hourly	0.017	0.012 / 0.893	0.018	0.008 / 0.800	0.015	0.008 / 0.783	0.015	0.003 / 0.594	0.013	0.007 / 0.792	0.018	0.006 / 0.778
m4_monthly	0.440	0.274 / 0.823	0.481	0.203 / 0.800	0.409	0.225 / 0.784	0.527	0.132 / 0.708	0.357	0.219 / 0.761	0.458	0.188 / 0.719
m4_quarterly	0.637	0.382 / 0.835	0.647	0.259 / 0.800	0.592	0.300 / 0.787	0.677	0.166 / 0.675	0.497	0.294 / 0.720	0.638	0.237 / 0.725
m4_weekly	0.365	0.244 / 0.834	0.368	0.194 / 0.799	0.345	0.192 / 0.783	0.457	0.062 / 0.537	0.303	0.158 / 0.754	0.366	0.159 / 0.724
M_DENSE_D	0.546	0.394 / 0.820	0.560	0.380 / 0.800	0.529	0.353 / 0.781	0.457	0.158 / 0.677	0.504	0.359 / 0.789	0.508	0.272 / 0.761
M_DENSE_H	0.688	0.458 / 0.824	0.679	0.407 / 0.800	0.678	0.407 / 0.791	0.619	0.271 / 0.699	0.683	0.455 / 0.821	0.579	0.326 / 0.750
saugeenday_D	0.711	0.498 / 0.870	0.700	0.303 / 0.798	0.636	0.301 / 0.783	0.651	0.158 / 0.698	0.724	0.387 / 0.879	0.640	0.227 / 0.689
saugeenday_M	3.784	1.674 / 0.821	3.416	1.317 / 0.810	3.425	1.316 / 0.810	3.144	1.386 / 0.821	3.393	1.804 / 0.839	3.548	1.244 / 0.774
saugeenday_W	2.806	1.521 / 0.825	2.753	0.900 / 0.794	2.782	0.905 / 0.781	2.348	0.625 / 0.700	3.638	2.513 / 0.922	2.461	0.800 / 0.794
solar_10T	0.257	0.177 / 0.871	0.275	0.121 / 0.800	0.228	0.122 / 0.769	0.303	0.040 / 0.664	0.315	0.085 / 0.754	0.269	0.080 / 0.666
solar_H	0.888	0.514 / 0.827	0.881	0.444 / 0.800	0.885	0.441 / 0.798	0.451	0.204 / 0.756	0.842	0.482 / 0.808	0.424	0.240 / 0.819
SZ_TAXI_15T	0.973	0.702 / 0.836	0.979	0.649 / 0.800	0.964	0.641 / 0.767	1.098	0.534 / 0.730	0.967	0.611 / 0.789	1.083	0.653 / 0.821
SZ_TAXI_H	0.758	0.546 / 0.830	0.758	0.500 / 0.800	0.751	0.494 / 0.790	0.824	0.445 / 0.768	0.740	0.476 / 0.786	0.805	0.481 / 0.781
temperature_rain	3.812	2.242 / 0.824	4.017	1.963 / 0.800	3.687	2.045 / 0.805	3.462	1.148 / 0.654	3.588	1.321 / 0.570	2.437	1.364 / 0.761
us_births_D	2.870	2.244 / 0.800	2.907	1.781 / 0.695	2.816	2.143 / 0.793	1.930	0.808 / 0.653	2.815	2.178 / 0.785	1.936	0.887 / 0.676
us_births_W	1.340	0.863 / 0.821	1.315	0.732 / 0.777	1.324	0.761 / 0.786	1.322	0.660 / 0.768	1.322	0.711 / 0.759	1.007	0.608 / 0.685
Average	1.083	0.675 / 0.846	1.075	0.529 / 0.794	1.039	0.553 / 0.788	0.954	0.350 / 0.675	1.072	0.619 / 0.785	0.911	0.406 / 0.736

Table 5: Full results for RareCP, native time series foundation model uncertainty, and MoE-CP at 80% target coverage. Cells report nWink and nW / Cov normalized by the target standard deviation. **bold** identifies the lowest nWink among the CP competitors and RareCP, while underlining identifies the lowest nWink overall across all three tables. Rankings use unrounded values and also apply to the unweighted *Average* rows. *Reimplemented following the original paper.

MoE-CP and Native TSFM UQ (1/2)

	MoE-CP*		TiRex		TiRex2		Chronos2		Moirai2	
Dataset	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑
electricity_15T	0.093	0.041 / 0.752	0.027	0.019 / 0.806	0.028	0.020 / 0.818	0.025	0.017 / 0.778	0.028	0.020 / 0.805
electricity_D	0.230	0.064 / 0.724	0.038	0.025 / 0.820	0.039	0.026 / 0.862	0.039	0.027 / 0.834	0.040	0.027 / 0.792
electricity_H	0.131	0.085 / 0.742	0.025	0.017 / 0.810	0.029	0.022 / 0.840	0.027	0.018 / 0.770	0.026	0.017 / 0.788
LOOP_SEATTLE_5T	1.231	0.514 / 0.656	0.940	0.526 / 0.729	0.911	0.552 / 0.756	0.926	0.525 / 0.706	0.891	0.573 / 0.771
LOOP_SEATTLE_H	1.349	0.551 / 0.757	0.960	0.571 / 0.784	0.970	0.556 / 0.805	0.944	0.540 / 0.736	1.104	0.517 / 0.781
m4_daily	0.247	0.121 / 0.810	0.072	0.047 / 0.814	0.079	0.061 / 0.875	0.084	0.063 / 0.852	0.078	0.047 / 0.779
m4_hourly	0.131	0.025 / 0.774	0.006	0.004 / 0.775	0.010	0.009 / 0.918	0.009	0.007 / 0.813	0.008	0.007 / 0.738
m4_monthly	0.930	0.254 / 0.610	0.283	0.180 / 0.783	0.288	0.198 / 0.822	0.295	0.191 / 0.778	0.317	0.199 / 0.733
m4_quarterly	0.794	0.272 / 0.763	0.392	0.262 / 0.824	0.396	0.267 / 0.842	0.416	0.278 / 0.811	0.437	0.275 / 0.725
m4_weekly	0.467	0.162 / 0.751	0.249	0.168 / 0.811	0.264	0.193 / 0.865	0.261	0.184 / 0.831	0.290	0.188 / 0.730
M_DENSE_D	0.464	0.166 / 0.775	0.357	0.179 / 0.788	0.358	0.171 / 0.761	0.328	0.169 / 0.778	0.354	0.191 / 0.814
M_DENSE_H	0.521	0.221 / 0.776	0.339	0.213 / 0.778	0.363	0.235 / 0.812	0.330	0.183 / 0.715	0.339	0.194 / 0.739
saugeenday_D	1.279	0.820 / 0.757	0.528	0.288 / 0.712	0.523	0.267 / 0.877	0.602	0.438 / 0.788	0.563	0.371 / 0.838
saugeenday_M	2.706	1.691 / 0.798	2.119	1.212 / 0.798	1.880	1.069 / 0.762	1.852	1.246 / 0.810	1.936	1.176 / 0.786
saugeenday_W	2.308	1.023 / 0.769	1.858	0.872 / 0.831	1.803	0.849 / 0.825	1.801	1.042 / 0.819	2.007	1.070 / 0.881
solar_10T	0.375	0.125 / 0.630	0.157	0.091 / 0.798	0.184	0.125 / 0.942	0.174	0.112 / 0.790	0.193	0.124 / 0.831
solar_H	0.434	0.132 / 0.802	0.226	0.160 / 0.849	0.373	0.320 / 0.961	0.351	0.314 / 0.912	0.326	0.202 / 0.856
SZ_TAXI_15T	1.152	0.627 / 0.784	0.900	0.576 / 0.780	0.905	0.544 / 0.786	0.899	0.574 / 0.766	0.903	0.592 / 0.804
SZ_TAXI_H	0.785	0.433 / 0.768	0.638	0.425 / 0.790	0.648	0.389 / 0.753	0.641	0.411 / 0.760	0.651	0.423 / 0.779
temperature_rain	3.633	0.718 / 0.587	3.238	1.282 / 0.685	3.482	1.171 / 0.706	3.229	1.345 / 0.665	3.518	1.260 / 0.639
us_births_D	1.905	1.451 / 0.778	1.276	0.740 / 0.822	0.765	0.561 / 0.852	0.953	0.502 / 0.785	1.132	0.641 / 0.838
us_births_W	1.293	0.908 / 0.732	0.955	0.579 / 0.777	1.042	0.524 / 0.714	0.882	0.508 / 0.750	0.898	0.297 / 0.536
Average	1.021	0.473 / 0.741	0.708	0.384 / 0.789	0.697	0.370 / 0.825	0.685	0.395 / 0.784	0.729	0.382 / 0.772

Native TSFM UQ (2/2)

	PatchTSTFM		TabPFN-TS3		TimesFM2.5		Toto2		TTM	
Dataset	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑
electricity_15T	0.026	0.018 / 0.765	0.030	0.023 / 0.833	0.029	0.022 / 0.830	0.024	0.017 / 0.806	0.039	0.032 / 0.906
electricity_D	0.040	0.025 / 0.753	0.052	0.039 / 0.858	0.040	0.030 / 0.900	0.038	0.024 / 0.819	0.049	0.043 / 0.920
electricity_H	0.027	0.017 / 0.776	0.034	0.025 / 0.815	0.028	0.021 / 0.863	0.024	0.017 / 0.814	0.037	0.033 / 0.898
LOOP_SEATTLE_5T	0.900	0.561 / 0.754	0.976	0.633 / 0.776	0.930	0.584 / 0.785	0.890	0.573 / 0.773	1.023	0.704 / 0.817
LOOP_SEATTLE_H	1.042	0.576 / 0.708	1.813	0.905 / 0.734	0.998	0.626 / 0.846	0.850	0.495 / 0.767	1.561	0.888 / 0.854
m4_daily	0.083	0.058 / 0.831	0.141	0.089 / 0.807	0.076	0.056 / 0.866	0.071	0.046 / 0.820	0.180	0.171 / 0.986
m4_hourly	0.008	0.006 / 0.775	0.013	0.012 / 0.918	0.009	0.008 / 0.896	0.006	0.004 / 0.815	0.029	0.027 / 0.952
m4_monthly	0.307	0.192 / 0.767	0.369	0.274 / 0.815	0.293	0.210 / 0.834	0.271	0.179 / 0.781	0.489	0.419 / 0.950
m4_quarterly	0.431	0.266 / 0.787	0.479	0.331 / 0.775	0.423	0.299 / 0.847	0.385	0.258 / 0.826	1.031	0.994 / 0.959
m4_weekly	0.275	0.186 / 0.800	0.332	0.247 / 0.835	0.256	0.187 / 0.859	0.240	0.167 / 0.826	0.383	0.351 / 0.941
M_DENSE_D	0.334	0.157 / 0.740	0.361	0.170 / 0.770	0.339	0.170 / 0.814	0.331	0.153 / 0.752	0.400	0.283 / 0.894
M_DENSE_H	0.330	0.190 / 0.732	0.536	0.303 / 0.782	0.351	0.228 / 0.828	0.320	0.204 / 0.788	0.447	0.317 / 0.859
saugeenday_D	0.584	0.379 / 0.775	0.950	0.612 / 0.838	0.547	0.351 / 0.883	0.489	0.255 / 0.752	0.965	0.863 / 0.958
saugeenday_M	1.948	1.138 / 0.738	1.845	1.437 / 0.869	2.069	1.189 / 0.798	1.640	1.205 / 0.798	1.955	1.412 / 0.857
saugeenday_W	1.865	1.016 / 0.856	1.903	1.050 / 0.844	2.079	0.828 / 0.812	1.813	0.870 / 0.850	2.129	1.412 / 0.944
solar_10T	0.182	0.114 / 0.669	0.228	0.171 / 0.927	0.205	0.133 / 0.913	0.154	0.085 / 0.896	0.491	0.398 / 0.911
solar_H	0.384	0.360 / 0.879	0.618	0.573 / 0.926	0.414	0.372 / 0.921	0.270	0.177 / 0.871	0.549	0.499 / 0.968
SZ_TAXI_15T	0.900	0.585 / 0.782	0.923	0.631 / 0.814	0.915	0.549 / 0.788	0.881	0.577 / 0.777	0.970	0.667 / 0.839
SZ_TAXI_H	0.641	0.410 / 0.758	0.663	0.442 / 0.784	0.640	0.445 / 0.815	0.633	0.416 / 0.780	0.696	0.557 / 0.878
temperature_rain	3.383	1.286 / 0.531	4.373	1.065 / 0.662	3.767	1.049 / 0.684	3.111	1.382 / 0.757	3.478	1.646 / 0.730
us_births_D	1.063	0.583 / 0.827	1.189	0.612 / 0.817	1.110	0.669 / 0.873	0.983	0.607 / 0.845	1.423	0.994 / 0.908
us_births_W	1.153	0.516 / 0.670	1.137	0.781 / 0.839	1.091	0.674 / 0.830	0.748	0.526 / 0.812	1.128	0.694 / 0.768
Average	0.723	0.393 / 0.758	0.862	0.474 / 0.820	0.755	0.396 / 0.840	0.644	0.375 / 0.806	0.884	0.609 / 0.895

RareCP (Ours)

	TiRex		TimesFM2.5		Toto2		ARIMA	
Dataset	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑	nWink ↓	nW ↓ / Cov ↑
electricity_15T	0.034	0.021 / 0.800	0.036	0.022 / 0.800	0.031	0.017 / 0.800	0.053	0.028 / 0.800
electricity_D	0.055	0.035 / 0.799	0.054	0.035 / 0.800	0.054	0.035 / 0.800	0.059	0.038 / 0.799
electricity_H	0.034	0.021 / 0.799	0.034	0.023 / 0.800	0.032	0.020 / 0.800	0.055	0.039 / 0.800
LOOP_SEATTLE_5T	**0.923**	0.636 / 0.800	0.965	0.645 / 0.800	0.969	0.666 / 0.799	1.045	0.722 / 0.800
LOOP_SEATTLE_H	**0.745**	0.491 / 0.800	0.957	0.578 / 0.800	0.934	0.567 / 0.800	1.079	0.706 / 0.800
m4_daily	**0.067**	0.041 / 0.800	0.074	0.043 / 0.800	0.075	0.046 / 0.800	0.078	0.048 / 0.800
m4_hourly	**0.006**	0.004 / 0.796	0.006	0.004 / 0.801	0.006	0.004 / 0.800	0.010	0.006 / 0.800
m4_monthly	**0.286**	0.187 / 0.795	0.331	0.202 / 0.800	0.334	0.194 / 0.793	0.400	0.245 / 0.798
m4_quarterly	**0.417**	0.267 / 0.797	0.467	0.263 / 0.800	0.485	0.284 / 0.796	0.537	0.314 / 0.799
m4_weekly	**0.255**	0.165 / 0.800	0.333	0.211 / 0.790	0.341	0.222 / 0.790	0.388	0.259 / 0.796
M_DENSE_D	**0.299**	0.199 / 0.798	0.398	0.203 / 0.796	0.436	0.224 / 0.795	0.489	0.271 / 0.795
M_DENSE_H	0.369	0.247 / 0.798	0.405	0.240 / 0.797	0.388	0.238 / 0.798	0.487	0.344 / 0.799
saugeenday_D	0.547	0.300 / 0.793	**0.534**	0.233 / 0.803	0.536	0.252 / 0.795	0.565	0.301 / 0.803
saugeenday_M	**2.071**	1.266 / 0.728	2.359	0.693 / 0.639	2.142	1.118 / 0.728	2.105	1.142 / 0.817
saugeenday_W	2.124	0.918 / 0.781	2.217	0.546 / 0.715	1.980	0.745 / 0.774	2.082	0.849 / 0.773
solar_10T	**0.167**	0.077 / 0.799	0.186	0.091 / 0.799	0.173	0.082 / 0.800	0.213	0.103 / 0.798
solar_H	**0.223**	0.141 / 0.800	0.306	0.198 / 0.800	0.289	0.171 / 0.800	0.368	0.225 / 0.799
SZ_TAXI_15T	**0.935**	0.639 / 0.800	0.945	0.632 / 0.801	0.952	0.638 / 0.800	1.026	0.734 / 0.800
SZ_TAXI_H	0.672	0.467 / 0.800	0.682	0.475 / 0.799	0.698	0.495 / 0.799	0.781	0.582 / 0.799
temperature_rain	**1.517**	1.063 / 0.799	2.724	1.732 / 0.799	2.842	1.769 / 0.787	2.732	1.654 / 0.799
us_births_D	**0.857**	0.601 / 0.787	0.891	0.494 / 0.775	1.036	0.505 / 0.780	1.243	0.865 / 0.776
us_births_W	0.723	0.515 / 0.795	0.769	0.493 / 0.729	**0.689**	0.483 / 0.824	0.991	0.743 / 0.816
Average	**0.606**	0.377 / 0.794	0.712	0.366 / 0.784	0.701	0.399 / 0.793	0.763	0.464 / 0.799

C.4 HYPERPARAMETER SETTINGS

C.4.1 RARECP

Table 6: RareCP hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Each backbone uses one configuration across all datasets and evaluation seeds. The values were selected by Optuna TPE with 100 trials. Continuous values are rounded to three decimals.

Hyperparameter	**Search range**	**TiREx**	**TimesFM**
Student and retrieval			
Student hidden width	32:16:128	64	64
Student hidden layers	1:1:4	2	4
Student activation	{softsign, tanh}	tanh	tanh
Student dropout	0:0.025:0.25	0.025	0.000
Target penalty λ	$[1, 20]_{\log}$	1.086	3.753
Student epochs	100:20:240	180	120
Student learning rate	$[5\times10^{-5}, 10^{-2}]_{\log}$	2.011×10^{-4}	2.906×10^{-3}
Retrieval size k	8:8:96	48	40
Hopfield scale β	$[0.5, 24]$	9.082	14.822
Initial temperature	$[0.03, 0.12]_{\log}$	0.095	0.053
Final temperature	$[5\times10^{-5}, 5\times10^{-4}]_{\log}$	3.668×10^{-4}	3.112×10^{-4}
Penalty temperature	$[2\times10^{-4}, 10^{-3}]_{\log}$	4.407×10^{-4}	5.955×10^{-4}
Training batch size	{128, 256, 512, 768, 1024}	128	1024
Gradient clip norm	$[0.1, 2]_{\log}$	0.666	0.291
Mixture gate and online calibration			
Gate hidden width	{1, 2, 4, 8, 16, 32, 64}	1	2
Gate activation	{tanh, softsign}	softsign	softsign
Gate temperature	$[0.75, 4]$	0.878	2.438
Gate epochs	10:5:100	15	25
Gate learning rate	$[5\times10^{-4}, 8\times10^{-2}]_{\log}$	2.546×10^{-3}	8.261×10^{-4}
Gate entropy penalty	$[0, 0.05]$	0.021	0.029
ACI step size γ	$[10^{-3}, 10^{-1}]_{\log}$	0.019	0.008
Other backbone settings			
Student distillation	fixed	enabled	disabled
ACI warm start	fixed	disabled	enabled
Calibration pool	fixed	per dataset	all datasets
Fixed across datasets			
Teacher bank	10 experts with latent dimension 65		
Student encoder	linear hypernetwork		
Retrieval similarity	cosine		
Online calibration	ACI with per dataset updates and rolling residual insertion		
Target miscoverage α	0.2		

Table 7: RareCP hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Each backbone uses one configuration across all datasets and evaluation seeds. The values were selected by Optuna TPE with 100 trials. Continuous values are rounded to three decimals.

Hyperparameter	**Search range**	**Toto**	**ARIMA**
Student and retrieval			
Student hidden width	32:16:128	128	48
Student hidden layers	1:1:4	3	3
Student activation	{softsign, tanh}	softsign	tanh
Student dropout	0:0.025:0.25	0.225	0.200
Target penalty λ	$[1, 20]_{\log}$	1.325	3.628
Student epochs	100:20:240	200	220
Student learning rate	$[5\times10^{-5}, 10^{-2}]_{\log}$	4.469×10^{-3}	4.513×10^{-3}
Retrieval size k	8:8:96	88	80
Hopfield scale β	$[0.5, 24]$	6.287	6.810
Initial temperature	$[0.03, 0.12]_{\log}$	0.053	0.035
Final temperature	$[5\times10^{-5}, 5\times10^{-4}]_{\log}$	1.187×10^{-4}	2.213×10^{-4}
Penalty temperature	$[2\times10^{-4}, 10^{-3}]_{\log}$	2.218×10^{-4}	5.112×10^{-4}
Training batch size	{128, 256, 512, 768, 1024}	128	256
Gradient clip norm	$[0.1, 2]_{\log}$	0.518	0.400
Mixture gate and online calibration			
Gate hidden width	{1, 2, 4, 8, 16, 32, 64}	4	2
Gate activation	{tanh, softsign}	softsign	softsign
Gate temperature	$[0.75, 4]$	2.593	3.485
Gate epochs	10:5:100	90	60
Gate learning rate	$[5\times10^{-4}, 8\times10^{-2}]_{\log}$	3.011×10^{-3}	9.396×10^{-3}
Gate entropy penalty	$[0, 0.05]$	0.028	0.023
ACI step size γ	$[10^{-3}, 10^{-1}]_{\log}$	0.016	0.006
Other backbone settings			
Student distillation	fixed	enabled	enabled
ACI warm start	fixed	disabled	disabled
Calibration pool	fixed	all datasets	all datasets
Fixed across datasets			
Teacher bank	10 experts with latent dimension 65		
Student encoder	linear hypernetwork		
Retrieval similarity	cosine		
Online calibration	ACI with per dataset updates and rolling residual insertion		
Target miscoverage α	0.2		

C.5 RARECP DCA-ABLATIONS

Table 8: Hyperparameter search spaces for the architecture ablation. We use $\mathcal{A} = \{0.20\} \cup \{0.20 \pm 0.02m\} : m = 1, \ldots, 5\}$ for the searched alpha sets.

Group	Parameter	Search space
Shared	Activation	$\{\mathrm{ReLU}, \tanh, \mathrm{Softsign}, \mathrm{LipSwish}\}$
Shared	Top-k	$\{8, 16, 24, 32, 48, 64, 96\}$
Shared	Batch size	$\{128, 256, 384, 512, 768, 1024, 2048, 4096\}$
Shared	Training epochs	$\{50, 100, 200, 300, 500\}$
Shared	Hopfield inverse temperature β	$\log \mathcal{U}(0.5, 24)$
Shared	Learning rate	$\log \mathcal{U}(2{\times}10^{-5}, 5{\times}10^{-3})$
Shared	Start temperature τ_{start}	$\log \mathcal{U}(0.005, 0.20)$
Shared	End temperature τ_{end}	$\log \mathcal{U}(2{\times}10^{-5}, 5{\times}10^{-3})$
Shared	Penalty temperature τ_{pen}	$\log \mathcal{U}(5{\times}10^{-5}, 3{\times}10^{-3})$
Shared	Alpha set	$\mathcal{A}$
Hypernetwork	Latent dimension d	$\{16, 24, 32, 48, 64, 96, 128\}$
Hypernetwork	Hidden dimension	$\{32, 64, 96, 128, 192\}$
Hypernetwork	Hidden layers	$\{1, 2, 3, 4\}$
MLP-N	Depth	$N \in \{2, 3, 4, 5\}$ with hidden layers $= N - 1$
MLP-N	Latent dimension d	$\{16, 24, 32, 48, 64, 96, 128, 192, 884\}$
MLP-N	Hidden dimension	d

C.5.1 UNIFORM BASELINE

Table 9: Uniform hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. The search range appears in the second header row.

Dataset	**TiREx**	**TimesFM**
	W	W
Search range	100:100:2000	100:100:2000
electricity_15T	400	400
electricity_D	300	300
electricity_H	100	100
LOOP_SEATTLE_5T	100	100
LOOP_SEATTLE_H	100	100
m4_daily	100	100
m4_hourly	200	200
m4_monthly	100	100
m4_quarterly	100	100
m4_weekly	100	100
M_DENSE_D	100	100
M_DENSE_H	100	100
saugeenday_D	400	400
saugeenday_M	400	400
saugeenday_W	700	700
solar_10T	100	100
solar_H	200	300
SZ_TAXI_15T	100	100
SZ_TAXI_H	100	400
temperature_rain	200	200
us_births_D	400	400
us_births_W	100	100
Fixed across datasets		
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$	
Residual weighting	uniform over the rolling window	
Interval form	symmetric empirical residual quantile around $\hat{\mu}(x)$	
Online update	append the observed residual after each prediction	
Target miscoverage α	0.2	

Table 10: Uniform hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. The search range appears in the second header row.

Dataset	**Toto**	**ARIMA**
	W	W
Search range	100:100:2000	100:100:2000
electricity_15T	400	400
electricity_D	300	100
electricity_H	100	100
LOOP_SEATTLE_5T	100	100
LOOP_SEATTLE_H	100	100
m4_daily	100	100
m4_hourly	200	200
m4_monthly	100	100
m4_quarterly	100	100
m4_weekly	100	100
M_DENSE_D	100	100
M_DENSE_H	100	100
saugeenday_D	400	400
saugeenday_M	1600	1200
saugeenday_W	600	700
solar_10T	300	100
solar_H	200	300
SZ_TAXI_15T	200	100
SZ_TAXI_H	400	400
temperature_rain	200	200
us_births_D	400	400
us_births_W	100	600
Fixed across datasets		
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$	
Residual weighting	uniform over the rolling window	
Interval form	symmetric empirical residual quantile around $\hat{\mu}(x)$	
Online update	append the observed residual after each prediction	
Target miscoverage α	0.2	

C.5.2 ACI BASELINE

Table 11: ACI hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals, with smaller values shown to four decimal places.

Dataset	**TiREx**		**TimesFM**	
	γ	W	γ	W
Search range	$[10^{-4}, .1]_{\log}$	100:100:2000	$[10^{-4}, .1]_{\log}$	100:100:2000
electricity_15T	0.085	100	0.071	100
electricity_D	0.099	100	0.068	200
electricity_H	0.086	300	0.082	200
LOOP_SEATTLE_5T	0.023	2000	0.033	900
LOOP_SEATTLE_H	0.016	900	0.010	1300
m4_daily	0.071	200	0.079	200
m4_hourly	0.097	100	0.098	100
m4_monthly	0.097	100	0.097	100
m4_quarterly	0.070	200	0.094	100
m4_weekly	0.074	100	0.068	100
M_DENSE_D	0.034	700	0.059	300
M_DENSE_H	0.037	800	0.039	800
saugeenday_D	0.096	100	0.098	100
saugeenday_M	0.077	600	0.0004	1900
saugeenday_W	0.100	500	0.049	1900
solar_10T	0.056	500	0.059	300
solar_H	0.022	1900	0.026	800
SZ_TAXI_15T	0.052	1300	0.058	1200
SZ_TAXI_H	0.038	900	0.043	1500
temperature_rain	0.095	200	0.081	200
us_births_D	0.071	1200	0.081	1800
us_births_W	0.009	1100	0.020	400
Fixed across datasets				
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$			
Interval form	symmetric empirical residual quantile around $\hat{\mu}(x)$			
Online update	update the residual window and adaptive miscoverage after each observation			
Target miscoverage α	0.2			

Table 12: ACI hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals, with smaller values shown to four decimal places.

Dataset	**Toto**		**ARIMA**	
	γ	W	γ	W
Search range	$[10^{-4}, .1]_{\log}$	100:100:2000	$[10^{-4}, .1]_{\log}$	100:100:2000
electricity_15T	0.082	100	0.087	200
electricity_D	0.065	300	0.077	200
electricity_H	0.081	400	0.087	200
LOOP_SEATTLE_5T	0.018	1700	0.022	2000
LOOP_SEATTLE_H	0.013	1700	0.016	700
m4_daily	0.099	200	0.094	200
m4_hourly	0.089	100	0.026	500
m4_monthly	0.079	100	0.096	100
m4_quarterly	0.066	200	0.100	100
m4_weekly	0.046	900	0.075	100
M_DENSE_D	0.026	1400	0.084	200
M_DENSE_H	0.036	400	0.054	100
saugeenday_D	0.100	100	0.099	100
saugeenday_M	0.054	1300	0.0001	1000
saugeenday_W	0.077	1800	0.064	1000
solar_10T	0.099	100	0.076	200
solar_H	0.014	1600	0.023	2000
SZ_TAXI_15T	0.060	1900	0.097	200
SZ_TAXI_H	0.037	1400	0.034	1100
temperature_rain	0.084	200	0.099	400
us_births_D	0.046	1000	0.002	1100
us_births_W	0.038	1800	0.080	1500
Fixed across datasets				
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$			
Interval form	symmetric empirical residual quantile around $\hat{\mu}(x)$			
Online update	update the residual window and adaptive miscoverage after each observation			
Target miscoverage α	0.2			

C.5.3 DTACI BASELINE

Table 13: DtACI hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals, with smaller values shown to four decimal places.

Dataset	**TiREx**		**TimesFM**	
	η	W	η	W
Search range	$[5\times10^{-5}, .1]_{\log}$	100:100:2000	$[5\times10^{-5}, .1]_{\log}$	100:100:2000
electricity_15T	0.0001	600	0.0001	600
electricity_D	0.0001	400	0.0001	400
electricity_H	0.001	400	0.0001	500
LOOP_SEATTLE_5T	0.002	1800	0.0003	1500
LOOP_SEATTLE_H	0.004	2000	0.009	2000
m4_daily	0.0001	600	0.0001	600
m4_hourly	0.043	1300	0.0001	1800
m4_monthly	0.0001	300	0.0001	200
m4_quarterly	0.002	800	0.002	1000
m4_weekly	0.0001	400	0.0001	400
M_DENSE_D	0.0001	100	0.015	1000
M_DENSE_H	0.0001	200	0.0001	100
saugeenday_D	0.001	400	0.002	800
saugeenday_M	0.051	600	0.058	500
saugeenday_W	0.0001	700	0.0001	700
solar_10T	0.0002	700	0.0001	400
solar_H	0.002	1300	0.0008	500
SZ_TAXI_15T	0.0001	1200	0.0001	1200
SZ_TAXI_H	0.0001	1300	0.0001	800
temperature_rain	0.007	1600	0.0001	200
us_births_D	0.0001	200	0.0001	600
us_births_W	0.095	1600	0.042	200
Fixed across datasets				
Nonconformity score	$\lvert y - \hat{\mu}(x)\rvert$			
Candidate step sizes	$\{0.001, 0.002, 0.004, \ldots, 0.128\}$			
Weight scale	$\sigma = 1.0$			
Interval form	weighted combination of candidate residual quantiles			
Online update	update candidate miscoverage levels and weights after each observation			
Target miscoverage α	0.2			

Table 14: DtACI hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals, with smaller values shown to four decimal places.

Dataset	**Toto**		**ARIMA**	
	η	W	η	W
Search range	$[5\times10^{-5}, .1]_{\log}$	100:100:2000	$[5\times10^{-5}, .1]_{\log}$	100:100:2000
electricity_15T	0.0001	400	0.002	600
electricity_D	0.0001	600	0.0001	400
electricity_H	0.0001	600	0.0001	600
LOOP_SEATTLE_5T	0.003	800	0.001	1400
LOOP_SEATTLE_H	0.006	1900	0.0004	1600
m4_daily	0.0002	600	0.0009	500
m4_hourly	0.014	1000	0.003	1000
m4_monthly	0.0001	300	0.0005	300
m4_quarterly	0.0003	800	0.001	700
m4_weekly	0.0003	300	0.0001	300
M_DENSE_D	0.0007	100	0.0005	100
M_DENSE_H	0.001	200	0.0001	100
saugeenday_D	0.001	900	0.001	1200
saugeenday_M	0.0001	1500	0.009	100
saugeenday_W	0.0001	700	0.0001	700
solar_10T	0.0002	400	0.0002	700
solar_H	0.0001	800	0.0009	900
SZ_TAXI_15T	0.0001	1400	0.0001	800
SZ_TAXI_H	0.0001	1900	0.0001	1000
temperature_rain	0.0001	900	0.014	700
us_births_D	0.001	1100	0.011	1500
us_births_W	0.029	900	0.065	700
Fixed across datasets				
Nonconformity score	$\lvert y - \hat{\mu}(x)\rvert$			
Candidate step sizes	$\{0.001, 0.002, 0.004, \ldots, 0.128\}$			
Weight scale	$\sigma = 1.0$			
Interval form	weighted combination of candidate residual quantiles			
Online update	update candidate miscoverage levels and weights after each observation			
Target miscoverage α	0.2			

C.5.4 NExCP BASELINE

Table 15: NExCP hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals.

Dataset	**TiREx**		**TimesFM**	
	ρ	W	ρ	W
Search range	$[.9, .9999]_{\log}$	100:100:2000	$[.9, .9999]_{\log}$	100:100:2000
electricity_15T	0.900	1900	0.900	1500
electricity_D	0.900	1100	0.900	1200
electricity_H	0.900	800	0.900	2000
LOOP_SEATTLE_5T	0.952	500	0.957	2000
LOOP_SEATTLE_H	0.978	1200	0.976	600
m4_daily	0.900	1700	0.901	700
m4_hourly	0.902	1100	0.902	700
m4_monthly	0.900	900	0.900	900
m4_quarterly	0.900	300	0.900	100
m4_weekly	0.900	700	0.900	600
M_DENSE_D	0.902	1800	0.900	800
M_DENSE_H	0.907	300	0.937	1700
saugeenday_D	0.900	1300	0.901	1600
saugeenday_M	0.955	400	0.989	200
saugeenday_W	0.900	500	0.900	100
solar_10T	0.903	900	0.902	1000
solar_H	0.970	1500	0.900	700
SZ_TAXI_15T	0.900	300	0.900	1000
SZ_TAXI_H	0.929	1400	0.921	1500
temperature_rain	0.900	100	0.900	300
us_births_D	0.926	1600	0.961	500
us_births_W	0.940	900	0.902	900
Fixed across datasets				
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$			
Recency weighting	exponentially decaying weights determined by ρ			
Interval form	symmetric weighted residual quantile around $\hat{\mu}(x)$			
Online update	append the observed residual after each prediction			
Target miscoverage α	0.2			

Table 16: NExCP hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals.

Dataset	**Toto**		**ARIMA**	
	ρ	W	ρ	W
Search range	$[.9, .9999]_{\log}$	100:100:2000	$[.9, .9999]_{\log}$	100:100:2000
electricity_15T	0.900	2000	0.900	1600
electricity_D	0.900	1800	0.900	2000
electricity_H	0.901	100	0.900	1000
LOOP_SEATTLE_5T	0.954	900	0.959	1900
LOOP_SEATTLE_H	0.981	2000	0.970	1600
m4_daily	0.900	1600	0.900	1000
m4_hourly	0.900	1700	0.902	1400
m4_monthly	0.900	900	0.901	1200
m4_quarterly	0.901	100	0.900	2000
m4_weekly	0.900	300	0.900	500
M_DENSE_D	0.903	1700	0.900	1700
M_DENSE_H	0.929	1400	0.934	900
saugeenday_D	0.900	1900	0.901	2000
saugeenday_M	0.986	1300	0.992	1100
saugeenday_W	0.901	1900	0.928	1400
solar_10T	0.900	800	0.900	1200
solar_H	0.989	1300	0.970	1600
SZ_TAXI_15T	0.900	1500	0.900	800
SZ_TAXI_H	0.940	900	0.940	900
temperature_rain	0.900	100	0.900	100
us_births_D	0.987	1100	1.000	400
us_births_W	0.901	800	0.907	1500

Fixed across datasets	
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$
Recency weighting	exponentially decaying weights determined by ρ
Interval form	symmetric weighted residual quantile around $\hat{\mu}(x)$
Online update	append the observed residual after each prediction
Target miscoverage α	0.2

C.5.5 KOWCPI BASELINE

Table 17: KOWCPI hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals. Kernels are Epa for Epanechnikov and Gau for Gaussian. Bandwidth modes are Med for the median heuristic and AICc for corrected Akaike selection.

Dataset	**TiREx**						**TimesFM**					
	w	K	n_β	B	s_B	W	w	K	n_β	B	s_B	W
Search range	10:10:100	2	$\{50, 100, 150, 200\}$	2	$[.1, 3]_{\log}$	100:100:2000	10:10:100	2	$\{50, 100, 150, 200\}$	2	$[.1, 3]_{\log}$	100:100:2000
electricity_15T	50	Gau	200	Med	0.203	1800	10	Epa	150	Med	0.964	1200
electricity_D	10	Epa	200	Med	0.125	200	10	Epa	150	Med	0.906	1700
electricity_H	10	Epa	200	Med	0.149	1500	10	Epa	150	Med	0.114	1100
LOOP_SEATTLE_5T	70	Gau	150	AICc	0.704	1400	30	Gau	100	Med	1.351	1500
LOOP_SEATTLE_H	60	Gau	100	Med	0.207	400	50	Gau	150	AICc	0.870	300
m4_daily	10	Epa	100	Med	0.257	1500	10	Epa	200	Med	0.250	1600
m4_hourly	10	Gau	100	Med	0.193	1100	40	Gau	150	Med	0.101	1500
m4_monthly	80	Epa	200	Med	0.439	300	50	Epa	100	AICc	0.437	700
m4_quarterly	30	Epa	200	Med	0.252	600	10	Epa	50	Med	0.811	1500
m4_weekly	10	Gau	50	Med	0.200	300	10	Epa	50	Med	0.888	400
M_DENSE_D	30	Epa	200	Med	0.827	800	10	Gau	50	Med	0.200	700
M_DENSE_H	60	Epa	50	Med	2.994	300	10	Gau	100	Med	1.963	1900
saugeenday_D	10	Gau	150	Med	0.209	400	10	Gau	50	Med	0.219	700
saugeenday_M	30	Gau	200	Med	0.552	900	20	Gau	100	Med	0.413	1800
saugeenday_W	20	Epa	50	Med	0.934	1500	80	Gau	100	Med	0.240	300
solar_10T	10	Gau	100	Med	0.212	400	10	Gau	200	Med	0.196	800
solar_H	60	Gau	100	Med	0.167	1800	80	Gau	100	Med	0.175	1500
SZ_TAXI_15T	20	Epa	150	AICc	0.117	1700	20	Epa	200	AICc	0.124	700
SZ_TAXI_H	80	Epa	50	Med	0.164	1600	10	Gau	200	Med	1.318	200
temperature_rain	10	Gau	200	Med	0.189	700	10	Gau	50	AICc	0.931	1700
us_births_D	80	Gau	150	Med	0.204	400	30	Gau	200	Med	0.264	1800
us_births_W	100	Gau	200	Med	0.103	900	20	Gau	150	Med	1.311	700
Fixed across datasets												
Nonconformity score	signed residual $y - \hat{\mu}(x)$											
Kernel weighting	weights residual histories by their similarity to the current residual history											
Interval form	narrowest asymmetric weighted residual interval among n_β tail allocations											
Online update	append the observed signed residual after each prediction											
Target miscoverage α	0.2											

Table 18: KOWCPI hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals. Kernels are Epa for Epanechnikov and Gau for Gaussian. Bandwidth modes are Med for the median heuristic and AICc for corrected Akaike selection.

Dataset	**Toto**						**ARIMA**					
	w	K	n_β	B	s_B	W	w	K	n_β	B	s_B	W
Search range	10:10:100	2	$\{50, 100, 150, 200\}$	2	$[.1, 3]_{\log}$	100:100:2000	10:10:100	2	$\{50, 100, 150, 200\}$	2	$[.1, 3]_{\log}$	100:100:2000
electricity_15T	10	Gau	100	Med	0.232	1100	40	Epa	100	Med	0.951	2000
electricity_D	10	Epa	100	Med	0.911	1800	10	Gau	50	Med	0.212	300
electricity_H	10	Epa	200	Med	0.101	2000	10	Gau	200	Med	0.205	1200
LOOP_SEATTLE_5T	40	Epa	50	Med	2.991	1100	40	Gau	200	Med	0.192	1100
LOOP_SEATTLE_H	60	Gau	200	AICc	0.882	1900	30	Gau	150	Med	0.200	1400
m4_daily	10	Epa	50	Med	0.200	1100	10	Gau	50	Med	0.192	2000
m4_hourly	10	Epa	150	Med	0.830	1000	10	Gau	150	Med	0.210	1700
m4_monthly	10	Epa	50	AICc	0.254	100	50	Epa	150	AICc	0.322	500
m4_quarterly	10	Epa	200	AICc	0.192	1200	10	Epa	50	AICc	0.159	1200
m4_weekly	10	Gau	50	Med	0.208	400	10	Epa	100	AICc	0.286	1700
M_DENSE_D	30	Epa	200	Med	0.905	400	10	Epa	150	Med	0.664	500
M_DENSE_H	10	Gau	50	Med	0.186	400	10	Epa	150	Med	0.769	1400
saugeenday_D	20	Gau	100	Med	0.206	500	10	Epa	50	Med	0.926	1100
saugeenday_M	30	Gau	200	Med	0.383	800	30	Gau	100	AICc	0.871	1500
saugeenday_W	20	Gau	100	Med	0.301	700	10	Gau	50	Med	0.259	900
solar_10T	10	Gau	50	Med	2.229	500	90	Gau	150	AICc	0.463	1900
solar_H	60	Gau	100	Med	0.171	1800	40	Gau	100	Med	0.173	700
SZ_TAXI_15T	10	Epa	200	AICc	0.101	1100	10	Epa	100	AICc	0.101	1700
SZ_TAXI_H	10	Gau	50	Med	0.708	500	10	Gau	50	Med	1.351	1700
temperature_rain	10	Gau	150	Med	0.193	600	10	Epa	200	Med	0.730	400
us_births_D	10	Gau	50	Med	0.501	100	20	Gau	150	Med	0.227	1800
us_births_W	100	Gau	200	AICc	2.823	400	40	Gau	50	AICc	0.101	1600

Fixed across datasets

Nonconformity score	signed residual $y - \hat{\mu}(x)$
Kernel weighting	weights residual histories by their similarity to the current residual history
Interval form	narrowest asymmetric weighted residual interval among n_β tail allocations
Online update	append the observed signed residual after each prediction
Target miscoverage α	0.2

C.5.6 COP BASELINE

Table 19: COP hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals. Here η is the learning rate, T_b is the burn in length, and s is the integrator scale.

Dataset	**TiREx**			**TimesFM**		
	η	T_b	s	η	T_b	s
Search range	$[10^{-3}, 3]_{\log}$	20:20:500	$[.1, 2]$	$[10^{-3}, 3]_{\log}$	20:20:500	$[.1, 2]$
electricity_15T	0.297	20	0.383	0.357	20	0.409
electricity_D	0.267	20	0.225	0.292	20	0.486
electricity_H	0.354	20	0.589	0.256	20	0.653
LOOP_SEATTLE_5T	0.042	60	1.853	0.064	60	1.790
LOOP_SEATTLE_H	0.013	60	1.809	0.013	420	1.710
m4_daily	0.162	20	0.569	0.270	20	0.104
m4_hourly	0.031	500	0.490	0.029	280	0.486
m4_monthly	0.124	40	0.783	0.253	20	0.471
m4_quarterly	0.116	120	0.714	0.083	320	0.324
m4_weekly	0.065	300	0.393	0.067	420	0.642
M_DENSE_D	0.020	120	0.567	0.096	40	0.666
M_DENSE_H	0.038	380	0.207	0.114	40	0.211
saugeenday_D	0.273	20	0.391	0.367	20	0.574
saugeenday_M	0.051	60	0.722	0.064	460	0.859
saugeenday_W	0.043	380	0.674	0.080	120	0.373
solar_10T	0.107	60	0.997	0.070	40	0.755
solar_H	0.079	320	0.160	0.162	480	0.214
SZ_TAXI_15T	0.074	440	0.715	0.236	20	0.754
SZ_TAXI_H	0.059	300	0.791	0.054	340	1.549
temperature_rain	0.127	220	0.291	0.141	220	0.105
us_births_D	0.034	40	1.234	0.039	80	0.507
us_births_W	0.008	280	0.998	0.010	440	0.854
Fixed across datasets						
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$					
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$					
Learning rate scaling	proportional to the recent score range					
Online update	coverage feedback with integral correction after burn in					
Target miscoverage α	0.2					

Table 20: COP hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals. Here η is the learning rate, T_b is the burn in length, and s is the integrator scale.

Dataset	**Toto**			**ARIMA**		
	η	T_b	s	η	T_b	s
Search range	$[10^{-3}, 3]_{\log}$	20:20:500	$[.1, 2]$	$[10^{-3}, 3]_{\log}$	20:20:500	$[.1, 2]$
electricity_15T	0.398	20	0.234	0.408	20	0.440
electricity_D	0.263	20	0.284	0.339	20	0.102
electricity_H	0.407	20	0.299	0.348	20	0.668
LOOP_SEATTLE_5T	0.032	60	0.302	0.058	60	1.166
LOOP_SEATTLE_H	0.012	100	1.208	0.015	220	1.604
m4_daily	0.152	60	0.590	0.194	40	0.515
m4_hourly	0.050	260	0.534	0.020	460	0.434
m4_monthly	0.182	40	0.362	0.222	40	0.265
m4_quarterly	0.328	20	0.257	0.310	40	0.811
m4_weekly	0.163	40	0.683	0.403	20	0.378
M_DENSE_D	0.022	260	0.620	0.060	220	0.178
M_DENSE_H	0.055	220	0.678	0.031	400	0.407
saugeenday_D	0.159	20	0.258	0.322	20	0.374
saugeenday_M	0.112	360	1.506	0.024	340	0.405
saugeenday_W	0.156	100	0.301	0.038	500	0.860
solar_10T	0.193	20	0.501	0.173	20	0.248
solar_H	0.009	140	1.306	0.032	480	0.602
SZ_TAXI_15T	0.096	240	0.569	0.087	440	0.530
SZ_TAXI_H	0.039	380	0.987	0.047	380	1.003
temperature_rain	0.113	220	0.224	0.137	300	0.338
us_births_D	0.001	140	1.486	0.013	100	1.832
us_births_W	0.012	20	1.376	0.089	340	1.864

Fixed across datasets	
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$
Learning rate scaling	proportional to the recent score range
Online update	coverage feedback with integral correction after burn in
Target miscoverage α	0.2

C.5.7 DECAY-OGD BASELINE

Table 21: Decay OGD hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals.

Dataset	**TiREx**		**TimesFM**	
	η	ϵ	η	ϵ
Search range	$[10^{-3}, 10]_{\log}$	$[.01, .3]$	$[10^{-3}, 10]_{\log}$	$[.01, .3]$
electricity_15T	0.411	0.023	0.833	0.118
electricity_D	1.232	0.158	2.611	0.198
electricity_H	1.314	0.167	0.516	0.028
LOOP_SEATTLE_5T	3.575	0.038	5.869	0.048
LOOP_SEATTLE_H	8.188	0.182	7.119	0.102
m4_daily	1.128	0.039	0.883	0.035
m4_hourly	0.309	0.170	0.977	0.295
m4_monthly	7.288	0.137	3.330	0.036
m4_quarterly	8.099	0.026	5.720	0.011
m4_weekly	4.394	0.107	9.916	0.175
M_DENSE_D	4.358	0.144	2.286	0.053
M_DENSE_H	3.209	0.066	3.304	0.064
saugeenday_D	7.549	0.057	8.755	0.070
saugeenday_M	7.707	0.278	1.518	0.186
saugeenday_W	7.275	0.102	8.099	0.015
solar_10T	6.476	0.178	7.840	0.222
solar_H	5.843	0.063	9.881	0.011
SZ_TAXI_15T	9.940	0.033	9.921	0.011
SZ_TAXI_H	9.983	0.037	9.854	0.010
temperature_rain	7.566	0.011	9.962	0.015
us_births_D	0.055	0.286	4.863	0.275
us_births_W	0.062	0.284	0.062	0.243
Fixed across datasets				
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$			
Learning rate schedule	$\eta_t = \eta\, t^{-1/2-\epsilon}$			
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$			
Online update	gradient update from the most recent coverage error			
Target miscoverage α	0.2			

Table 22: Decay OGD hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals.

Dataset	**Toto**		**ARIMA**	
	η	ϵ	η	ϵ
Search range	$[10^{-3}, 10]_{\log}$	$[.01, .3]$	$[10^{-3}, 10]_{\log}$	$[.01, .3]$
electricity_15T	1.025	0.121	0.643	0.062
electricity_D	0.572	0.051	2.874	0.220
electricity_H	0.337	0.021	6.700	0.256
LOOP_SEATTLE_5T	6.084	0.087	9.288	0.085
LOOP_SEATTLE_H	9.729	0.196	3.618	0.045
m4_daily	1.485	0.070	1.382	0.053
m4_hourly	0.692	0.275	0.881	0.251
m4_monthly	9.079	0.154	3.891	0.056
m4_quarterly	8.009	0.025	6.409	0.011
m4_weekly	6.359	0.117	4.218	0.070
M_DENSE_D	1.623	0.036	9.766	0.130
M_DENSE_H	3.641	0.089	9.656	0.132
saugeenday_D	7.166	0.027	7.060	0.030
saugeenday_M	0.471	0.294	0.541	0.214
saugeenday_W	9.398	0.084	8.187	0.092
solar_10T	3.753	0.070	9.079	0.158
solar_H	9.029	0.028	6.307	0.065
SZ_TAXI_15T	9.806	0.011	9.860	0.017
SZ_TAXI_H	9.165	0.032	9.770	0.060
temperature_rain	8.515	0.012	9.622	0.012
us_births_D	0.037	0.200	0.109	0.202
us_births_W	0.034	0.125	0.155	0.277
Fixed across datasets				
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$			
Learning rate schedule	$\eta_t = \eta\, t^{-1/2-\epsilon}$			
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$			
Online update	gradient update from the most recent coverage error			
Target miscoverage α	0.2			

C.5.8 ECI BASELINE

Table 23: ECI hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals. Here η is the learning rate, c controls the error correction, and T_b is the burn in length.

Dataset	**TiREx**			**TimesFM**		
Search range	η $[10^{-3}, 3]_{\log}$	c $[.1, 50]_{\log}$	T_b 20:20:500	η $[10^{-3}, 3]_{\log}$	c $[.1, 50]_{\log}$	T_b 20:20:500
electricity_15T	0.201	1.400	20	0.297	0.394	20
electricity_D	0.326	0.389	20	0.290	0.656	20
electricity_H	0.080	34.721	160	0.298	2.183	20
LOOP_SEATTLE_5T	0.044	0.756	60	0.066	2.475	60
LOOP_SEATTLE_H	0.017	6.389	60	0.025	2.981	40
m4_daily	0.178	23.839	20	0.181	0.147	20
m4_hourly	0.034	49.562	280	0.016	0.890	500
m4_monthly	0.147	11.408	40	0.252	0.225	20
m4_quarterly	0.250	4.736	20	0.244	7.684	20
m4_weekly	0.176	3.200	20	0.197	2.466	20
M_DENSE_D	0.190	0.588	20	0.090	0.986	20
M_DENSE_H	0.076	5.473	60	0.036	2.528	400
saugeenday_D	0.168	1.519	20	0.215	16.032	20
saugeenday_M	0.150	2.187	400	0.313	0.550	220
saugeenday_W	0.088	1.408	160	0.074	0.565	440
solar_10T	0.139	4.915	20	0.057	31.139	80
solar_H	0.165	12.548	80	0.392	3.004	40
SZ_TAXI_15T	0.067	1.762	440	0.064	1.906	440
SZ_TAXI_H	0.055	2.869	440	0.052	0.514	280
temperature_rain	0.106	2.421	260	0.108	1.542	340
us_births_D	0.001	20.473	200	0.017	3.435	80
us_births_W	0.011	4.395	360	0.010	12.355	220

Fixed across datasets	
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$
Learning rate scaling	proportional to the recent score range
Error correction	smooth correction controlled by c
Online update	coverage gradient with error correction after each observation
Target miscoverage α	0.2

Table 24: ECI hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals. Here η is the learning rate, c controls the error correction, and T_b is the burn in length.

Dataset	**Toto**			**ARIMA**		
	η	c	T_b	η	c	T_b
Search range	$[10^{-3}, 3]_{\log}$	$[.1, 50]_{\log}$	20:20:500	$[10^{-3}, 3]_{\log}$	$[.1, 50]_{\log}$	20:20:500
electricity_15T	0.413	2.834	20	0.487	8.768	20
electricity_D	0.289	4.974	20	0.303	0.584	20
electricity_H	0.308	0.194	20	0.204	2.385	20
LOOP_SEATTLE_5T	0.042	1.622	60	0.057	5.042	40
LOOP_SEATTLE_H	0.012	8.369	60	0.016	1.781	140
m4_daily	0.205	5.164	20	0.044	0.335	460
m4_hourly	0.016	0.160	500	0.016	37.551	480
m4_monthly	0.183	2.767	20	0.176	2.236	40
m4_quarterly	0.249	2.414	20	0.125	4.756	80
m4_weekly	0.206	2.453	20	0.038	15.471	380
M_DENSE_D	0.037	2.568	60	0.114	1.866	20
M_DENSE_H	0.030	2.476	460	0.029	3.904	300
saugeenday_D	0.163	0.638	20	0.138	39.973	20
saugeenday_M	0.250	2.743	120	0.022	3.283	180
saugeenday_W	0.044	2.252	420	0.044	3.040	500
solar_10T	0.112	5.496	40	0.096	15.660	100
solar_H	0.094	8.940	320	0.035	2.901	500
SZ_TAXI_15T	0.082	3.048	120	0.072	2.610	360
SZ_TAXI_H	0.041	2.980	320	0.045	2.942	460
temperature_rain	0.091	0.816	400	0.108	0.477	500
us_births_D	0.008	2.186	80	0.013	2.688	100
us_births_W	0.010	14.476	20	0.059	10.700	260

Fixed across datasets	
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$
Learning rate scaling	proportional to the recent score range
Error correction	smooth correction controlled by c
Online update	coverage gradient with error correction after each observation
Target miscoverage α	0.2

C.5.9 PID BASELINE

Table 25: PID hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals. Here η is the learning rate, K_I is the integral gain, and T_b is the burn in length.

Dataset	**TiREx**			**TimesFM**		
Search range	η $[10^{-5}, 3]_{\log}$	K_I $[10^{-3}, 100]_{\log}$	T_b 20:20:500	η $[10^{-5}, 3]_{\log}$	K_I $[10^{-3}, 100]_{\log}$	T_b 20:20:500
electricity_15T	0.247	0.131	20	0.317	0.016	20
electricity_D	0.180	0.022	40	0.297	0.609	20
electricity_H	0.386	0.015	20	0.193	0.950	40
LOOP_SEATTLE_5T	0.039	0.415	80	0.037	1.183	60
LOOP_SEATTLE_H	0.008	2.136	220	0.012	0.028	240
m4_daily	0.248	0.243	20	0.112	0.001	100
m4_hourly	0.012	0.015	500	0.016	0.046	500
m4_monthly	0.132	0.445	40	0.251	0.178	20
m4_quarterly	0.244	0.041	20	0.152	1.252	40
m4_weekly	0.167	0.026	20	0.066	0.015	120
M_DENSE_D	0.020	0.004	120	0.098	0.370	20
M_DENSE_H	0.121	0.403	20	0.160	0.084	20
saugeenday_D	0.246	0.862	20	0.140	0.808	20
saugeenday_M	0.050	0.021	300	0.127	0.239	460
saugeenday_W	0.042	0.007	380	0.066	0.482	340
solar_10T	0.133	0.004	20	0.137	1.413	20
solar_H	0.115	0.003	160	0.146	0.519	440
SZ_TAXI_15T	0.136	0.020	60	0.166	7.473	60
SZ_TAXI_H	0.059	0.251	340	0.040	0.007	320
temperature_rain	0.098	0.778	440	0.114	0.003	300
us_births_D	0.074	0.003	40	0.017	0.004	80
us_births_W	0.013	0.002	460	0.014	0.001	140

Fixed across datasets	
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$
Learning rate scaling	proportional to the recent score range
Saturation constant	$C_{\text{sat}} = 1$
Online update	coverage gradient with logarithmic integral correction
Target miscoverage α	0.2

Table 26: PID hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals. Here η is the learning rate, K_I is the integral gain, and T_b is the burn in length.

Dataset	**Toto**			**ARIMA**		
Search range	η $[10^{-5}, 3]_{\log}$	K_I $[10^{-3}, 100]_{\log}$	T_b 20:20:500	η $[10^{-5}, 3]_{\log}$	K_I $[10^{-3}, 100]_{\log}$	T_b 20:20:500
electricity_15T	0.368	0.002	20	0.318	2.549	20
electricity_D	0.252	0.003	20	0.311	0.503	20
electricity_H	0.285	0.065	20	0.331	0.062	20
LOOP_SEATTLE_5T	0.030	0.596	60	0.032	0.006	260
LOOP_SEATTLE_H	0.007	0.053	400	0.014	0.487	140
m4_daily	0.095	0.005	100	0.209	0.009	20
m4_hourly	0.221	0.035	20	0.019	0.002	500
m4_monthly	0.091	0.499	80	0.043	5.389	260
m4_quarterly	0.218	0.119	20	0.262	1.804	20
m4_weekly	0.034	1.099	400	0.265	1.739	20
M_DENSE_D	0.021	0.032	160	0.054	0.259	160
M_DENSE_H	0.048	0.014	80	0.033	0.101	400
saugeenday_D	0.028	6.021	460	0.223	0.875	20
saugeenday_M	0.698	0.024	120	0.019	0.003	80
saugeenday_W	0.053	0.224	420	0.037	0.020	360
solar_10T	0.151	0.037	20	0.236	1.013	20
solar_H	0.009	0.015	220	0.031	0.376	320
SZ_TAXI_15T	0.185	0.002	40	0.061	0.241	500
SZ_TAXI_H	0.036	0.085	360	0.090	0.003	20
temperature_rain	0.107	24.770	220	0.143	12.567	220
us_births_D	0.001	0.071	160	0.011	0.028	220
us_births_W	0.007	0.001	360	0.070	0.018	40

Fixed across datasets	
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$
Learning rate scaling	proportional to the recent score range
Saturation constant	$C_{\text{sat}} = 1$
Online update	coverage gradient with logarithmic integral correction
Target miscoverage α	0.2

C.5.10 SAOCP BASELINE

Table 27: SAOCP hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. The search range appears in the second header row.

Dataset	**TiREx**	**TimesFM**
	G	G
Search range	$\{2, 4, 8, 16, 32, 64\}$	$\{2, 4, 8, 16, 32, 64\}$
electricity_15T	64	64
electricity_D	64	64
electricity_H	64	64
LOOP_SEATTLE_5T	64	64
LOOP_SEATTLE_H	64	64
m4_daily	64	64
m4_hourly	64	64
m4_monthly	64	64
m4_quarterly	64	64
m4_weekly	64	64
M_DENSE_D	64	64
M_DENSE_H	64	64
saugeenday_D	64	64
saugeenday_M	32	64
saugeenday_W	64	64
solar_10T	64	64
solar_H	64	64
SZ_TAXI_15T	64	64
SZ_TAXI_H	64	64
temperature_rain	64	64
us_births_D	64	64
us_births_W	32	32
Fixed across datasets		
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$	
Expert scale	$\sqrt{3}$ times the largest calibration score	
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$	
Online update	weighted ensemble of quantile experts with multiple lifetimes	
Target miscoverage α	0.2	

Table 28: SAOCP hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. The search range appears in the second header row.

Dataset	**Toto**	**ARIMA**
	G	G
Search range	$\{2, 4, 8, 16, 32, 64\}$	$\{2, 4, 8, 16, 32, 64\}$
electricity_15T	64	64
electricity_D	64	64
electricity_H	64	64
LOOP_SEATTLE_5T	64	64
LOOP_SEATTLE_H	64	64
m4_daily	64	64
m4_hourly	64	64
m4_monthly	64	64
m4_quarterly	64	64
m4_weekly	64	64
M_DENSE_D	64	64
M_DENSE_H	64	64
saugeenday_D	64	64
saugeenday_M	4	32
saugeenday_W	64	64
solar_10T	64	64
solar_H	64	64
SZ_TAXI_15T	64	64
SZ_TAXI_H	64	64
temperature_rain	64	64
us_births_D	64	64
us_births_W	64	32
Fixed across datasets		
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$	
Expert scale	$\sqrt{3}$ times the largest calibration score	
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$	
Online update	weighted ensemble of quantile experts with multiple lifetimes	
Target miscoverage α	0.2	

C.5.11 SF-OGD BASELINE

Table 29: SF OGD hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. The search range appears in the second header row. Continuous values are rounded to three decimals.

Dataset	**TiREx**	**TimesFM**
	η	η
Search range	$[10^{-4}, 3]_{\log}$	$[10^{-4}, 3]_{\log}$
electricity_15T	0.210	0.176
electricity_D	0.215	0.286
electricity_H	0.189	0.225
LOOP_SEATTLE_5T	1.672	2.614
LOOP_SEATTLE_H	0.957	1.196
m4_daily	0.453	0.350
m4_hourly	0.035	0.042
m4_monthly	1.057	1.424
m4_quarterly	2.990	2.998
m4_weekly	1.249	1.154
M_DENSE_D	0.839	0.823
M_DENSE_H	1.120	1.378
saugeenday_D	2.468	2.737
saugeenday_M	1.381	0.616
saugeenday_W	2.978	2.288
solar_10T	0.675	0.761
solar_H	1.743	2.995
SZ_TAXI_15T	2.961	2.968
SZ_TAXI_H	2.926	2.883
temperature_rain	2.995	2.999
us_births_D	0.016	0.294
us_births_W	0.040	0.035
Fixed across datasets		
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$	
Learning rate normalization	$\eta_t = \eta / \sqrt{\sum_{i=1}^{t} g_i^2}$	
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$	
Online update	normalized gradient update from the latest coverage error	
Target miscoverage α	0.2	

Table 30: SF OGD hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. The search range appears in the second header row. Continuous values are rounded to three decimals.

Dataset	**Toto**	**ARIMA**
	η	η
Search range	$[10^{-4}, 3]_{\log}$	$[10^{-4}, 3]_{\log}$
electricity_15T	0.191	0.221
electricity_D	0.204	0.251
electricity_H	0.192	0.446
LOOP_SEATTLE_5T	1.653	1.685
LOOP_SEATTLE_H	0.943	1.381
m4_daily	0.461	0.497
m4_hourly	0.037	0.056
m4_monthly	1.295	1.509
m4_quarterly	2.869	2.988
m4_weekly	1.265	1.369
M_DENSE_D	0.797	2.006
M_DENSE_H	0.902	1.752
saugeenday_D	2.913	3.000
saugeenday_M	0.610	0.199
saugeenday_W	2.656	2.775
solar_10T	1.221	1.773
solar_H	2.724	2.859
SZ_TAXI_15T	2.888	2.969
SZ_TAXI_H	2.949	2.591
temperature_rain	2.993	2.949
us_births_D	0.014	0.047
us_births_W	0.033	0.358
Fixed across datasets		
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$	
Learning rate normalization	$\eta_t = \eta / \sqrt{\sum_{i=1}^{t} g_i^2}$	
Interval form	symmetric adaptive interval around $\hat{\mu}(x)$	
Online update	normalized gradient update from the latest coverage error	
Target miscoverage α	0.2	

C.5.12 HopCPT baseline

Table 31: HopCPT hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. One configuration is selected for each backbone using an Optuna TPE study with 100 trials and is shared across datasets and evaluation seeds. Past plus forecast denotes univariate past context together with the backbone forecast.

Hyperparameter	**Search range**	**TiREx**	**TimesFM**
Hopfield scale β	$[.5, 24]_{\log}$	0.860	2.131
Batch size	$\{64, 128, 256, 512, 1024\}$	64	64
Learning rate	$[.00002, .005]_{\log}$	0.00024	0.00003
Context mode	4 modes	past plus forecast	past plus forecast
Past window	$\{1, 2, 4, 8, 16\}$	1	1
Encoder hidden layers	$\{0, 1, 2\}$	0	1
Encoder dropout	$\{0, .05, .1, .2, .5\}$	n.a.	0.05
ReLU before Hopfield layer	$\{\mathrm{no}, \mathrm{yes}\}$	no	yes
Residual memory size	$\{200, 400, 800, 1500, 3000\}$	1500	3000
Calibration window	$\{500, 1000, 2000, 4000\}$	2000	2000
Calibration pool size	$\{1000, 2000, 4000, 8000\}$	1000	8000
Residual in stored pattern	$\{\mathrm{no}, \mathrm{yes}\}$	no	no
Position encoding	$\{\mathrm{none}, \mathrm{rel.\ simple}, \mathrm{rel.\ linear}\}$	none	rel. linear
Encoder epochs	$\{0, 25, 50, 100, 200\}$	200	200
Fixed across datasets			
Implementation	upstream MSE with absolute residual prediction and one shared head		
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$		
Calibration scope	per series		
Interval construction	Hopfield weighted residual quantile around $\hat{\mu}(x)$		
Test time encoder training	disabled		
Target miscoverage α	0.2		

Table 32: HopCPT hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. One configuration is selected for each backbone using an Optuna TPE study with 100 trials and is shared across datasets and evaluation seeds. Past plus forecast denotes univariate past context together with the backbone forecast.

Hyperparameter	**Search range**	**Toto**	**ARIMA**
Hopfield scale β	$[.5, 24]_{\log}$	2.131	2.131
Batch size	$\{64, 128, 256, 512, 1024\}$	64	64
Learning rate	$[.00002, .005]_{\log}$	0.00003	0.00003
Context mode	4 modes	past plus forecast	past plus forecast
Past window	$\{1, 2, 4, 8, 16\}$	1	1
Encoder hidden layers	$\{0, 1, 2\}$	1	1
Encoder dropout	$\{0, .05, .1, .2, .5\}$	0.05	0.05
ReLU before Hopfield layer	$\{\mathrm{no}, \mathrm{yes}\}$	yes	yes
Residual memory size	$\{200, 400, 800, 1500, 3000\}$	3000	3000
Calibration window	$\{500, 1000, 2000, 4000\}$	2000	2000
Calibration pool size	$\{1000, 2000, 4000, 8000\}$	8000	8000
Residual in stored pattern	$\{\mathrm{no}, \mathrm{yes}\}$	no	no
Position encoding	$\{\mathrm{none}, \mathrm{rel.\ simple}, \mathrm{rel.\ linear}\}$	rel. linear	rel. linear
Encoder epochs	$\{0, 25, 50, 100, 200\}$	200	200
Fixed across datasets			
Implementation	upstream MSE with absolute residual prediction and one shared head		
Nonconformity score	$\lvert y - \hat{\mu}(x)\rvert$		
Calibration scope	per series		
Interval construction	Hopfield weighted residual quantile around $\hat{\mu}(x)$		
Test time encoder training	disabled		
Target miscoverage α	0.2		

C.5.13 RESCP BASELINE

Table 33: ResCP hyperparameter configuration used in the CP methods result tables for the TiREx and TimesFM backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals and are shared across all three reservoir seeds. Similarities are Cos for cosine and Euc for Euclidean. Setting $W = 0$ disables FIFO truncation.

Dataset	**TiREx**						**TimesFM**					
	d_{res}	ρ	s_{in}	τ	S	W	d_{res}	ρ	s_{in}	τ	S	W
Search range	20:20:200	$[.5, .99]$	$[.05, 1]_{\log}$	$[.01, 2]_{\log}$	$\{\text{Cos}, \text{Euc}\}$	0:50:500	20:20:200	$[.5, .99]$	$[.05, 1]_{\log}$	$[.01, 2]_{\log}$	$\{\text{Cos}, \text{Euc}\}$	0:50:500
electricity_15T	200	0.990	0.165	0.157	Euc	250	200	0.923	0.094	0.284	Euc	50
electricity_D	140	0.966	0.508	0.345	Euc	50	140	0.932	0.196	0.354	Cos	50
electricity_H	140	0.769	0.723	0.898	Euc	50	200	0.935	0.104	1.556	Euc	50
LOOP_SEATTLE_5T	200	0.975	0.226	0.144	Euc	450	160	0.954	0.430	0.184	Euc	250
LOOP_SEATTLE_H	200	0.990	0.123	0.128	Cos	450	200	0.973	0.127	0.176	Euc	250
m4_daily	180	0.967	0.510	0.144	Cos	350	200	0.953	0.750	0.193	Euc	100
m4_hourly	180	0.984	0.082	0.173	Cos	200	160	0.980	0.463	0.133	Cos	400
m4_monthly	180	0.912	0.210	0.858	Euc	50	160	0.945	0.669	0.816	Euc	50
m4_quarterly	100	0.970	0.873	0.133	Cos	400	140	0.964	0.510	0.571	Euc	50
m4_weekly	180	0.810	0.968	1.689	Cos	50	180	0.989	0.342	0.439	Cos	50
M_DENSE_D	120	0.968	0.198	0.150	Euc	250	100	0.960	0.128	0.102	Cos	350
M_DENSE_H	60	0.851	0.546	1.978	Cos	100	60	0.839	0.205	0.809	Euc	100
saugeenday_D	140	0.766	0.376	0.363	Cos	50	40	0.989	0.422	0.111	Cos	0
saugeenday_M	20	0.541	0.395	0.294	Euc	450	100	0.615	0.472	0.211	Euc	350
saugeenday_W	100	0.961	0.649	0.208	Cos	400	140	0.941	0.080	0.187	Cos	500
solar_10T	40	0.533	0.056	0.435	Euc	450	100	0.537	0.067	1.485	Euc	100
solar_H	200	0.779	0.111	0.138	Cos	450	160	0.938	0.076	0.055	Cos	0
SZ_TAXI_15T	40	0.989	0.662	0.108	Euc	0	180	0.990	0.663	0.542	Euc	150
SZ_TAXI_H	120	0.988	0.323	0.445	Euc	500	120	0.514	0.260	0.818	Cos	0
temperature_rain	200	0.979	0.166	0.095	Cos	400	200	0.931	0.144	0.137	Euc	450
us_births_D	140	0.919	0.095	0.219	Euc	0	100	0.963	0.105	0.641	Cos	500
us_births_W	60	0.989	0.058	0.171	Euc	500	180	0.760	0.638	1.622	Euc	350

Fixed across datasets	
Residual input	signed residual $y - \hat{\mu}(x)$
Reservoir dynamics	sparse $\tanh$ reservoir with connectivity 0.3 and no leakage
State normalization	unit ℓ_2 norm
Time weighting	linear recency ramp
Interval construction	narrowest weighted residual interval among 100 tail allocations
Target miscoverage α	0.2

Table 34: ResCP hyperparameter configuration used in the CP methods result tables for the Toto and ARIMA backbones. Best values come from Optuna TPE with 100 trials per dataset and backbone pair ranked by mean nWink on the validation split. Search ranges appear in the second header row. Continuous values are rounded to three decimals and are shared across all three reservoir seeds. Similarities are Cos for cosine and Euc for Euclidean. Setting $W = 0$ disables FIFO truncation.

Dataset	**Toto**						**ARIMA**					
	d_{res}	ρ	s_{in}	τ	S	W	d_{res}	ρ	s_{in}	τ	S	W
Search range	20:20:200	$[.5, .99]$	$[.05, 1]_{\log}$	$[.01, 2]_{\log}$	$\{\text{Cos}, \text{Euc}\}$	0:50:500	20:20:200	$[.5, .99]$	$[.05, 1]_{\log}$	$[.01, 2]_{\log}$	$\{\text{Cos}, \text{Euc}\}$	0:50:500
electricity_15T	200	0.636	0.107	1.998	Cos	50	160	0.987	0.726	0.196	Euc	150
electricity_D	160	0.965	0.210	0.528	Euc	50	200	0.911	0.097	0.322	Euc	50
electricity_H	120	0.989	0.995	0.191	Cos	250	200	0.977	0.082	0.182	Cos	150
LOOP_SEATTLE_5T	160	0.973	0.230	0.260	Euc	300	200	0.986	0.229	0.210	Euc	100
LOOP_SEATTLE_H	200	0.974	0.105	0.203	Cos	100	160	0.940	0.103	0.135	Cos	150
m4_daily	60	0.946	0.170	0.233	Euc	100	100	0.829	0.069	0.170	Euc	500
m4_hourly	60	0.917	0.623	0.195	Cos	100	160	0.956	0.338	0.201	Euc	350
m4_monthly	20	0.978	0.884	1.324	Euc	50	200	0.939	0.108	0.459	Euc	50
m4_quarterly	60	0.941	0.491	1.298	Cos	50	100	0.960	0.223	0.581	Cos	50
m4_weekly	180	0.936	0.070	0.874	Euc	50	200	0.976	0.101	0.570	Euc	50
M_DENSE_D	160	0.987	0.244	0.156	Euc	300	140	0.949	0.090	0.175	Euc	300
M_DENSE_H	60	0.954	0.187	1.319	Cos	100	200	0.925	0.117	0.174	Euc	300
saugeenday_D	200	0.990	0.872	0.115	Cos	0	140	0.901	0.182	0.375	Cos	50
saugeenday_M	160	0.869	0.257	0.426	Cos	150	200	0.516	0.518	0.223	Euc	500
saugeenday_W	120	0.990	0.188	0.217	Euc	500	180	0.904	0.261	0.314	Cos	500
solar_10T	60	0.977	0.116	1.528	Euc	50	40	0.627	0.294	0.480	Euc	100
solar_H	200	0.984	0.175	0.096	Cos	450	160	0.988	0.592	0.073	Euc	0
SZ_TAXI_15T	60	0.959	0.221	0.239	Cos	0	100	0.890	0.693	0.057	Cos	0
SZ_TAXI_H	140	0.966	0.404	0.524	Cos	500	180	0.965	0.240	0.369	Cos	500
temperature_rain	200	0.906	0.214	0.099	Cos	250	100	0.891	0.102	0.103	Cos	400
us_births_D	140	0.831	0.234	0.726	Cos	0	140	0.969	0.213	0.115	Euc	500
us_births_W	20	0.829	0.302	1.571	Cos	200	100	0.925	0.175	0.171	Euc	350

Fixed across datasets

Residual input	signed residual $y - \hat{\mu}(x)$
Reservoir dynamics	sparse $\tanh$ reservoir with connectivity 0.3 and no leakage
State normalization	unit ℓ_2 norm
Time weighting	linear recency ramp
Interval construction	narrowest weighted residual interval among 100 tail allocations
Target miscoverage α	0.2

C.5.14 MoE-CP BASELINE

Table 35: MoE CP hyperparameter configuration used in the CP methods result tables. MoE CP trains its own point predictor and is tuned separately for each dataset. The values were selected by Optuna TPE with 100 trials ranked by mean nWink on the validation split. Divergences are KL for Kullback Leibler, Jef for Jeffreys, Rén for Rényi of order 0.5, Hel for Hellinger, Euc for Euclidean, and Cos for cosine.

Dataset *Search range*	K $\{2, 3, 4, 5\}$	τ $\{50, 60, \ldots, 300\}$	D 6 divergences
electricity_15T	3	300	Cos
electricity_D	3	280	KL
electricity_H	4	80	Rén
LOOP_SEATTLE_5T	4	50	Cos
LOOP_SEATTLE_H	2	80	Cos
m4_daily	5	200	Cos
m4_hourly	5	250	Jef
m4_monthly	3	240	Rén
m4_quarterly	2	280	Rén
m4_weekly	2	140	Rén
M_DENSE_D	3	90	Rén
M_DENSE_H	4	210	Rén
saugeenday_D	5	80	Rén
saugeenday_M	5	120	Cos
saugeenday_W	5	70	Rén
solar_10T	4	60	Rén
solar_H	2	50	Rén
SZ_TAXI_15T	3	190	Cos
SZ_TAXI_H	5	230	Cos
temperature_rain	2	50	Cos
us_births_D	2	90	Rén
us_births_W	2	150	Rén

Fixed across datasets	
Expert and gate network	MLP, 3×64 hidden units, ReLU, dropout 0.1
Optimizer	Adam, lr 10^{-4}, batch size 64, MSE loss
Early stopping	max. 2000 epochs, patience 20, 10% holdout
Training set size	up to 30,000 windows
Nonconformity score	$\lvert y - \hat{\mu}(x) \rvert$
Target miscoverage α	0.2
Infinite width intervals	capped at $10\times$ the largest finite width

Table 35 lists the configuration of our MoE-CP implementation. Unlike the other baselines, MoE-CP Kong et al. (2026) does not use a backbone forecast. It fits its own mixture of K expert MLPs, with a softmax gate that assigns each input a vector of expert weights. When building an interval, MoE-CP assigns more weight to calibration points whose gate vectors are similar to that of the test point. We tune three hyperparameters per dataset. K is the number of experts. We train one model for each value of K and reuse it across trials. τ controls how sharply the weights favour similar calibration points. D is the divergence used to compare gate vectors. We search over the six options from the original paper. The fixed settings also follow the original paper. MoE-CP can return an infinite interval when too little calibration weight falls on similar points. For evaluation, we replace these with a large finite width.

C.6 RUNTIME ANALYSIS

Table 36: Wall-clock runtime of the conformal layer on the benchmark with the TiREx backbone (backbone inference excluded), reported as the mean over 10 seed runs. †Excludes a one-time teacher bank (168.67 s) shared across seeds. *Our own implementation, as no public code release is available.

Method	Train (s/sample) ↓	Train (s/epoch) ↓	Inference (s/sample) ↓
Uniform	—	—	5.68×10^{-5}
ACI	—	—	6.16×10^{-5}
DtACI	—	—	1.33×10^{-4}
NExCP	—	—	6.64×10^{-5}
KOWCPI	—	—	3.08×10^{-2}
ResCP	—	—	5.57×10^{-3}
COP	—	—	1.93×10^{-5}
Decay-OGD	—	—	1.70×10^{-6}
ECI	—	—	9.55×10^{-6}
PID	—	—	1.10×10^{-5}
SAOCP	—	—	1.02×10^{-2}
SF-OGD	—	—	4.34×10^{-6}
MoE-CP*	2.64×10^{-3}	24.28	1.44×10^{-3}
HopCPT	8.73×10^{-6}	0.052	2.07×10^{-5}
RareCP *(Ours)*	6.29×10^{-2} †	208.57	3.64×10^{-4}

All wall-clock measurements were collected on a single NVIDIA RTX A6000 GPU (48 GB VRAM) with a 64-core CPU and 256 GB RAM. Every method reads from the same cache of precomputed TiREx predictions, so backbone inference is excluded from all entries and only the conformal layer is timed. The training-free baselines (Uniform, ACI, DtACI, NExCP, KOWCPI, ResCP, COP, Decay-OGD, ECI, PID, SAOCP, and SF-OGD) have no offline training stage, so their entire cost lies in the prediction loop. For each dataset, we time the call that produces the intervals and divide by the number of test samples to obtain the per-sample inference cost reported in Table 36. For MoE-CP, HopCPT, and RareCP, we additionally report the offline training cost, both normalised per training sample and per epoch.

C.7 PARAMETER COUNT

Table 37: Parameter counts of the foundation models and the RareCP wrapper used throughout the evaluation. Counts include native uncertainty heads where applicable. The RareCP wrapper adds at most $\sim$6.13M trainable parameters on top of the backbone, approximately 2% of the parameter count of the largest FM baseline.

Model / wrapper	**Parameters**
TiRex	35.3M
TiRex2	82.5M
Chronos2	119.5M
Moirai2	11.4M
PatchTST-FM	257.9M
TabPFN-TS3	58.3M
TimesFM2.5	231.3M
Toto2	312.7M
TTM	5.0M
RareCP (10 experts)	2.25–6.13M†

† Per backbone: TiRex 2.96M, TimesFM2.5 5.20M, Toto2 6.13M, ARIMA 2.25M.

C.8 VISUALIZATIONS

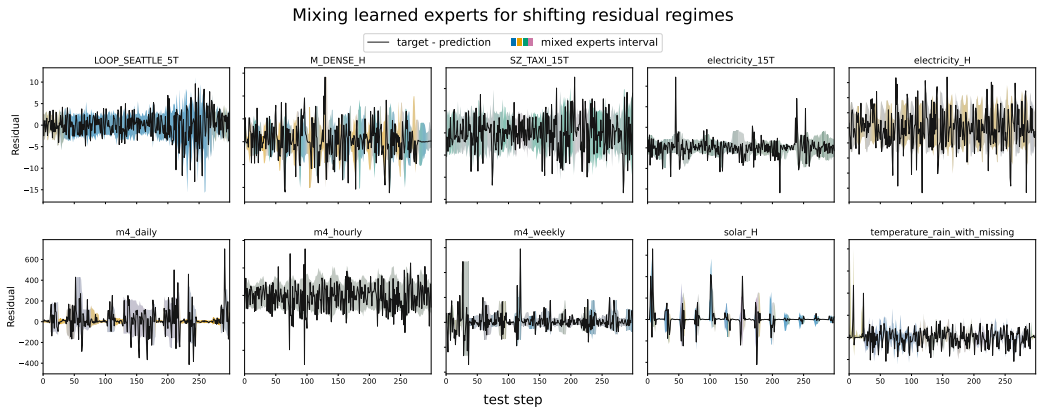


Figure 6: Visualization of 80% intervals using the gating mixture on the 10 datasets.

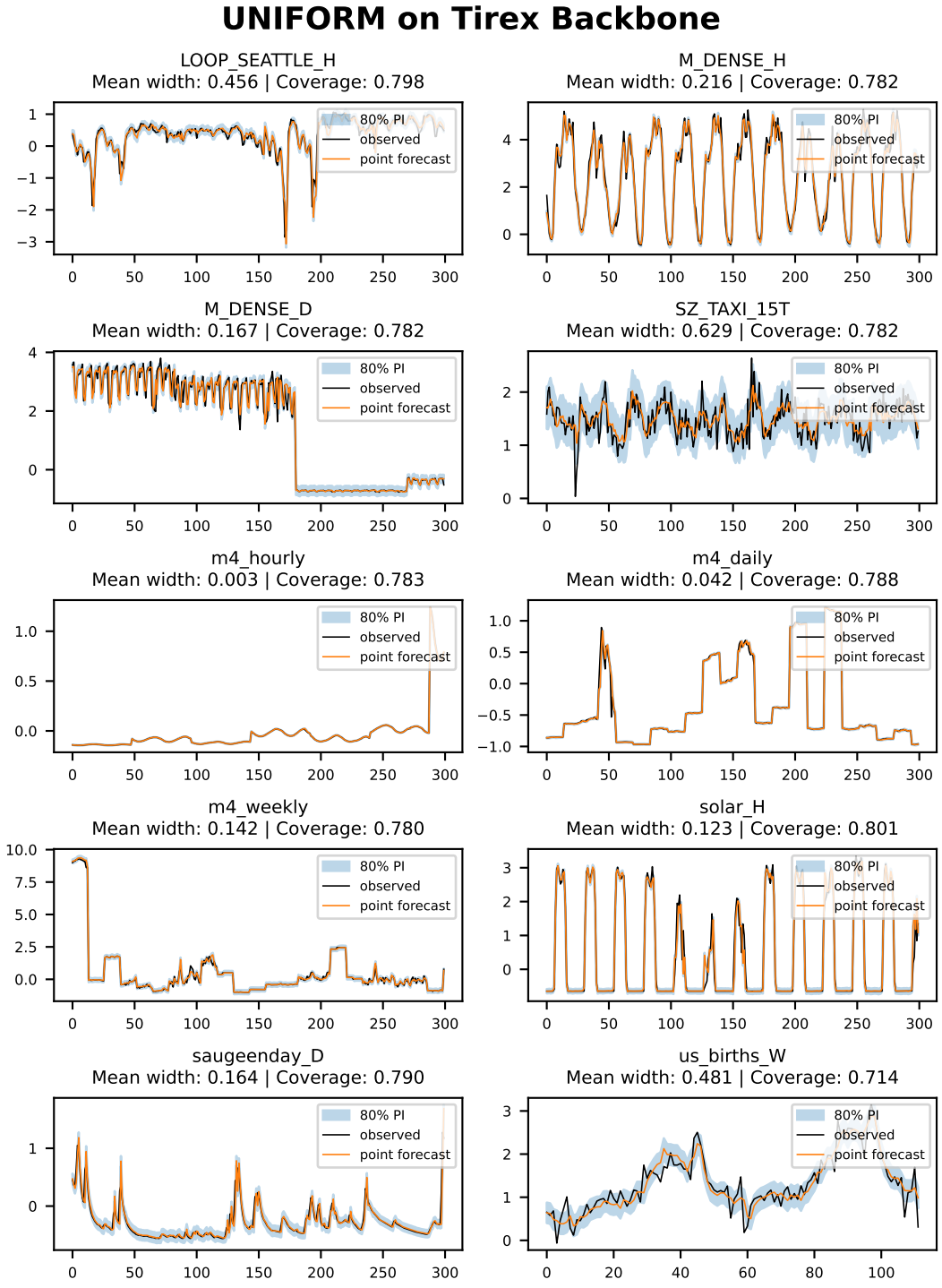


Figure 7: 80% prediction intervals from Uniform on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

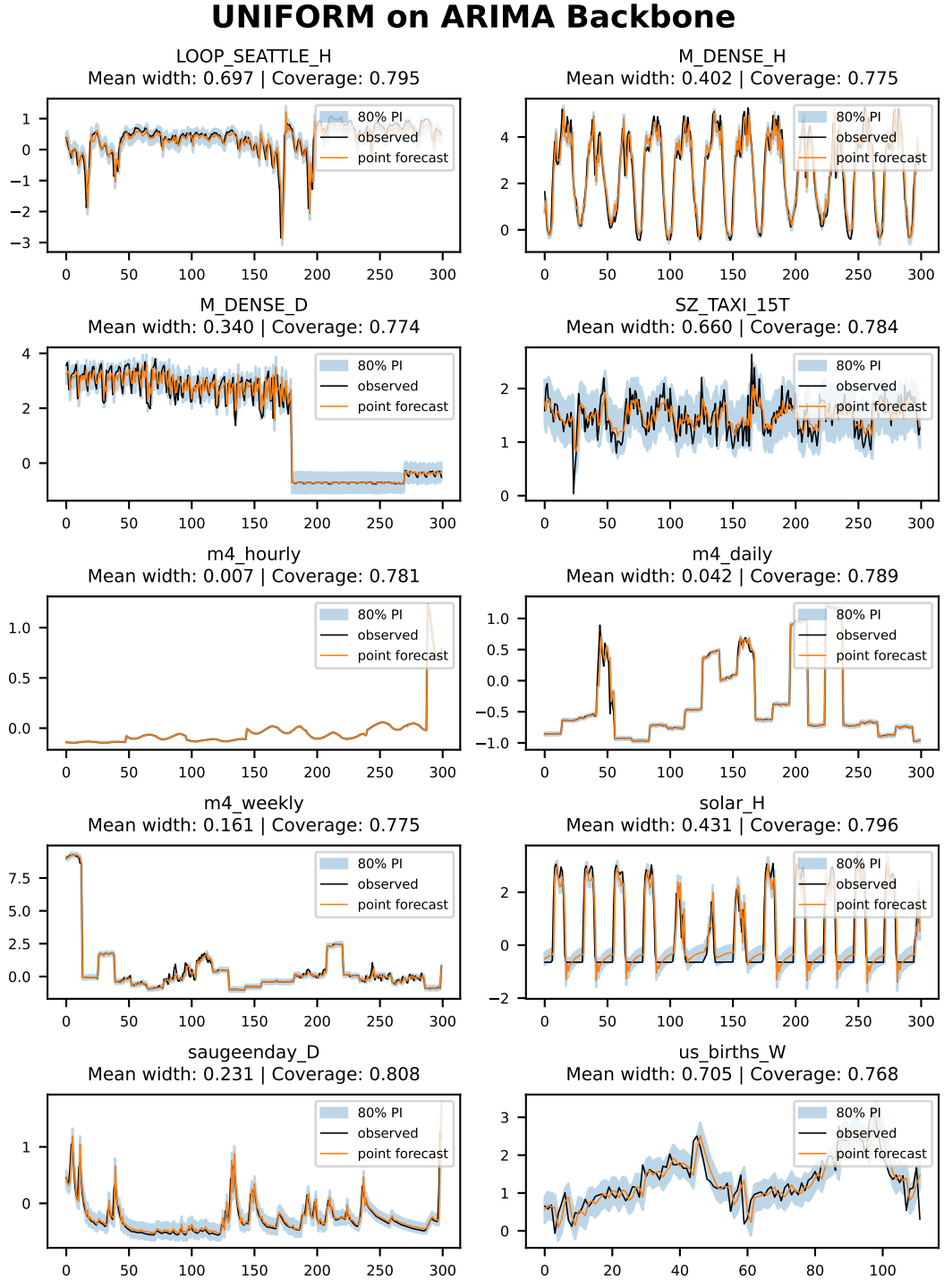


Figure 8: 80% prediction intervals from Uniform on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

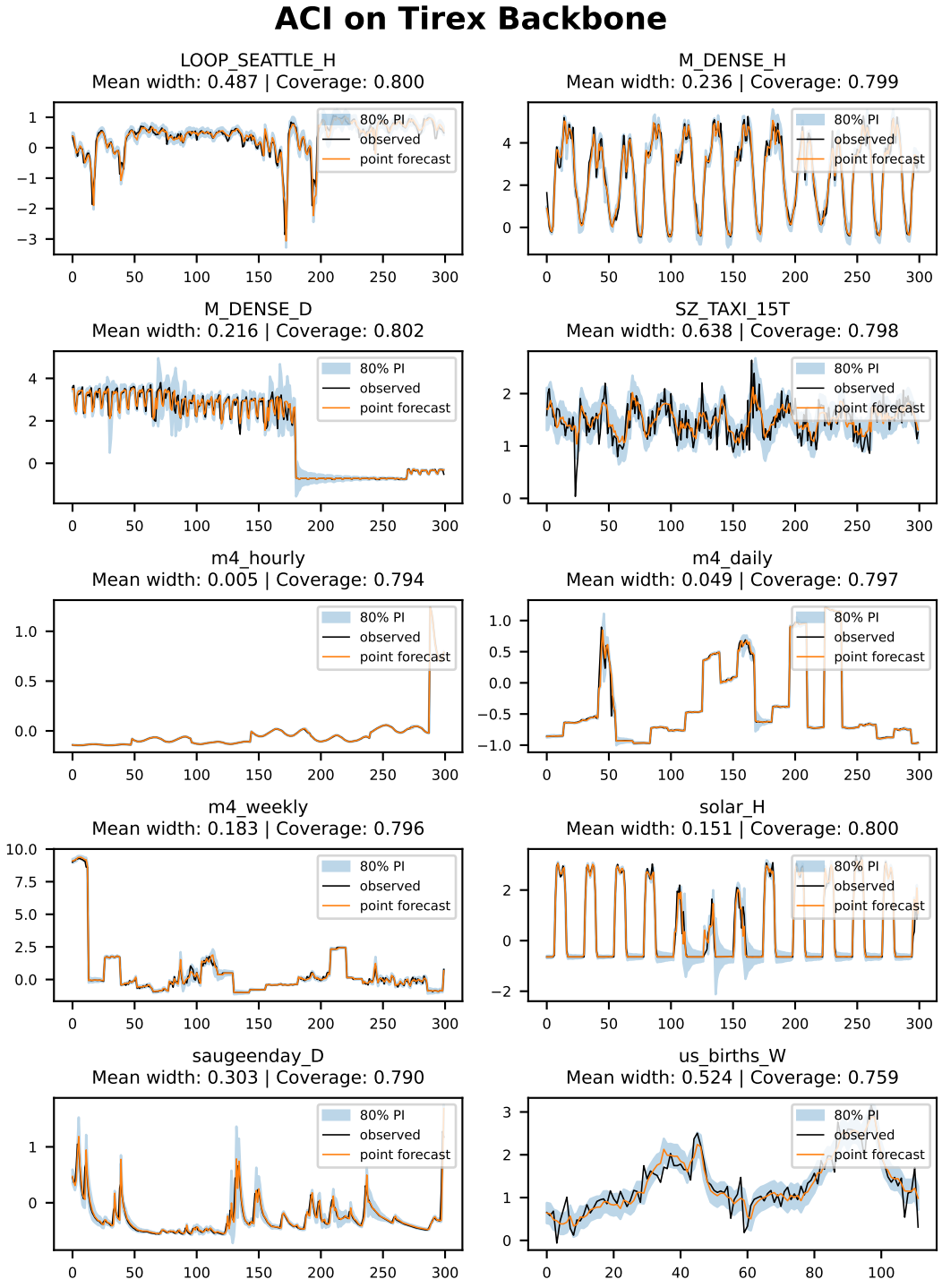


Figure 9: 80% prediction intervals from ACI on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

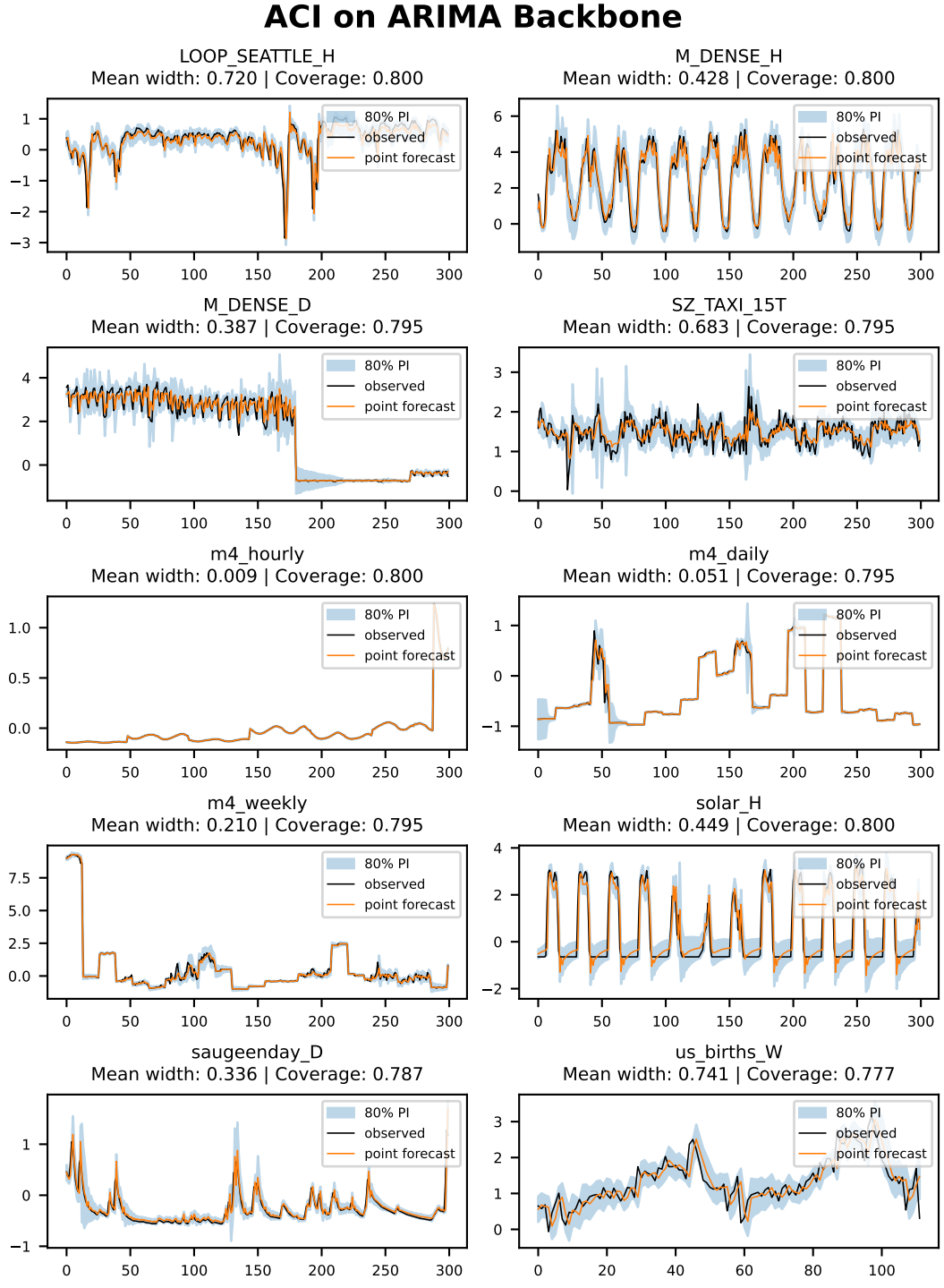


Figure 10: 80% prediction intervals from ACI on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

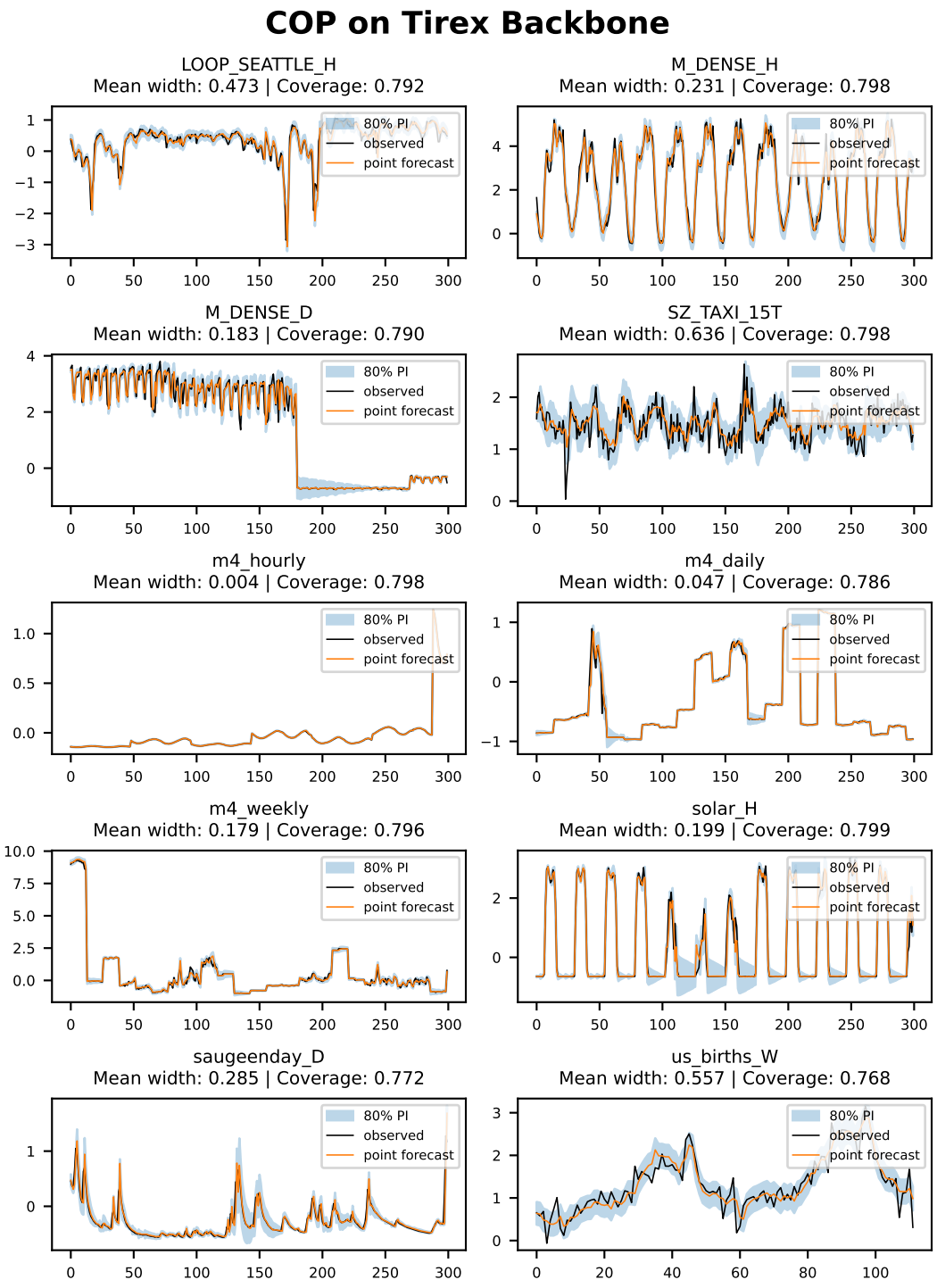


Figure 11: 80% prediction intervals from COP on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

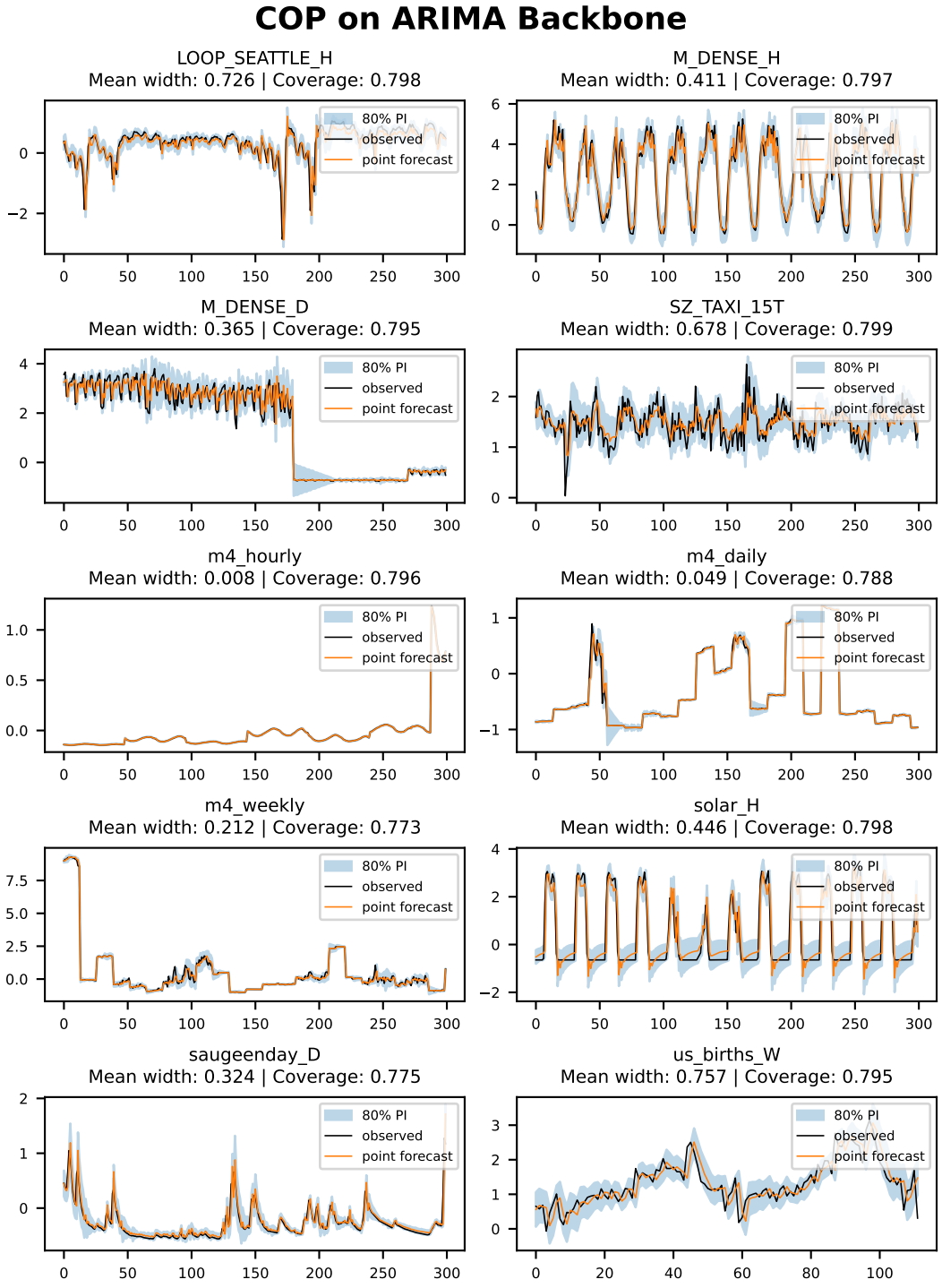


Figure 12: 80% prediction intervals from COP on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

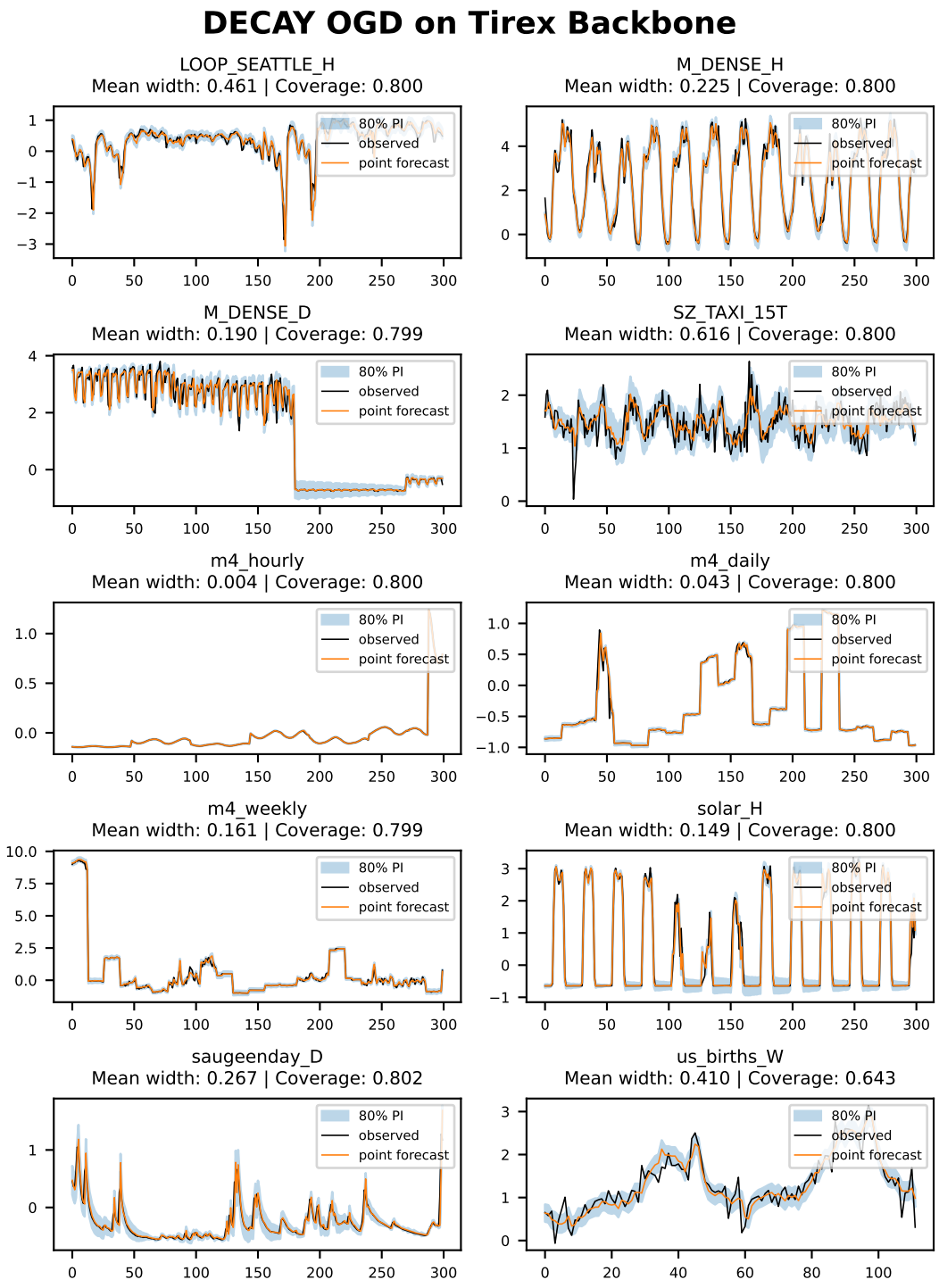


Figure 13: 80% prediction intervals from Decay-OGD on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

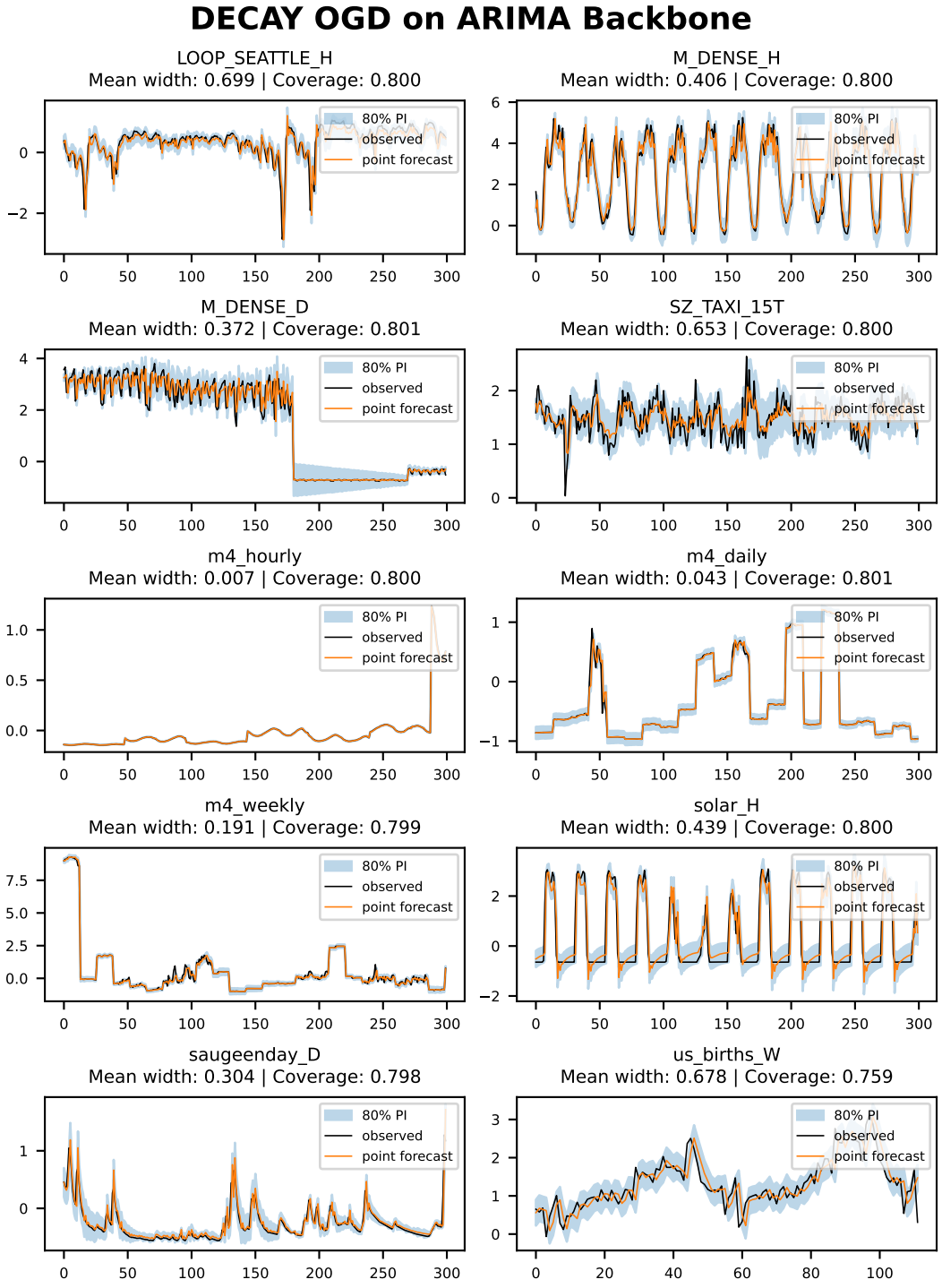


Figure 14: 80% prediction intervals from Decay-OGD on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

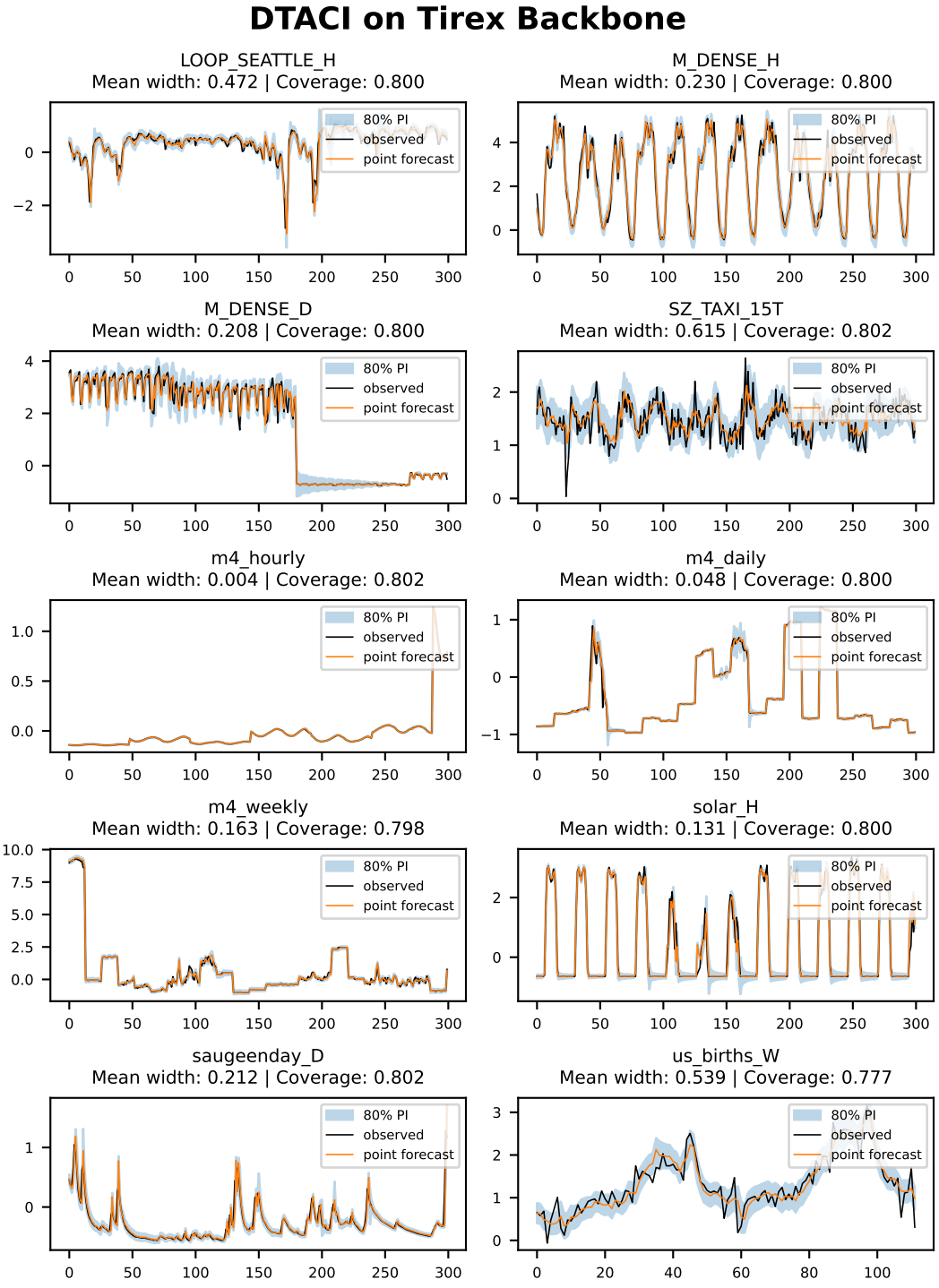


Figure 15: 80% prediction intervals from DtACI on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

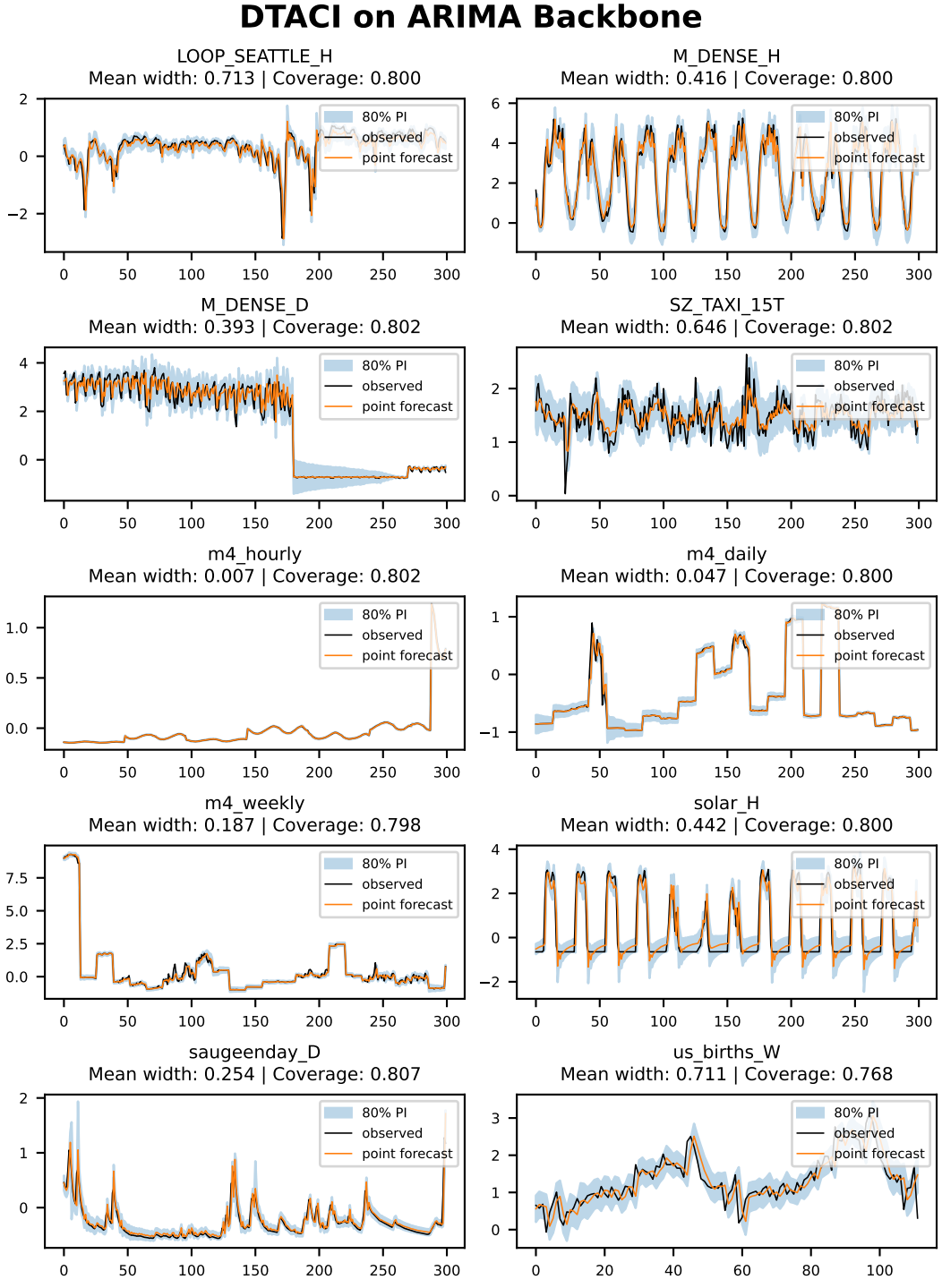


Figure 16: 80% prediction intervals from DtACI on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

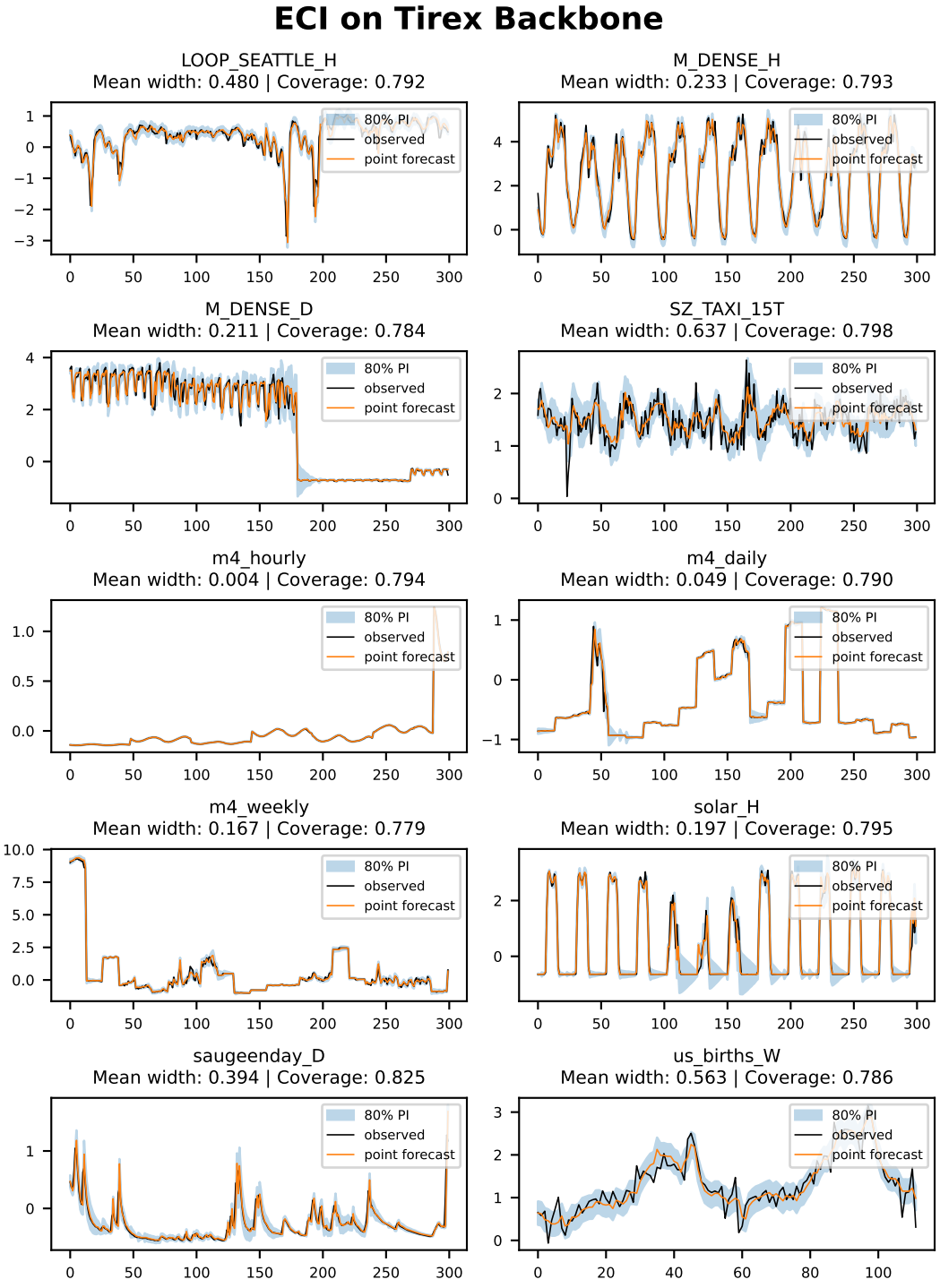


Figure 17: 80% prediction intervals from ECI on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

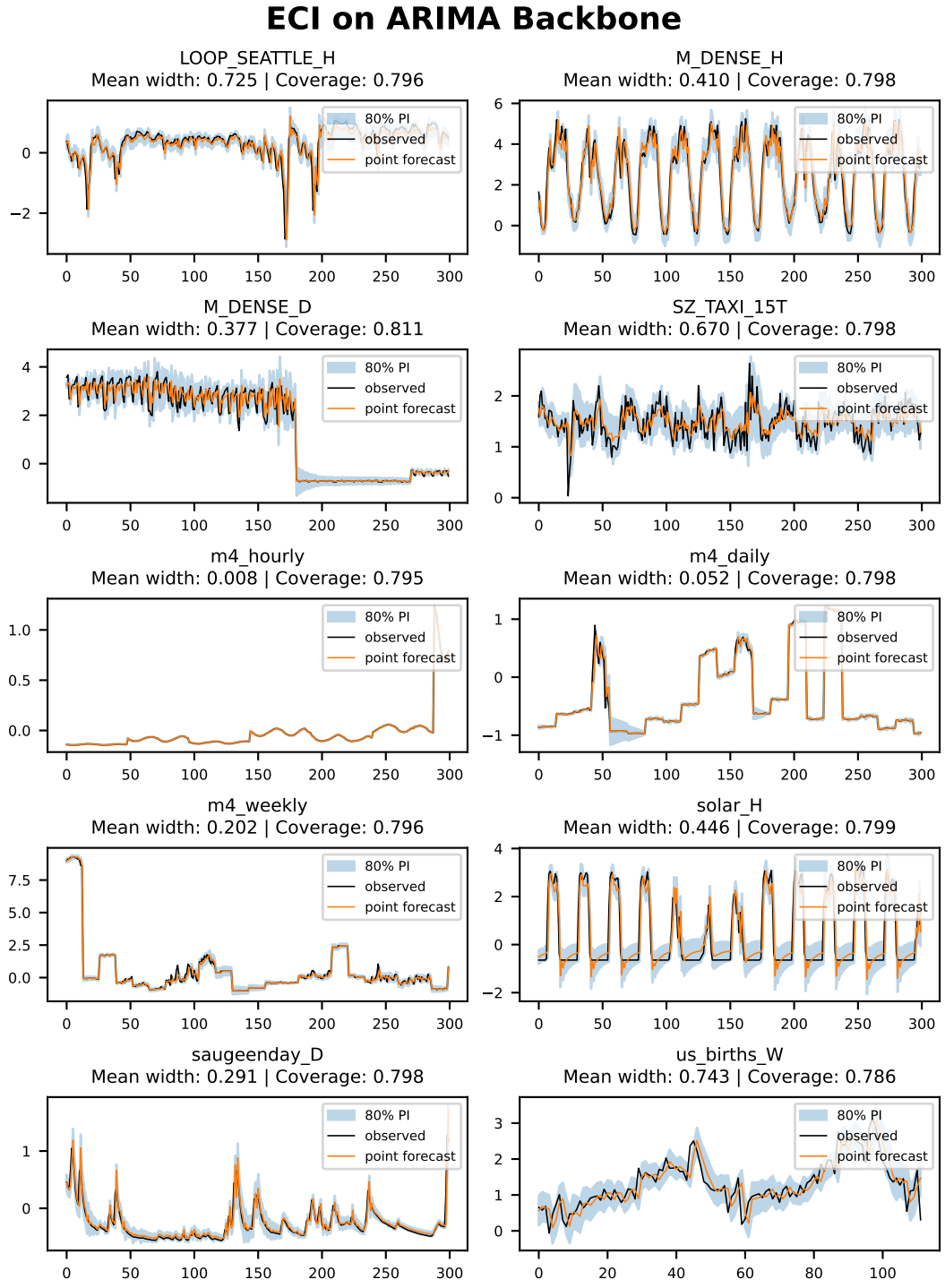


Figure 18: 80% prediction intervals from ECI on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

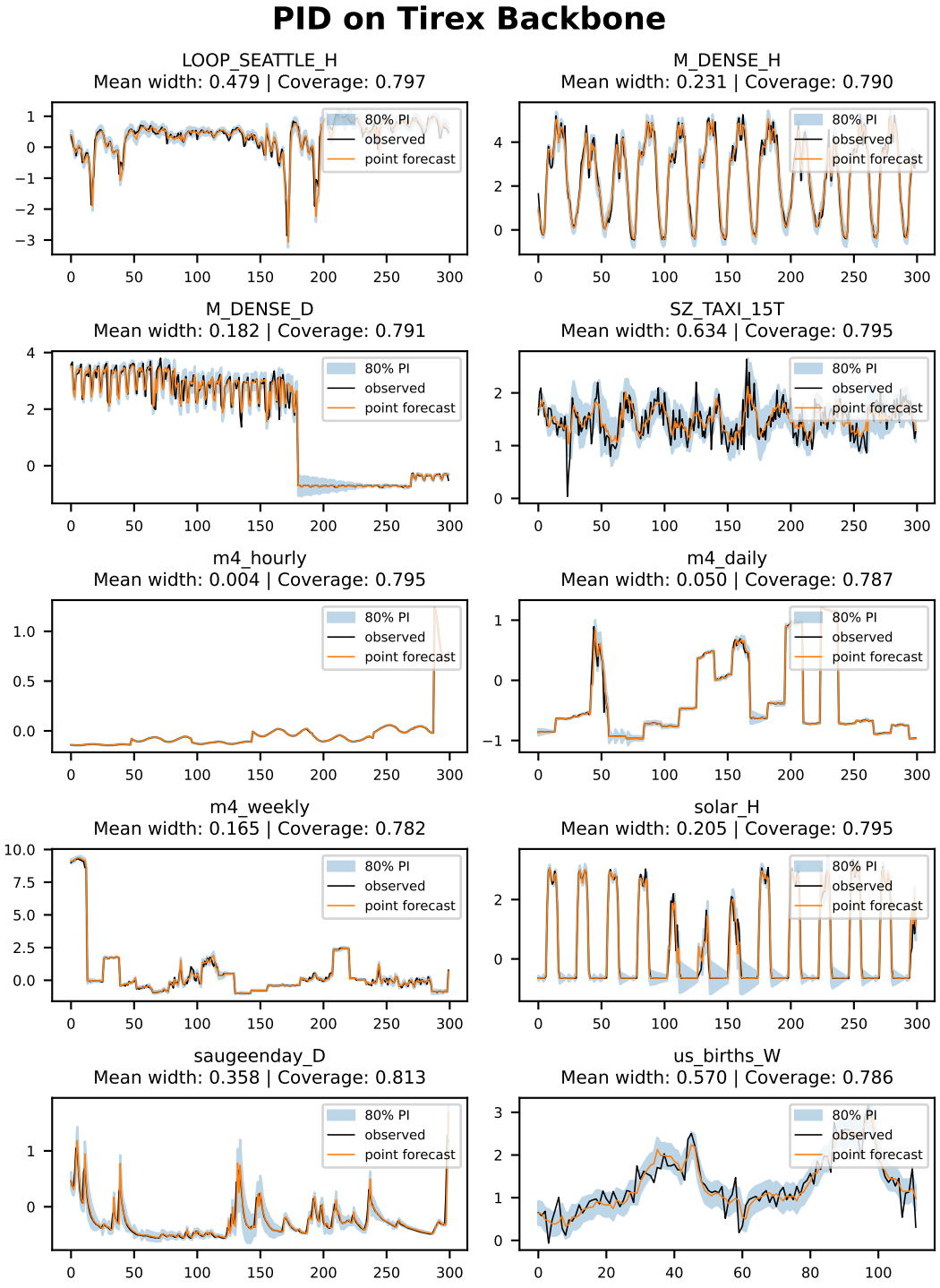


Figure 19: 80% prediction intervals from PID on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

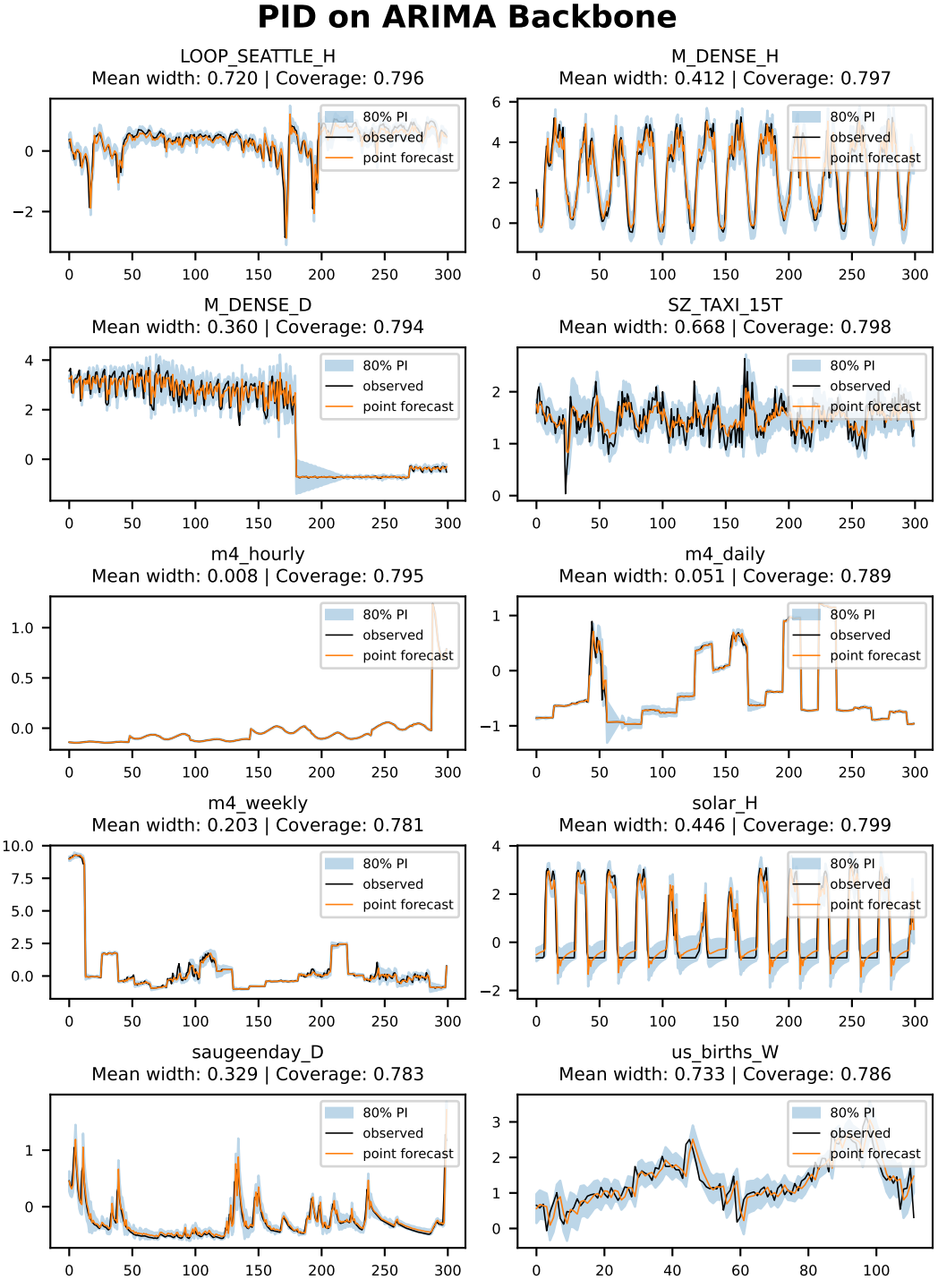


Figure 20: 80% prediction intervals from PID on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

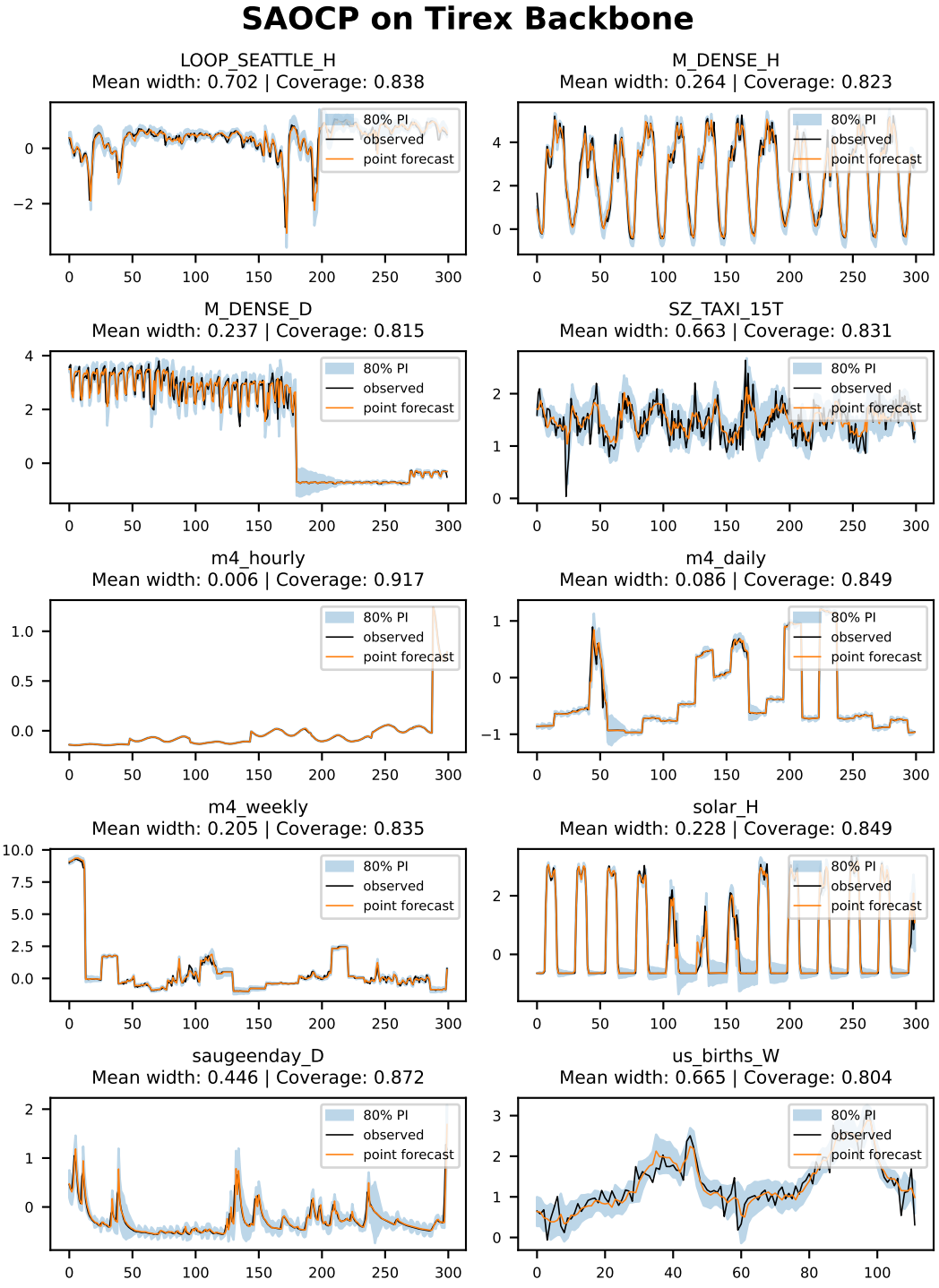


Figure 21: 80% prediction intervals from SAOCP on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

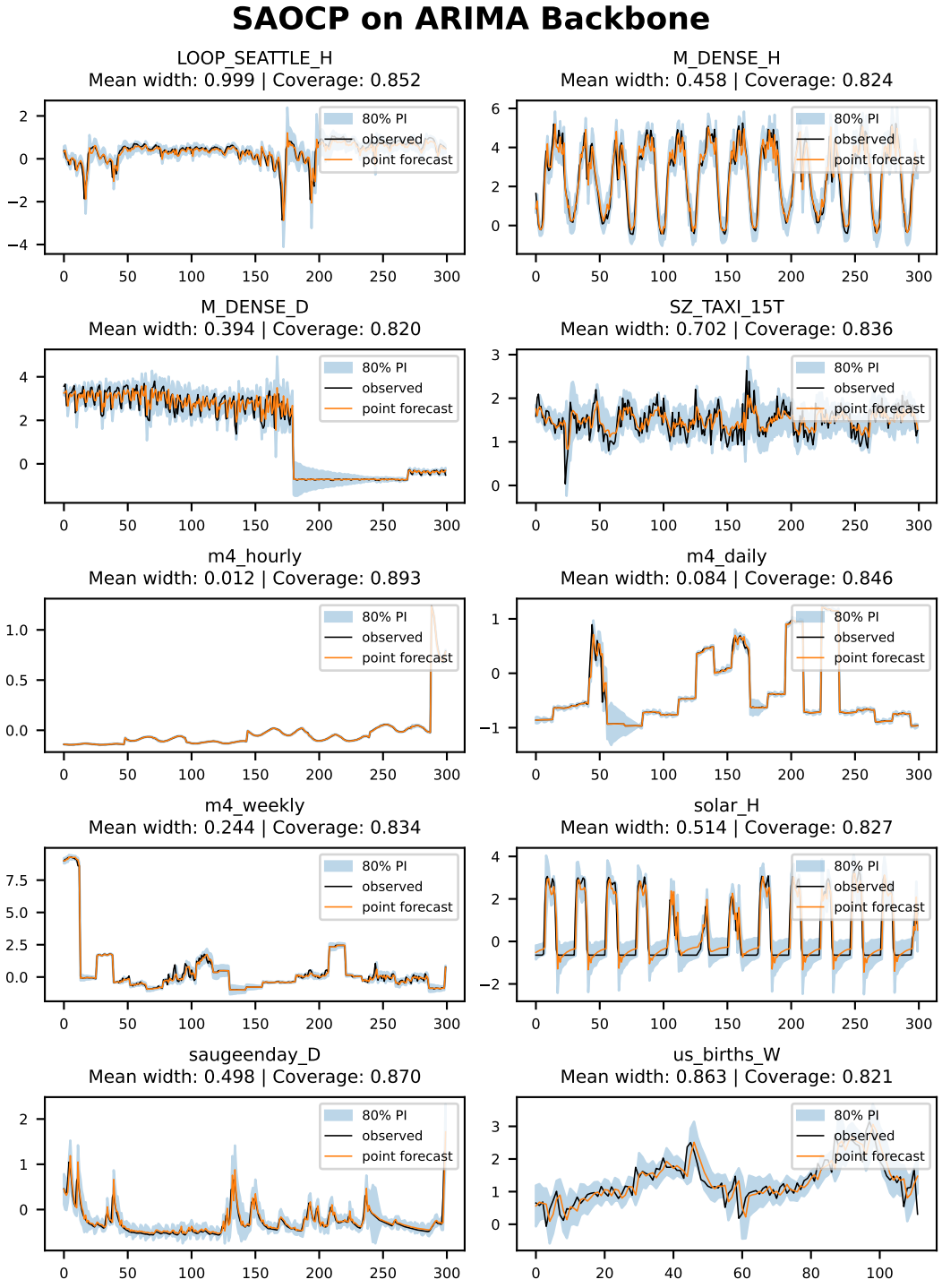


Figure 22: 80% prediction intervals from SAOCP on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

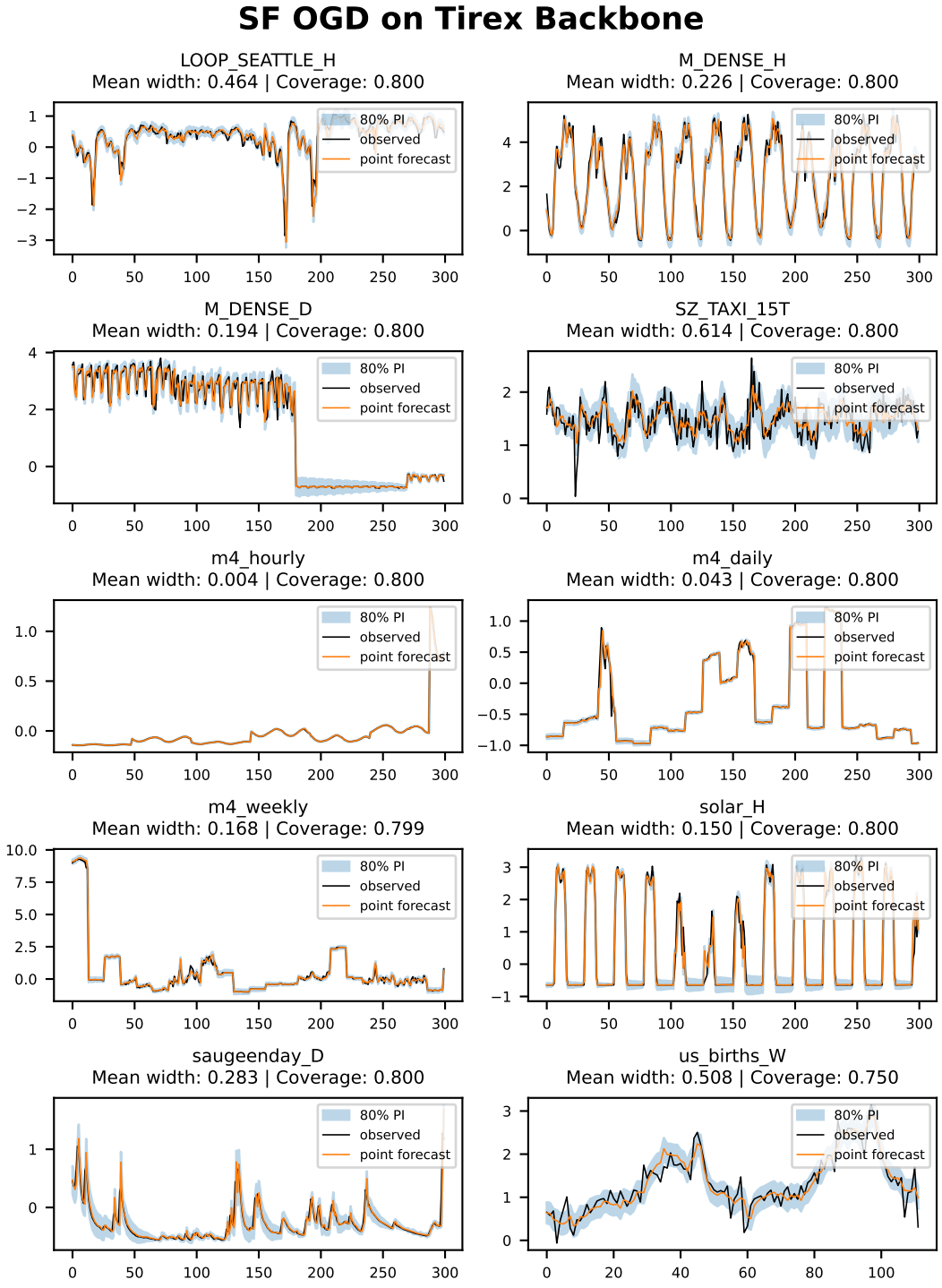


Figure 23: 80% prediction intervals from SF-OGD on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

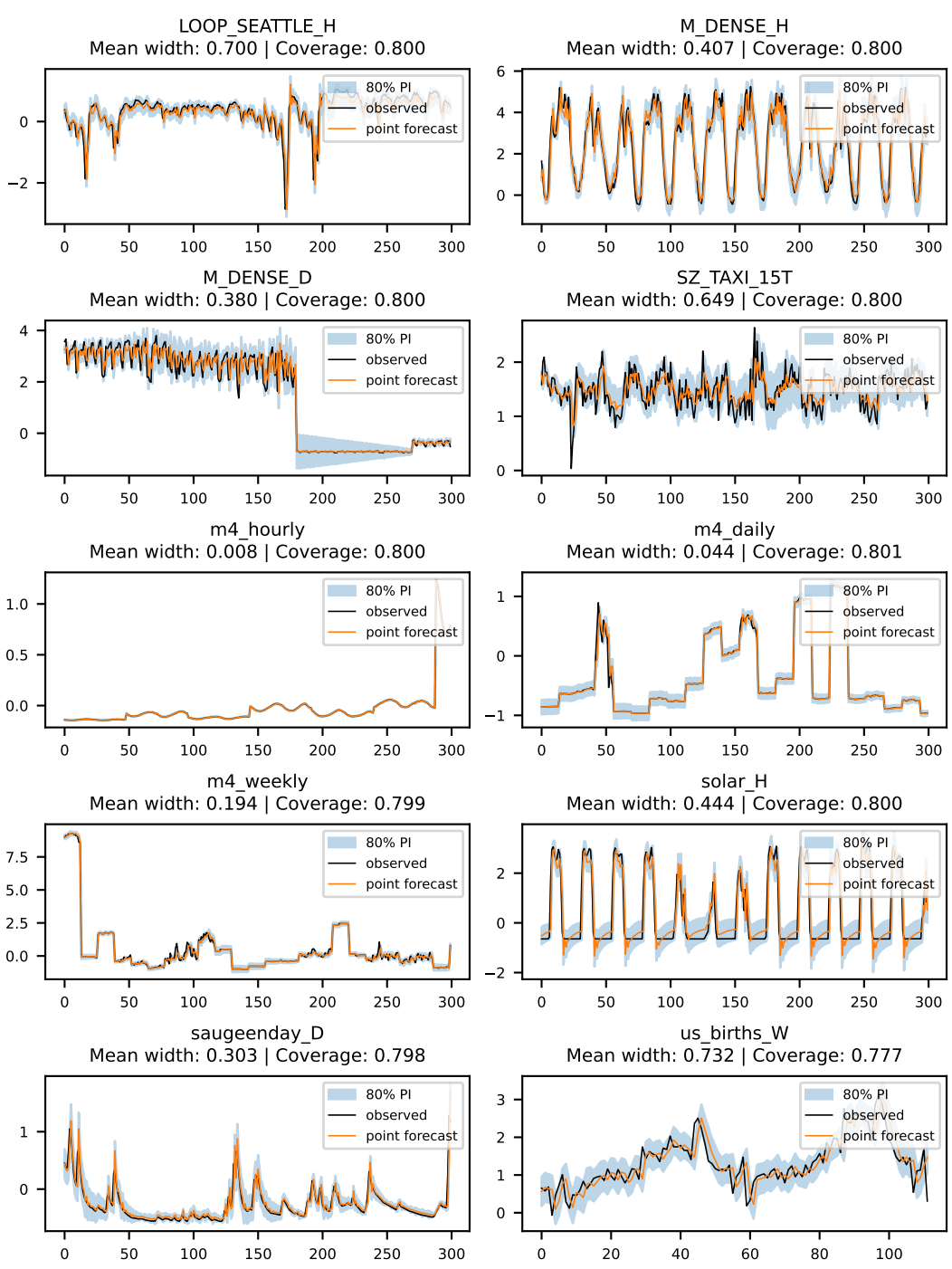


Figure 24: 80% prediction intervals from SF-OGD on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

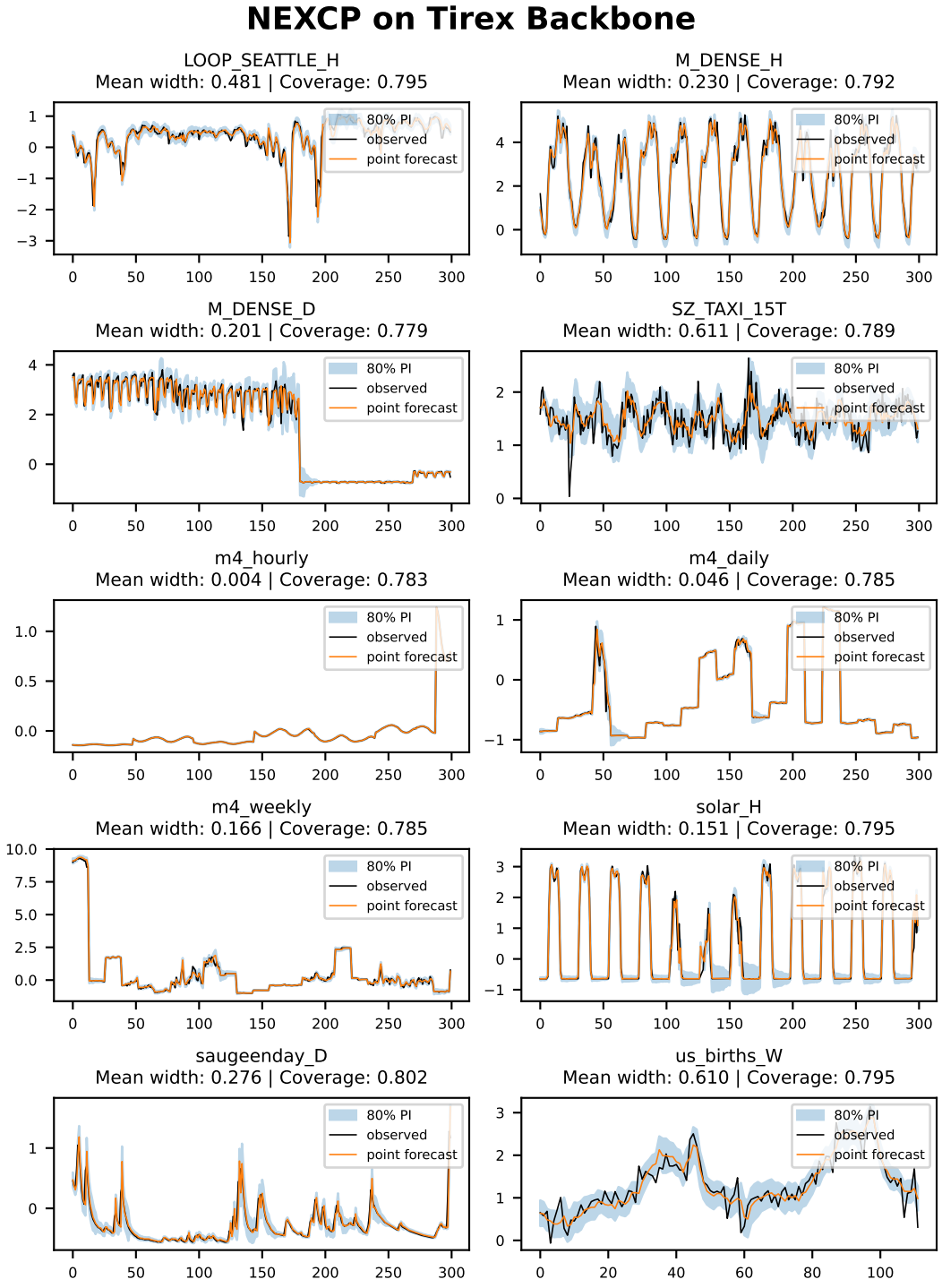


Figure 25: 80% prediction intervals from NexCP on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

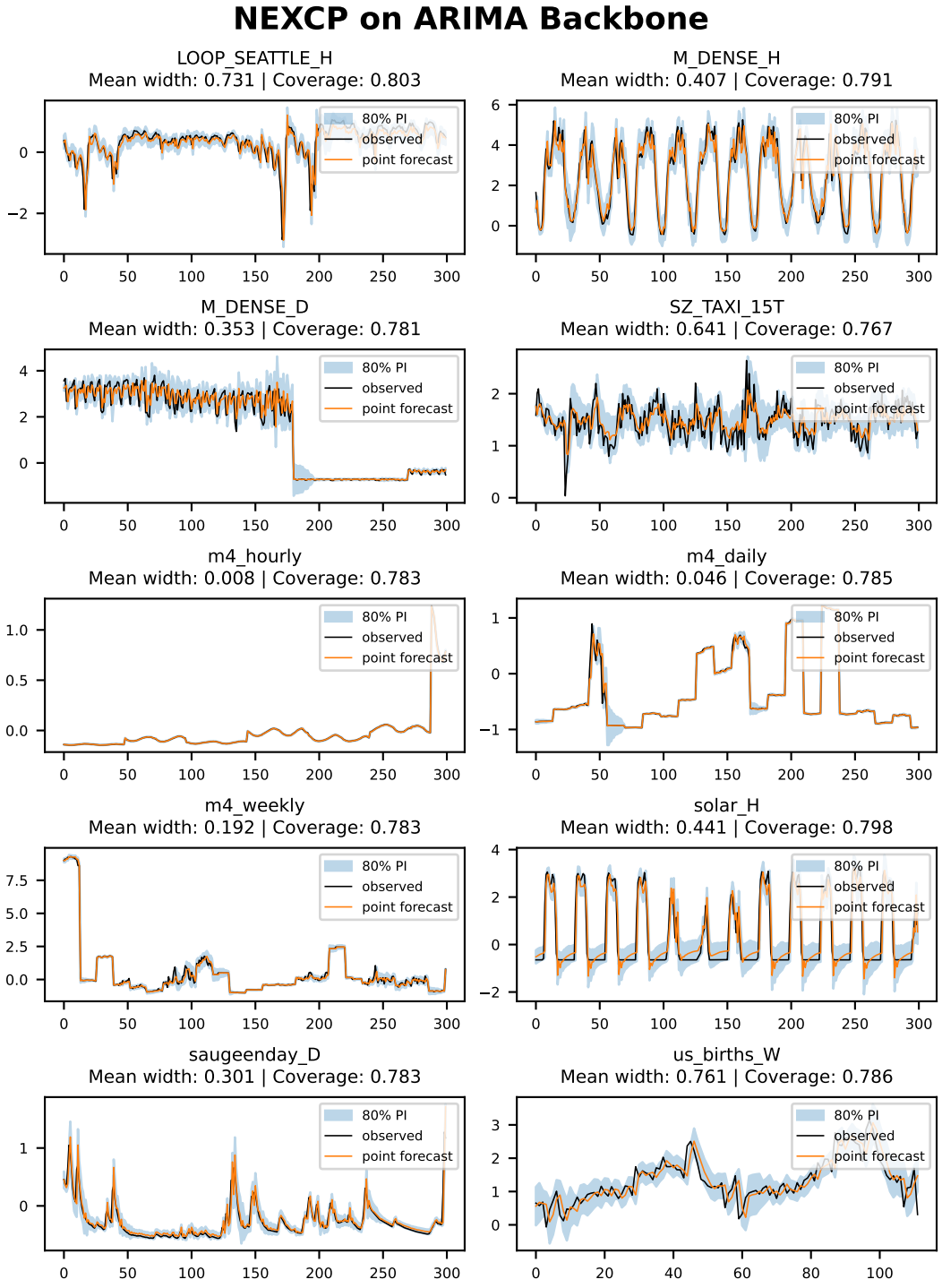


Figure 26: 80% prediction intervals from NexCP on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

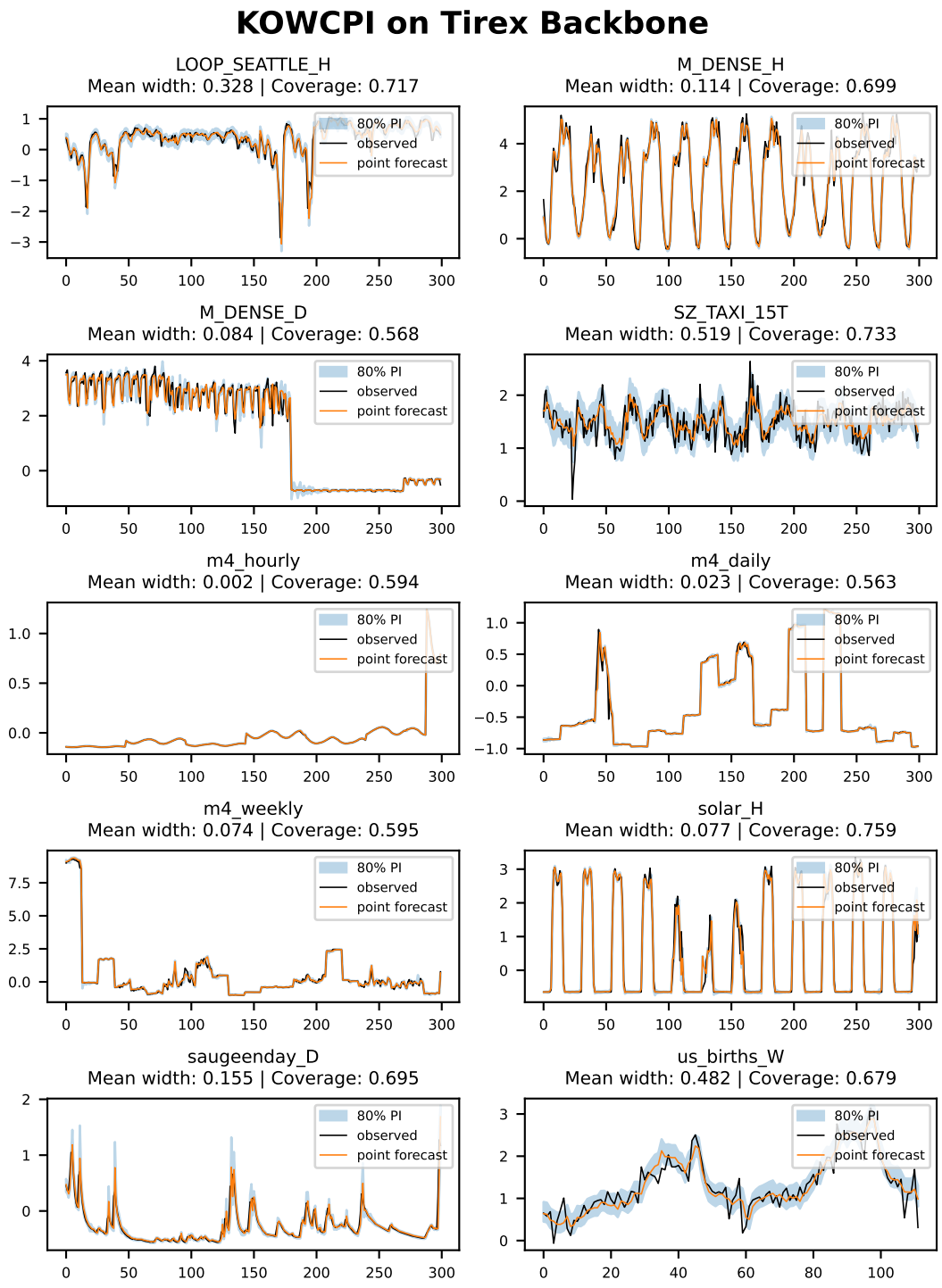


Figure 27: 80% prediction intervals from KOWCPI on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

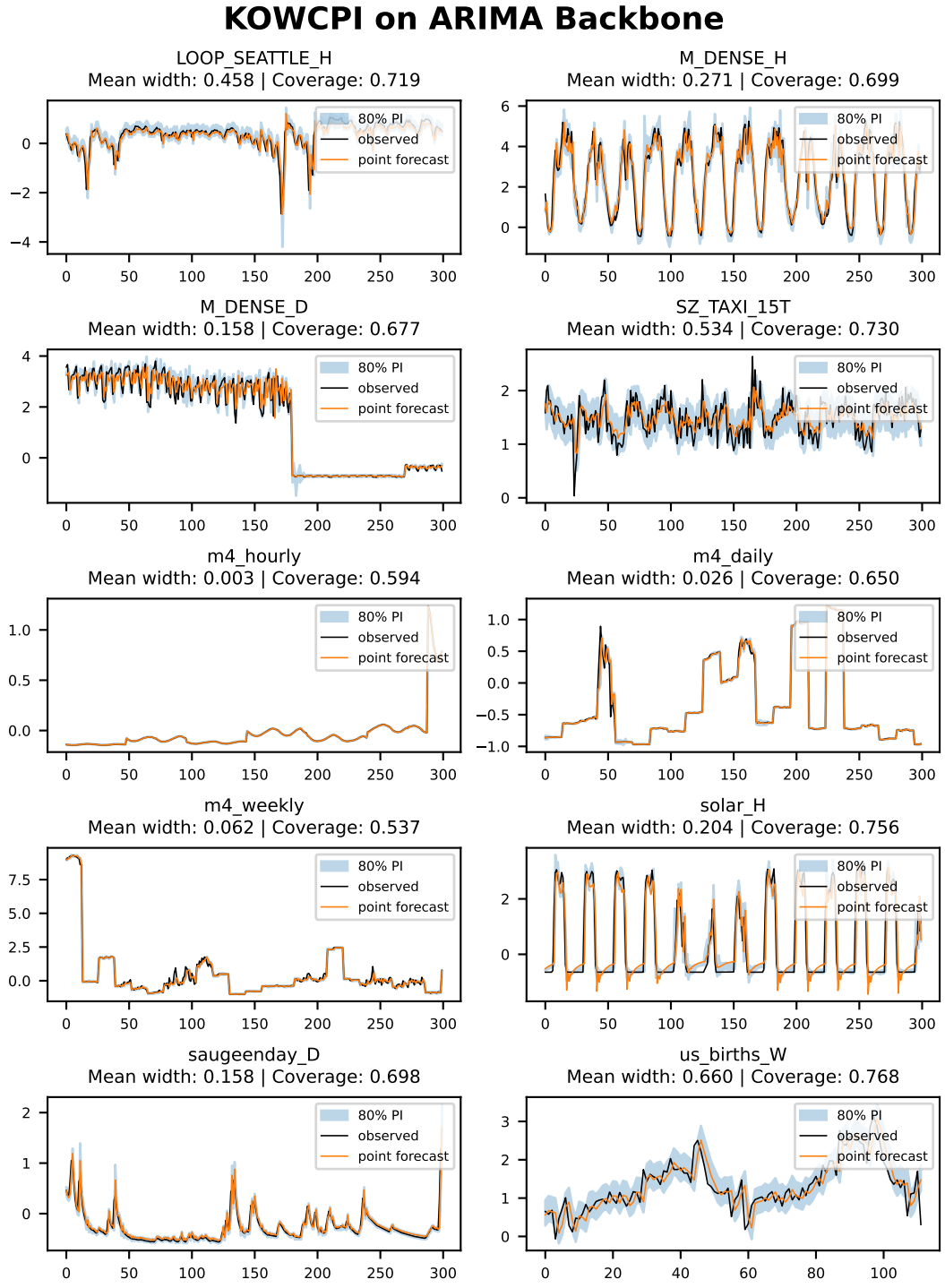


Figure 28: 80% prediction intervals from KOWCPI on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

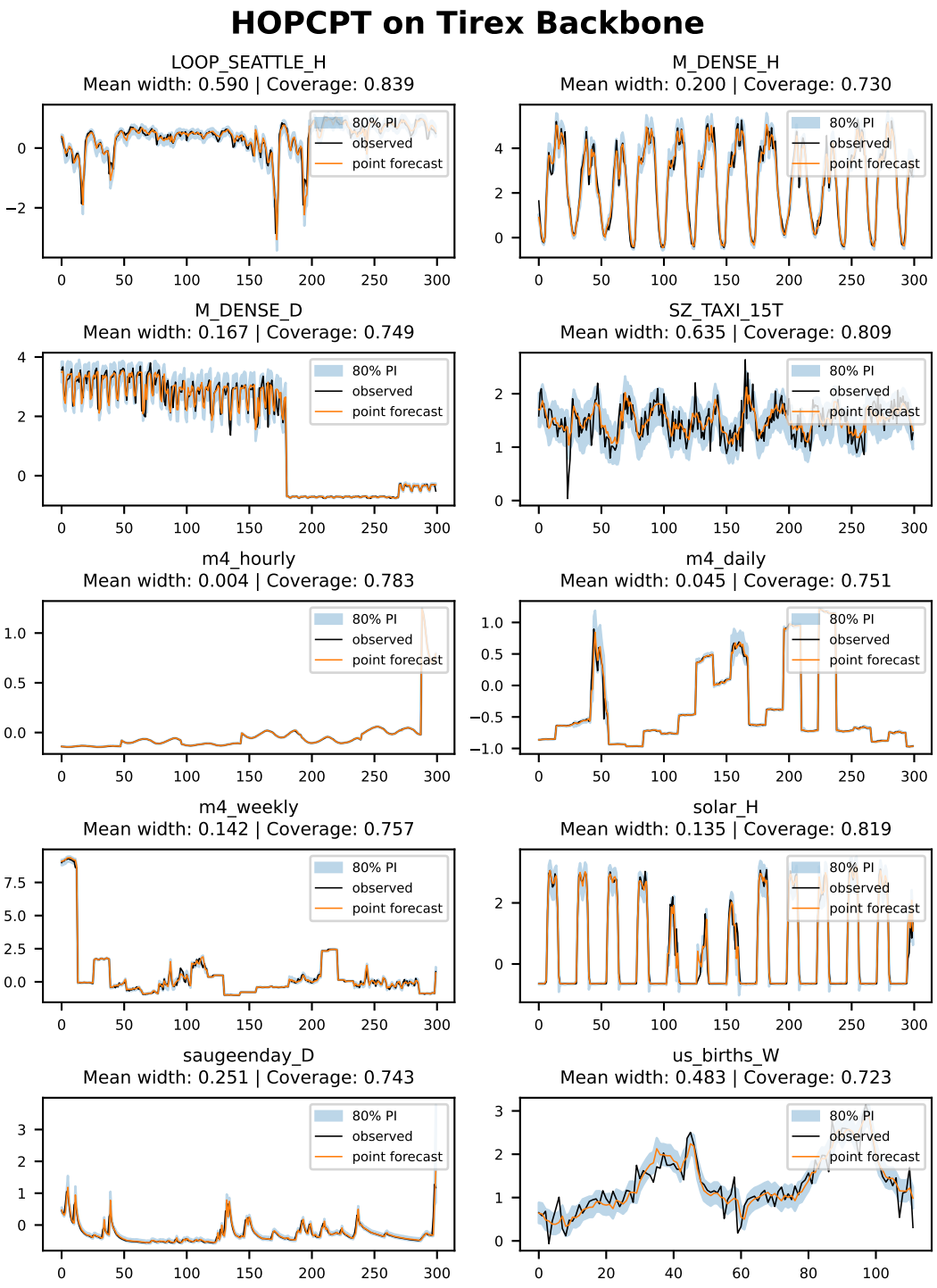


Figure 29: 80% prediction intervals from HopCPT on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

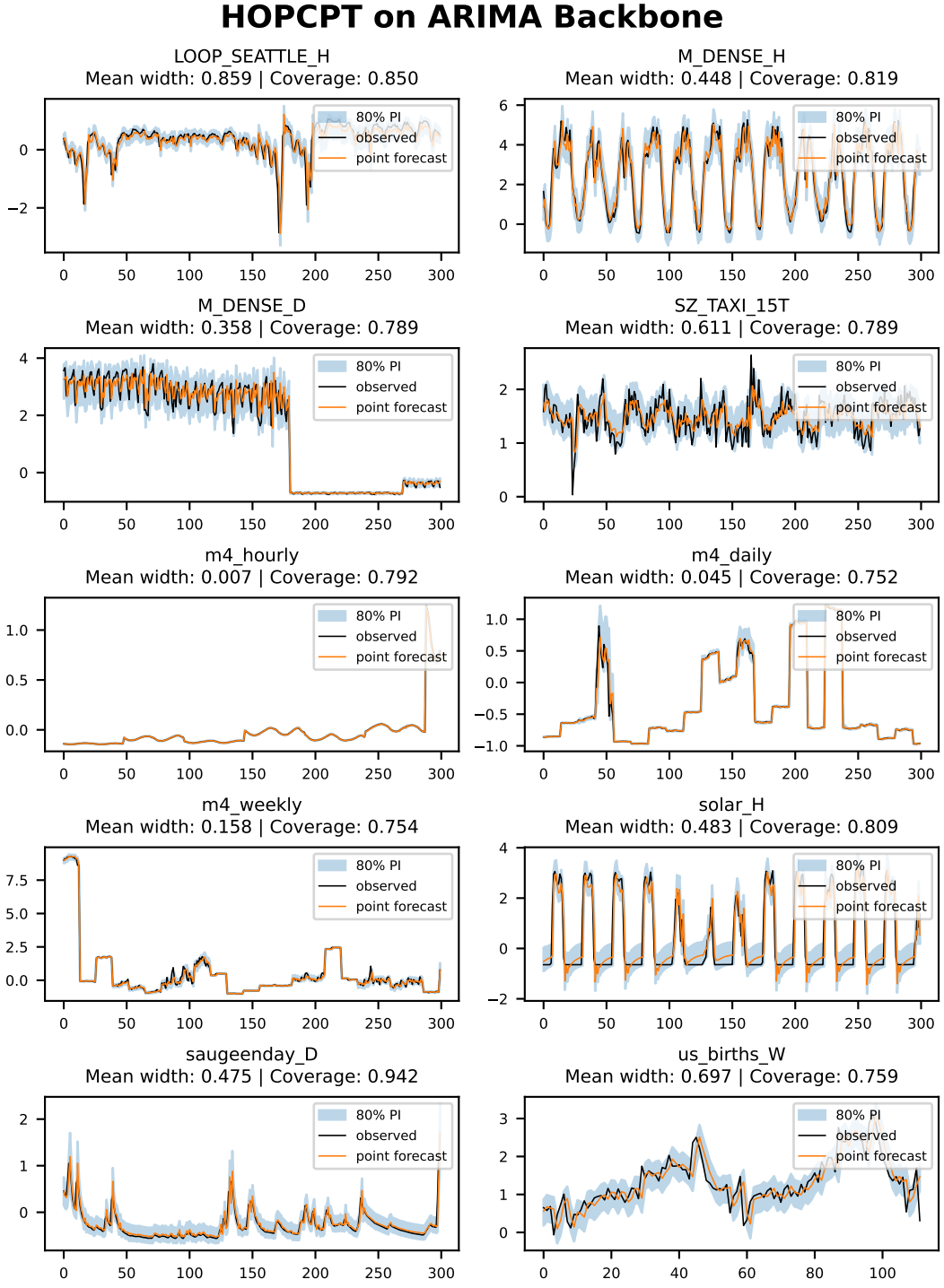


Figure 30: 80% prediction intervals from HopCPT on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

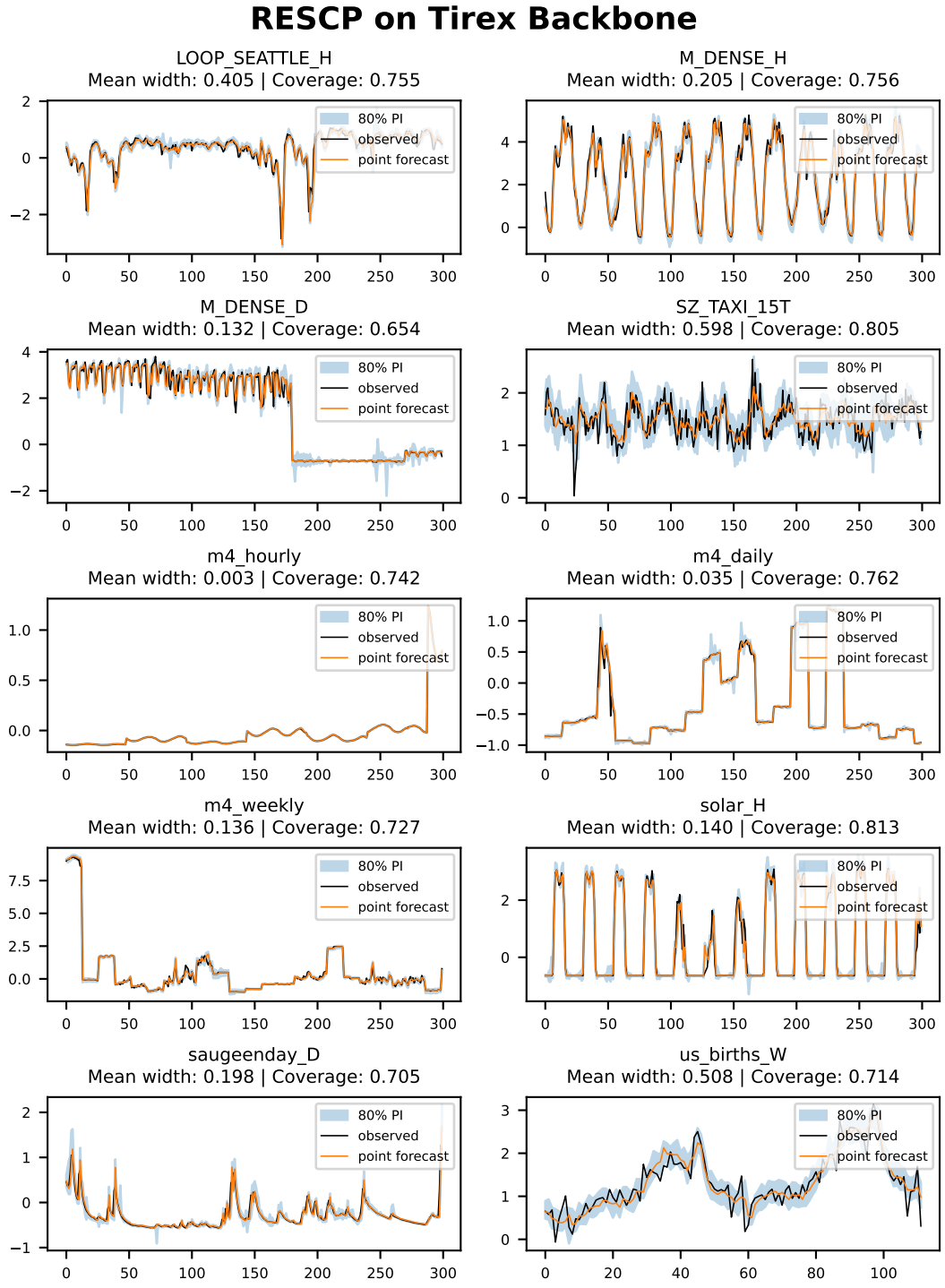


Figure 31: 80% prediction intervals from ResCP on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

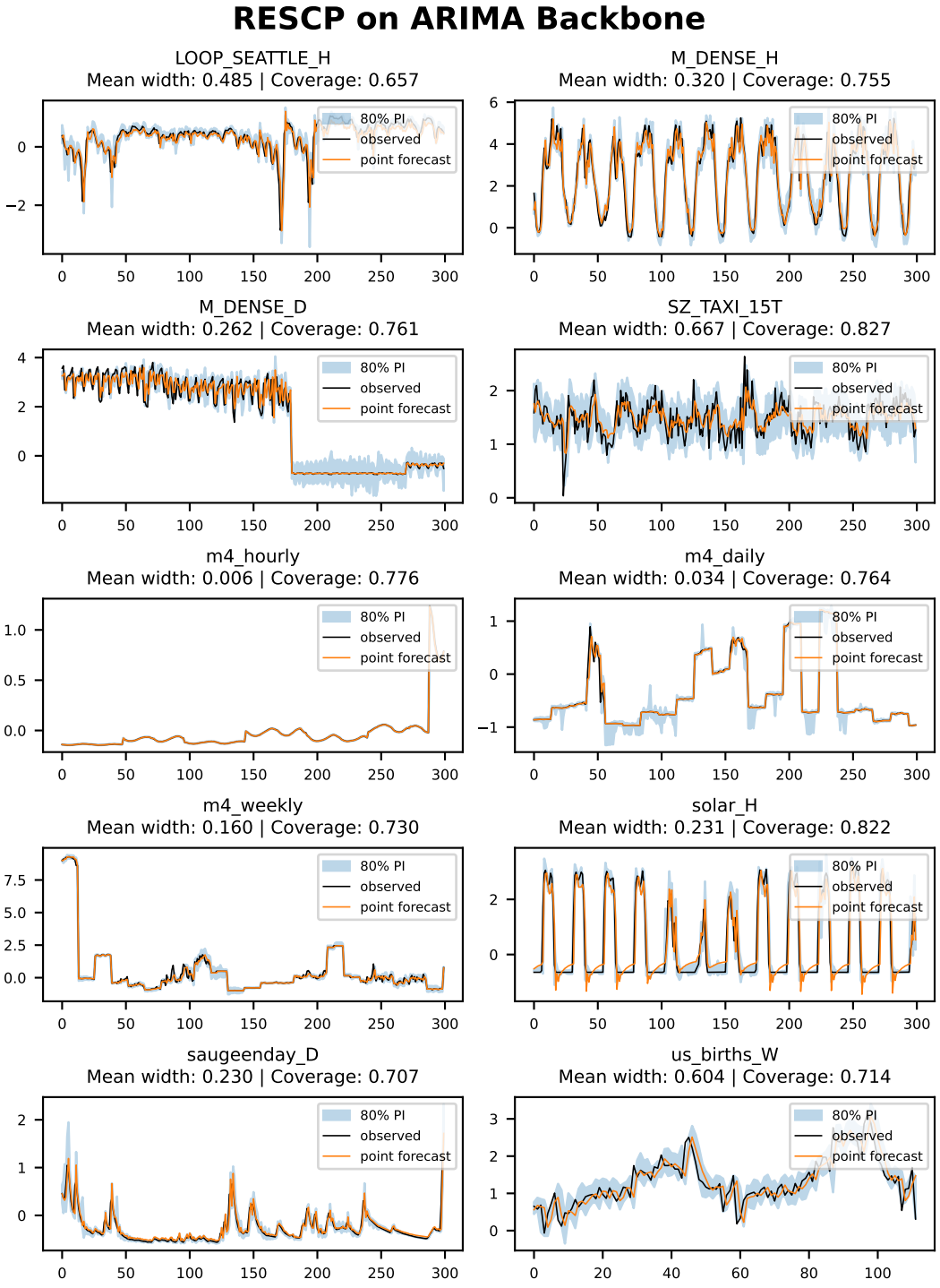


Figure 32: 80% prediction intervals from ResCP on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

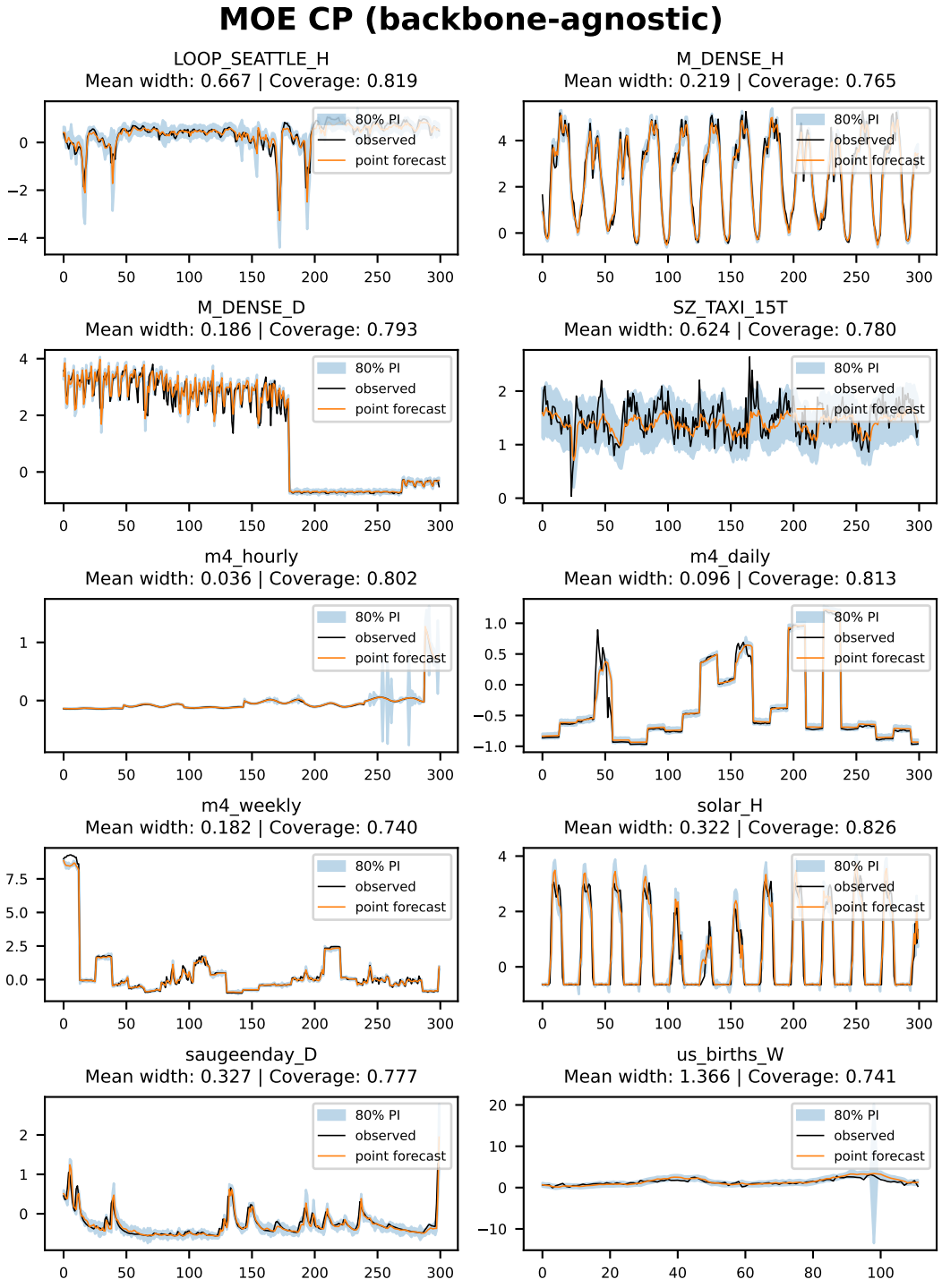


Figure 33: 80% prediction intervals from MoE-CP. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

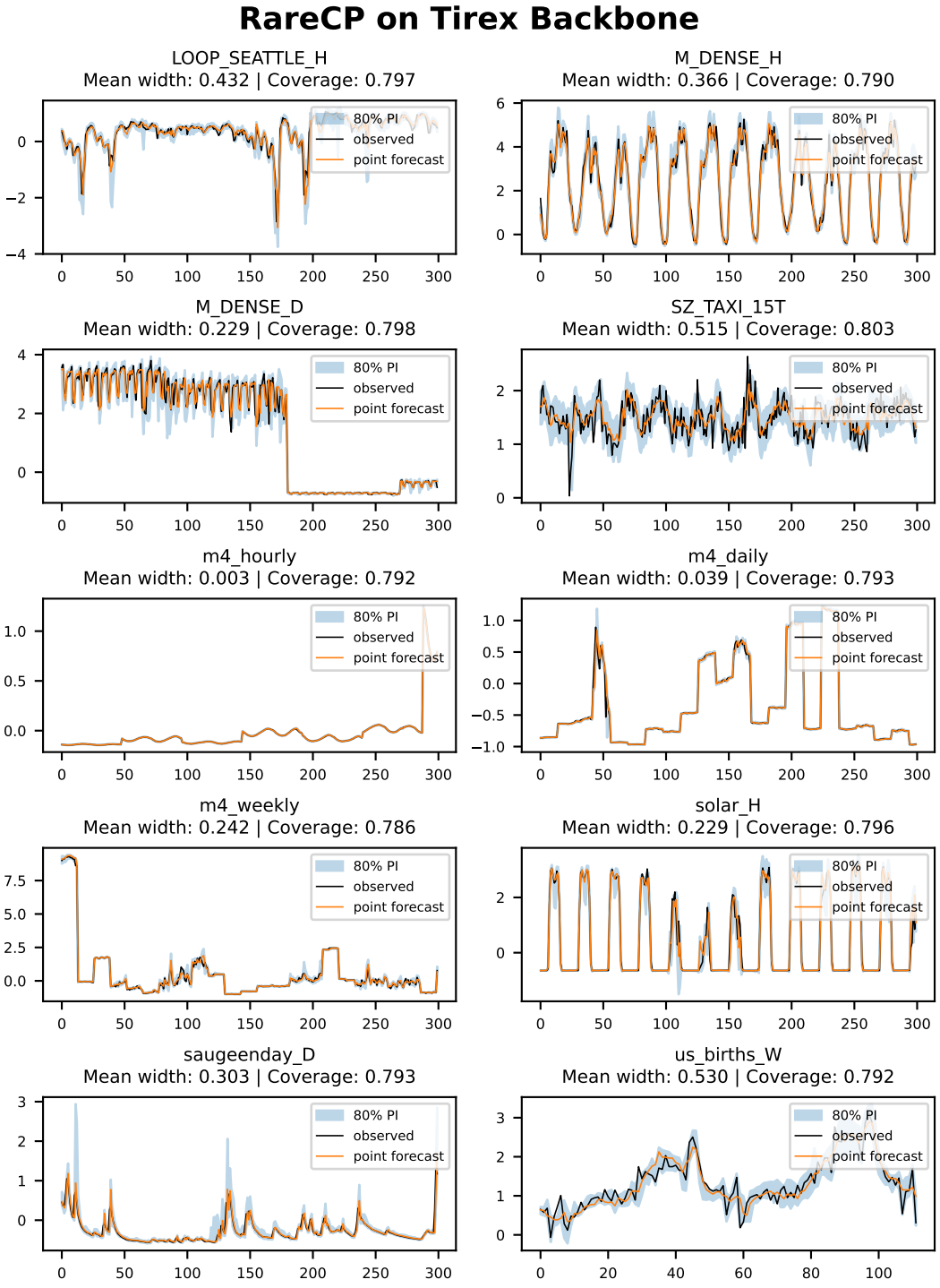


Figure 34: 80% prediction intervals from RareCP on the TiRex backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.

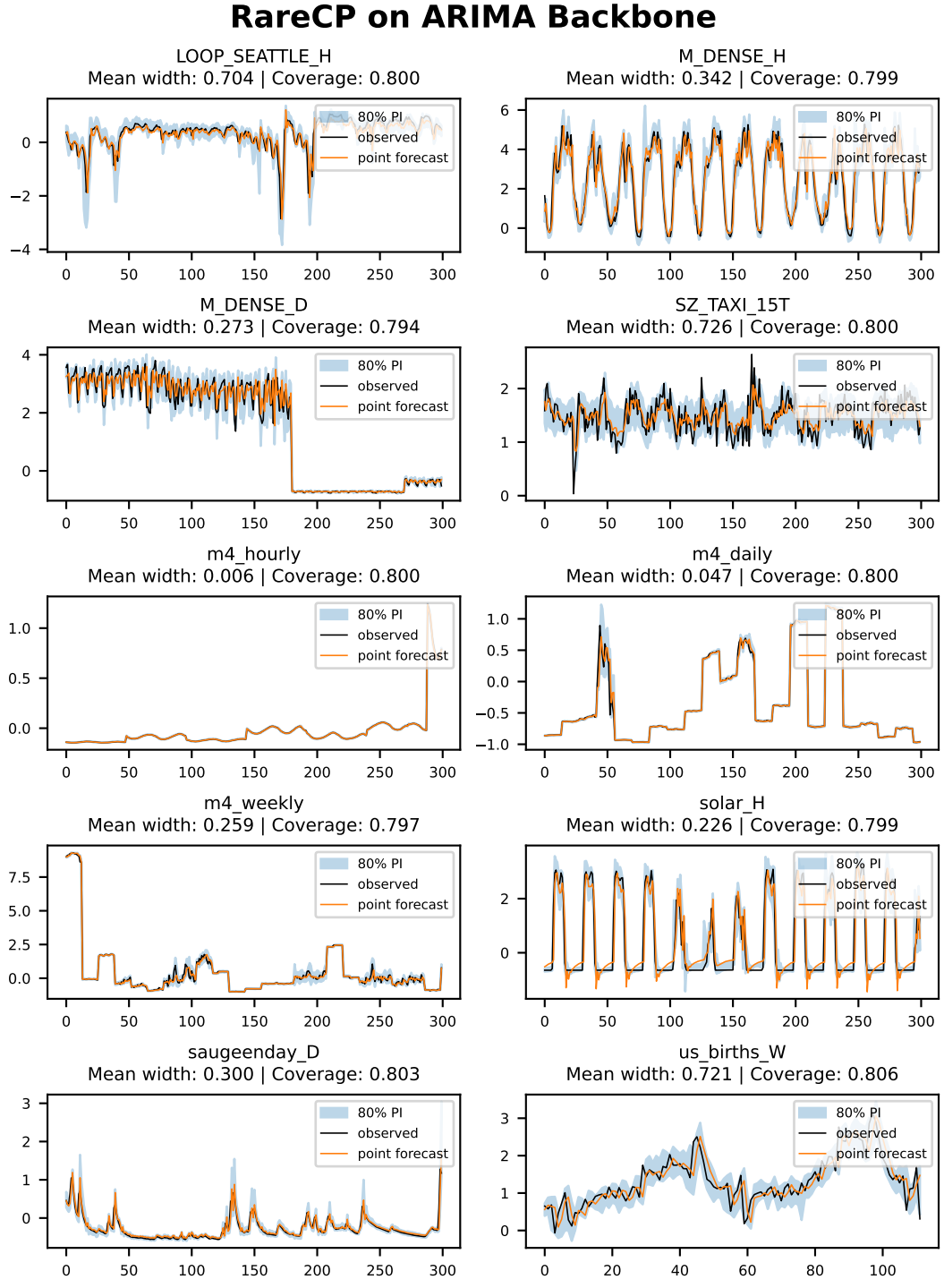


Figure 35: 80% prediction intervals from RareCP on the ARIMA backbone. Each panel shows up to 300 test targets (black), forecasts (orange), and intervals (blue), with mean width and empirical coverage reported above.