UCLA

UAI 2026
Submission #488

Model-Agnostic Online Certificate-Driven Calibration for Time Series Forecasting Under Distribution Shift

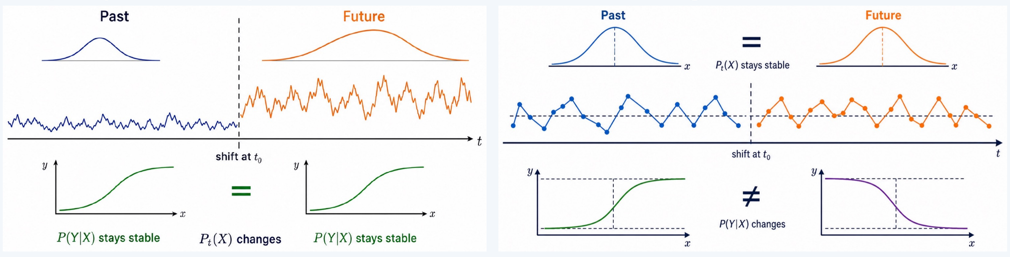
Chenfeng Huang · Zixuan Ma · George Michailidis
Department of Statistics and Data Science, University of California, Los Angeles

Online Martingale PAC-Bayes (OMPB) framework senses shift from unlabeled target windows, then conservatively calibrates only a gated Bayesian head as labels arrive.

1. **Martingale certificate** — variance-adaptive bounds for dependent streams
2. **Label-free shift sensing** — posterior disagreement on antecedent windows
3. **Conservative adaptation** — gated head-only updates; backbone fallback

1 The Problem

Time-series forecasters must stay reliable when deployment dynamics drift from training conditions:



Covariate shift **Concept shift**

Temporal dependence: serial correlation breaks the i.i.d. sampling behind standard bounds.

‣ **The operational challenge: react before the current target outcome is known.**

2 PAC-Bayes and Its Limits

$$\underbrace{R_{\mathcal{T}}(h_\rho)}_{\substack{\text{True risk}\\(\text{target domain } \mathcal{T})}} \le \underbrace{\hat{R}_S(h_\rho)}_{\substack{\text{Empirical risk}\\(\text{training sample } S)}} + \underbrace{\sqrt{\frac{KL(\rho \parallel \pi) + \ln\left(\frac{2\sqrt{m}}{\delta}\right)}{2m}}}_{\substack{\text{Complexity term}\\(\text{PAC-Bayes penalty})}}$$

Classical PAC-Bayes: risk ≤ empirical risk + complexity.

Limit 1 — i.i.d. only: the complexity term requires independent samples; autocorrelation breaks it.

Limit 2 — source only: a bound certified on ETTh1 does not automatically bound risk on a shifted target stream (ETTh2):

ETTh1 sample $S_1 \sim D^m$ → **Learn posterior** $\rho = \rho(S_1)$ → **Source-domain** bound certified → **✗ No bound on** shifted ETTh2 risk

3 Predict-Then-Update Protocol

Strict temporal causality at every step t — no look-ahead, no target labels in advance:

1. **History**
   $z_1, \dots, z_{t-1}$ and unlabeled target windows arrive before time t
2. **Predict**
   draw $h_t \sim \rho_{t-1}$, emit $\hat{y}_t = h_t(x_t)$ while y_t is hidden
3. **Reveal**
   y_t is disclosed; the loss is scored only after the forecast is committed
4. **Update**
   minimize $\mathcal{L}_{cal}$ to get ρ_t; append z_t and repeat at $t+1$

4 Martingale PAC-Bayes Bounds

Idea 1 — use martingale (not i.i.d.) concentration: each loss = predictable μ_t + zero-mean innovation D_t:

$$L_t = \mu_t + D_t, \qquad D_t := L_t - \mu_t$$

$$M_m := \sum_{t=1}^{m} D_t, \quad V_m := \sum_{t=1}^{m} E[D_t^2 \mid \mathcal{F}_{t-1}]$$

$$R_m^{pred}(\rho) - \hat{R}_m(\rho) = -\frac{1}{m} E_{h\sim\rho}[M_m(h)] \le \Gamma_m(\rho,\pi,\delta)$$

$$\Gamma_m^{sub\gamma}(\rho,\pi,\delta) = \sqrt{\frac{2\, E_{h\sim\rho}[V_m^{sub\gamma}]PB_m(\rho,\pi,\delta)}{m^2}}) + \frac{\bar{C}}{m}PB_m(\rho,\pi,\delta)$$

$$V_m^{sub\gamma}(h) := \sum_{t=1}^{m} v_t(h)$$

The empirical−predictable risk gap is the mean cumulative innovation, bounded by the sub-gamma complexity $\Gamma_m^{sub\gamma}$ — computable online, valid under temporal dependence.

5 The Online Certificate

Idea 2 — the domain-adaptation bound adds a mismatch term that needs **no target labels**:

$$R_{\mathcal{T}}(h_\rho) \le \hat{R}_S(h_\rho) + \Gamma_m^{sub\gamma}(\rho,\pi,\delta) + \frac{1}{2} dis_\rho(\mathcal{S},\mathcal{T}) + \lambda_\rho$$

Posterior disagreement on a target window w

$$\tilde{d}_{\tau_d}(h,h';w) = \min(1, \frac{\|h(w) - h'(w)\|_2^2}{HC\tau_d^2}) \in [0,1]$$

Scale τ_d selected once from the source replay buffer; q = 0.5

$$\tau_{\text{auto}} = \sqrt{\text{Quantile}_q\left(\frac{\|h^{(k)}(w) - h^{(k')}(w)\|_2^2}{HC}\right)}$$

Per-window estimate via unordered posterior pairs

$$\hat{d}_\rho(w) = \frac{2}{K(K-1)} \sum_{1\le k<k'\le K} \tilde{d}_{(\tau_d)}(h^k, h^{k'}; w)$$

Online source–target mismatch term

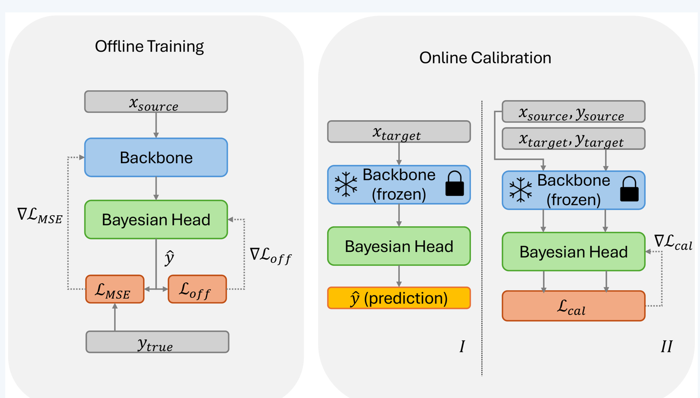
$$\widehat{dis}_\rho(t) = |\frac{1}{m}\sum_{t'=1}^{m} \hat{d}_\rho(w_{t'}) - \frac{1}{|W_{T,t}|}\sum_{w'\in W_{T,t}} \hat{d}_\rho(w')|$$

TIME-INDEXED ONLINE CERTIFICATE

$$\widehat{\mathrm{PB}}_\gamma(t) = \hat{R}_{S,t}(h_\rho) + \Gamma_m^{\mathrm{sub}\,\gamma}(\rho,\pi,\delta) + \tfrac{1}{2}\widehat{\mathrm{dis}}_\rho(t),$$

λ_ρ needs target labels → omitted online. The certificate monitors bounded proxies — not a raw MAE/MSE bound.

6 The Gated Bayesian Head



Adapt less, on purpose. Gate s = σ(α) scales a low-rank residual on the frozen backbone; s ≈ 0 returns the base forecast — conservative under stale labels.

ONLINE OBJECTIVE

$$\mathcal{L}_{\text{cal}}(t;\rho) = \widehat{\mathrm{PB}}_\gamma(t;\rho) + \hat{R}_{T,t}^{\text{sup}}(\rho).$$

7 Experimental Setup

ETTh — covariate shift (ETTh1 → ETTh2)
ILI — concept shift (pre-COVID → COVID)
WEATHER-5K — covariate shift (Miami → Atlanta → NYC)

Backbones: TCN · Autoformer · GPT4TS (model-agnostic)
Baselines: SOLID · OneNet · PROCEED (1-step delay)

All online hyperparameters are fixed before target evaluation — never tuned on target labels.

Dataset	Feat.	Freq.	Input	Train	Test
ETTh	7	Hourly	96	13,745	3,293
ILI	11	Weekly	24	443	108
WEATHER-5K	6	Hourly	96	78,764	8,646×2

8 Main Results

Best MAE in all 36 backbone × horizon × ID/OOD cells on ETTh + ILI (MSE: 35 of 36) — and lowest error on WEATHER-5K at every horizon and location.

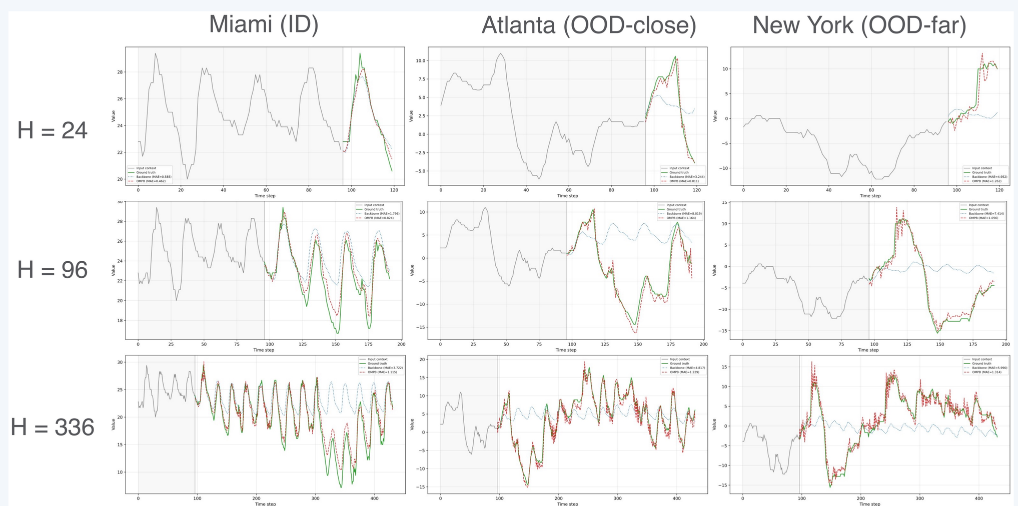
Table 1: ETTh forecasting results (MAE | MSE) under In-Distribution (ID) and Out-of-Distribution (OOD) settings. Lower score values indicate better performance, and the best score is in **bold**.

Backbone	Method	H=24		H=96		H=336	
		ID (ETTh1)	OOD (ETTh2)	ID (ETTh1)	OOD (ETTh2)	ID (ETTh1)	OOD (ETTh2)
TCN	Original	0.4588 | 0.4521	2.5145 | 8.9095	0.5330 | 0.5715	2.5635 | 8.9076	0.6193 | 0.7304	3.1380 | 13.7387
	SOLID	0.4741 | 0.4333	0.7553 | 1.0175	0.5363 | 0.5389	0.8886 | 1.4424	0.7110 | 0.9710	0.9488 | 1.7629
	OneNet	0.5838 | 0.6739	0.8234 | 1.3034	0.6606 | 0.8488	1.0073 | 1.9592	0.7082 | 0.9191	1.2228 | 3.3373
	PROCEED	0.4917 | 0.5161	1.1002 | 0.7768	0.5889 | 0.7086	0.9608 | 1.8119	0.6490 | 0.8225	1.1293 | 2.5799
	OMPB	**0.3635 | 0.2724**	**0.5292 | 0.5659**	**0.3813 | 0.2800**	**0.5570 | 0.5912**	**0.4484 | 0.3674**	**0.5981 | 0.6782**
Autoformer	Original	0.5147 | 0.5605	1.1207 | 1.9961	0.5325 | 0.5651	1.3298 | 2.6778	0.6830 | 0.8846	1.1962 | 2.3453
	SOLID	0.4899 | 0.4813	0.8410 | 1.2997	0.5446 | 0.5629	0.9501 | 1.6828	0.5937 | 0.6374	1.0386 | 2.0287
	OneNet	0.4786 | 0.4815	0.7311 | 0.9580	0.5571 | 0.6026	0.8395 | 1.2940	0.6140 | 0.6861	0.9812 | 1.8274
	PROCEED	0.5119 | 0.5967	0.7580 | 1.1151	0.6827 | 1.0923	1.0509 | 2.7848	0.8767 | 1.8344	1.5051 | 4.6450
	OMPB	**0.3814 | 0.3018**	**0.6785 | 0.8535**	**0.3845 | 0.2858**	**0.6033 | 0.6882**	**0.4373 | 0.3481**	**0.6461 | 0.7764**
GPT4TS	Original	0.3794 | 0.3217	0.5485 | 0.7119	0.4473 | 0.4167	0.6738 | 1.0241	0.5320 | 0.5588	0.8547 | 1.5203
	SOLID	0.3763 | 0.3218	0.6172 | 0.7745	0.4451 | 0.4160	0.7499 | 1.1224	0.5333 | 0.5625	0.9131 | 1.6823
	OneNet	0.3740 | 0.3184	0.6115 | 0.7512	0.4466 | 0.4146	0.7489 | 1.1222	0.5364 | 0.5572	0.8975 | 1.6239
	PROCEED	0.4843 | 0.5082	0.7210 | 1.0268	0.6237 | 0.8802	0.9364 | 1.8907	0.6911 | 1.0597	1.1482 | 2.8742
	OMPB	**0.3156 | 0.2165**	**0.5242 | 0.6616**	**0.3535 | 0.2457**	**0.5708 | 0.7377**	**0.4095 | 0.3158**	**0.5719 | 0.7304**

Table 2: ILI forecasting results (MAE | MSE) under In-Distribution (ID) and Out-of-Distribution (OOD) settings. Lower score values indicate better performance, and the best score is in **bold**.

Backbone	Method	H=24		H=48		H=72	
		ID (Pre-COVID)	OOD (COVID)	ID (Pre-COVID)	OOD (COVID)	ID (Pre-COVID)	OOD (COVID)
TCN	Original	1.6734 | 5.8091	4.5372 | 38.1840	1.2409 | 3.2224	2.6306 | 11.5707	1.9092 | 8.1936	4.0072 | 32.0801
	SOLID	1.2047 | 3.5700	1.7447 | 7.4170	1.5927 | 5.3366	3.1461 | 32.3844	1.5807 | 5.4009	2.2503 | 13.2387
	OneNet	1.6995 | 6.9002	6.7105 | 81.2448	1.8328 | 7.6513	7.3912 | 115.1614	1.7588 | 6.7818	7.9340 | 143.7589
	PROCEED	1.4870 | 4.6585	1.8515 | 10.9656	1.2260 | 3.1388	1.7905 | 9.2732	1.3271 | 3.7264	1.9005 | 10.6453
	OMPB	**0.7102 | 1.1751**	**1.2672 | 6.3903**	**0.5740 | 0.9482**	**0.7109 | 1.8331**	**0.5960 | 1.2583**	**0.7113 | 2.1360**
Autoformer	Original	1.2800 | 3.3924	2.7942 | 15.7648	2.0135 | 8.1263	2.6485 | 15.4891	2.0193 | 8.5962	2.6667 | 15.5638
	SOLID	1.2591 | 3.5508	4.0567 | 64.7122	1.2642 | 3.5153	2.3409 | 19.6094	1.4241 | 4.4793	2.2399 | 14.5178
	OneNet	1.5966 | 5.5776	2.4633 | 19.4883	1.4029 | 4.6296	2.1933 | 16.2578	1.4493 | 4.5256	2.2957 | 18.3962
	PROCEED	1.7888 | 7.1304	1.9178 | 11.4278	1.4210 | 4.1995	1.5856 | 7.5495	1.3469 | 3.8370	2.2889 | 17.9593
	OMPB	**0.9916 | 2.3576**	**1.3498 | 3.9055**	**1.1288 | 3.1272**	**1.2460 | 4.2611**	**1.2825 | 3.9065**	**1.4256 | 4.8619**
GPT4TS	Original	0.9667 | 2.3908	1.3245 | 4.7387	0.8972 | 1.9836	1.3013 | 4.2179	1.0006 | 2.1929	1.4140 | 5.0265
	SOLID	1.1803 | 3.5820	1.4930 | 6.8183	1.1177 | 3.2670	1.8180 | 13.2240	1.1957 | 3.6045	1.3675 | 6.7276
	OneNet	1.2891 | 4.0408	1.9931 | 14.5083	1.1923 | 3.5878	1.9970 | 14.2298	1.3108 | 4.0760	2.2679 | 17.8664
	PROCEED	1.6876 | 7.4868	2.2322 | 16.6780	1.1823 | 3.3912	1.8764 | 10.6957	1.2604 | 3.8947	2.1143 | 13.9342
	OMPB	**0.8543 | 1.9178**	**1.0906 | 3.3870**	**0.6248 | 1.0503**	**0.9263 | 2.6351**	**0.6909 | 1.3602**	**1.0130 | 2.9065**

9 Qualitative Forecasts

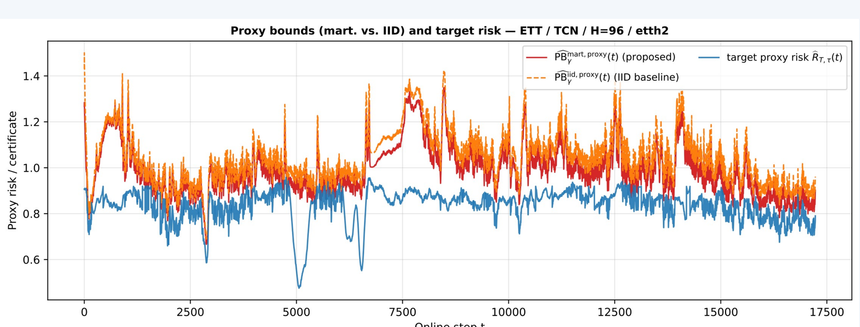


WEATHER-5K, GPT4TS backbone: OMPB (red) re-locks onto the ground truth (green) where the backbone (blue) drifts.

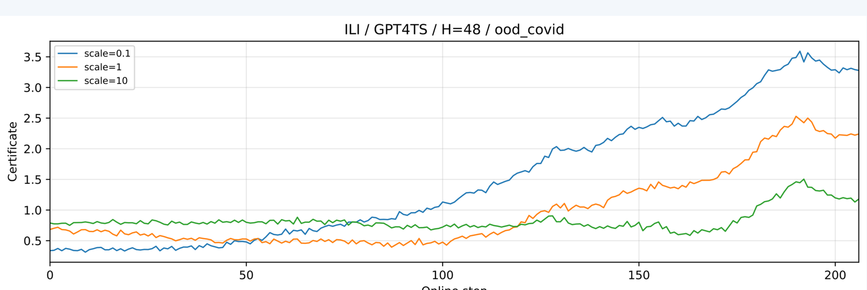
−27.3%	−53.2%	−59.2%
Miami	Atlanta	New York

MAE vs. the unadapted backbone (Original → OMPB).

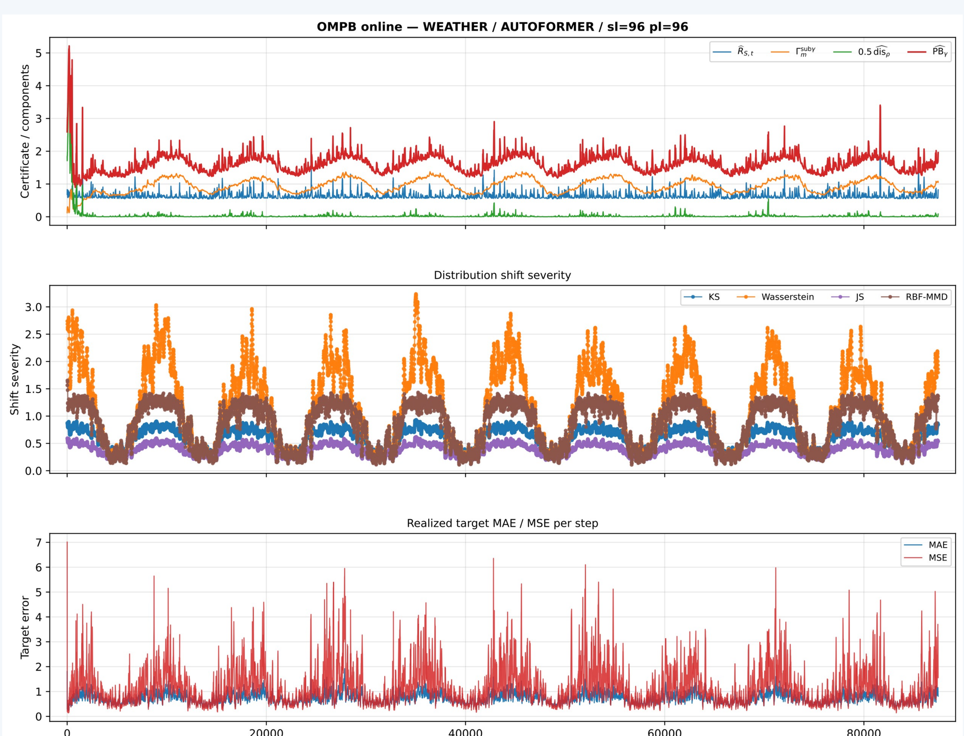
10 Certificate Behavior



Martingale proxy (red) is tighter than i.i.d. (orange).



Mis-scaled posteriors inflate the certificate — miscalibration is visible online.



WEATHER-5K / Autoformer: the certificate rises with shift severity and responds to realized error spikes.

Note: Certificate is a deployment-time monitoring signal for shift — not a per-step MAE/MSE upper bound

11 Ablations

Table 4: Partial-backbone adaptation diagnostic for OMPB. "Trainable" reports additional backbone parameters updated online beyond the Bayesian head.

Setting	Mode	Trainable	MAE	MSE
ETTh / TCN OOD ETTh2, $H = 96$	Frozen	0	**0.5570**	**0.5912**
	Last layer	22,176	0.7124	0.8834
	LoRA	2,816	0.7410	1.0474
ILI / GPT4TS OOD COVID, $H = 48$	Frozen	0	**0.9263**	**2.6351**
	Last layer	110,640	1.4808	6.8875
	LoRA	9,408	1.6591	9.3536
WEATHER-5K / Autoformer OOD far NYC, $H = 96$	Frozen	0	**0.4344**	**0.4720**
	Last layer	780	0.7401	0.9791
	LoRA	8,192	0.9681	1.5443

Frozen backbone + gated head beats last-layer and LoRA online updates on all three OOD streams — extra online flexibility hurts under delayed supervision.

Feedback delay: accuracy degrades gracefully from 1-step to 2H label delay (MAE 0.56 → 1.86 on OOD ETTh2), staying below the unadapted backbone.

Paper

Code

The work of GM is partially supported by NSF ATD 2319552 and DMS 2348640.